ORCA: Open-ended Response Correctness Assessment for Audio Question Answering

Šimon Sedláček[1*], **Sara Barahona**[2*], **Bolaji Yusuf**[1*], **Laura Herrera-Alarcón**[2*], **Santosh Kesiraju**[1*], **Cecilia Bolaños**[3], **Alicia Lozano-Diez**[2], **Sathvik Udupa**[1], **Fernando López**[2], **Allison Ferner**[4], **Ramani Duraiswami**[5], **Jan Černocký**[1]

[1]Speech@FIT, Brno University of Technology, Czechia [2]Universidad Autónoma de Madrid, Spain
[3]University of Buenos Aires, Argentina [4]Tufts University, USA [5] University of Maryland, USA
{isedlacek,iyusuf,kesiraju}@fit.vut.cz

Abstract

Evaluating open-ended responses from large audio language models (LALMs) is challenging because human annotators often genuinely disagree on answer correctness due to multiple valid interpretations, partial correctness, and subjective judgment. Traditional metrics reporting only mean scores fail to capture this uncertainty. We present ORCA (Open-ended Response Correctness Assessment), a framework that models the variability in human judgments using Beta distributions to predict both expected correctness and uncertainty. Our three-stage annotation framework combines human judgment with structured feedback and iterative refinement to simultaneously curate training data and improve benchmark quality. We collected 11,721 annotations across 3,580 question-answer pairs from 15 LALMs on two audio QA benchmarks, achieving inter-annotator agreement of 0.82 (Krippendorff's alpha). ORCA achieves 0.91 Spearman correlation with mean human judgments, matching or outperforming LLM-judge baselines while providing uncertainty estimates and requiring significantly less compute. We release our models, code, and curated dataset[1].

1 Introduction

Large audio language models (LALMs) are emerging as powerful tools for audio understanding, capable of answering questions about speech, music, and environmental sounds (Goel et al., 2025; Gemma Team, 2025b). To advance these models, efficient, and reliable evaluation is essential. Most LALM evaluations use the multiple-choice question (MCQ) framework (Sakshi et al., 2024; Bhattacharya et al., 2025), as it enables fast and automatic assessment. While MCQ evaluation is convenient, open-ended responses are more natural and better reflect real-world usage (Balepur et al., 2025). However, evaluating open-ended responses poses significant challenges: (a) multiple semantically equivalent answers are possible (Bulian et al., 2022), (b) partial correctness is common, and (c) human judgments – often considered the gold standard – exhibit genuine variability due to subjectivity or ambiguity.

An important aspect of audio question answering (QA) evaluation is that human annotators often genuinely disagree on answer correctness. Audio QA often involves subjective judgments about acoustic events, musical characteristics, emotional tone, or sarcasm – domains where multiple valid interpretations exist. Traditional evaluation metrics that report only mean scores fail to capture this dimension of uncertainty, conflating high-consensus scenarios with legitimate disagreement.

Figure 1 illustrates this phenomenon: two question-answer pairs receive similar mean ratings ($\approx$ 3.0 on a 1-5 scale), yet one exhibits high annotator agreement (variance=0.2) while the other shows genuine disagreement (variance=2.7). The first involves inferring a relationship from specific dialogue cues that annotators interpret consistently; the second asks for a subjective acoustic description where valid perspectives diverge. Traditional metrics lose this distinction, treating both scenarios identically despite their qualitatively different nature. To our knowledge, no prior work has modeled the full distribution of human correctness judgments for audio QA evaluation.

Existing evaluation metrics fail to capture this distributional information, treating all scenarios as point estimates. To address this limitation, we present ORCA (**O**pen-ended **R**esponse **C**orrectness **A**ssessment) for audio QA, a framework that models the variability in human judgments using Beta distributions. For practical scalability, we adopt a text-only evaluation approach

[1]https://github.com/BUTSpeechFIT/ORCA

Q: What is the link between the speakers?
Rationale: The female speaker addresses the male speaker as "coach," clearly indicating a coaching relationship. Their discussion focuses on "jumping technique," "keeping my heels down," and "timing on the jumps," which are all terms specific to equestrian sports, confirming the roles of an equestrian coach and their rider.
Reference answer: equestrian coach-rider
Candidate answer: They are a coach and a student.
Human ratings: [3, 3, 3, 3, 4] ($\mu = 3.2, \sigma^2 = 0.2$)

Q: How would you describe the sound texture of the audio?
Rationale: The audio features a loud, indistinct roar of what sounds like a crowd or a large group of people chanting. The low fidelity and overlapping sounds make it difficult to discern individual elements, creating a muffled and chaotic texture.
Reference answer: Muffled and chaotic
Candidate answer: The sound texture of the audio is complex and layered, with multiple sound events occurring simultaneously.
Human ratings: [1, 3, 3, 5] ($\mu = 3.0, \sigma^2 = 2.7$)

Figure 1: Two examples with similar mean ratings ($\mu \approx 3$) but different variances: low variance (top) indicates consensus; high variance (bottom) reveals disagreement.

in which metrics assess the correctness of the answer augmented by textual grounding (rationales and transcriptions) (Yang et al., 2024), avoiding the circular dependency of using audio models to judge audio models. This framework applies uniformly to human annotators, ORCA, and LLM-judge baselines.

The contributions of this work are as follows.

- We propose ORCA, a lightweight model-based answer correctness assessment framework that predicts both the mean and variance of human correctness judgments using Beta distribution, enabling uncertainty quantification for audio QA evaluation.
- We introduce a three-stage annotation framework that combines human judgment with structured feedback and iterative refinement, yielding high-quality training data while improving the underlying benchmark quality.
- We collected 11,721 human annotations across 3,580 question-answer pairs from 15 state-of-the-art LALMs on two benchmarks (MMAU and MMAR), achieving Krippendorff's alpha of 0.82 after filtering.
- We conducted extensive experiments that demonstrate the effectiveness of ORCA compared to multiple LLM-as-a-judge baselines, achieving a spearman correlation of 0.91 and the lowest mean absolute error to average human judgments.
- We will release our trained models, source code, and curated annotation dataset to facilitate further research in this area.

2 Related work

Audio question answering benchmarks have evolved from foundational perception tasks (Yang et al., 2024; Wang et al., 2025) to complex reasoning scenarios (Sakshi et al., 2024; Ma et al., 2025; Kumar et al., 2025), evaluating large audio language models (LALMs) across diverse capabilities.

2.1 Open-ended answer evaluation

Evaluating open-ended responses presents challenges due to semantic equivalence, partial correctness, and subjective interpretation. Traditional lexical matching (exact match, F1, BLEU) does not capture semantic similarity (Bulian et al., 2022). Recent work explores LLM-based judges such as Prometheus (Kim et al., 2024) for text-based QA, while audio QA benchmarks like AIR-Bench (Yang et al., 2024) and AudioBench (Wang et al., 2025) employ API-based LLM judges with textual representations to avoid circular dependency. However, these approaches assume correctness of reference answers and rationales, and LLM judges suffer from prompt sensitivity (Nalbandyan et al., 2025), lack of calibration, high computational costs, and reproducibility concerns.

Our work addresses these limitations through our three-stage annotation framework and the flexibility of the proposed ORCA model. With the former, we systematically validate and refine textual grounding with structured feedback mechanisms and this process serves dual purposes: generating calibrated training data for ORCA while improving benchmark quality.

2.2 Human annotation variability

Human annotation disagreement in NLP tasks often reflects genuine ambiguity rather than noise (Plank, 2022; Sandri et al., 2023). Several works have explored modeling label distributions and uncertainty quantification in text-based tasks (Liu et al., 2023; Wu et al., 2024; Leonardelli et al., 2023), no previous work has applied distributional modeling to assess answer correctness for audio QA.

3 Framework for human annotation collection

Figure 2 illustrates our three-stage process for collecting high-quality human judgments for text-only evaluation while systematically improving benchmark quality.

3.1 Stage 1: Data preparation

We prepare data through two parallel operations.

Stage 1a: Rationale and transcript generation To enable text-only evaluation models, we augment the benchmark data with textual grounding information (also referred to as *context*). A *rationale* is a textual justification explaining why the reference answer is correct given the audio and question. We generate rationales using Gemini-2.5-Flash (Gemini Team, 2025), prompting it with the audio, question, and reference answer. The rationale typically includes an audio description, reasoning steps, and when applicable, external knowledge. Gemini sometimes generates rationales based on textual cues rather than audio content, producing non-informative justifications. We address this through human feedback in Stage 2, with corrections in Stage 3.

For questions from speech-based categories, we also generate transcriptions using Whisper-large-v3 (Radford et al., 2023).

Stage 1b: Candidate answer generation We generate candidate answers by prompting multiple LALMs with the question and audio. This yields diverse responses with different reasoning patterns, error modes, and correctness levels. The specific models used are detailed in Section 5.

3.2 Stage 2: Annotation and feedback collection

Human annotators evaluate answer correctness given five pieces of textual information: the question, reference answer, rationale, audio transcript (for speech questions), and the candidate answer. When textual context proves insufficient, annotators can listen to the original audio directly.

Annotators assign correctness scores on a 1-5 scale (detailed rating definitions in Appendix A) and provide structured feedback identifying issues through predefined categories: **Q** (incomplete question), **A** (insufficient rationale), **R** (incorrect reference), **U** (ambiguous), **E** (lacking expertise). Additional free-form comments capture issues not covered by these categories. This feedback enables targeted corrections in Stage 3.

LLM-judges perform parallel evaluations with the same textual inputs. Statistics, rating scales, and protocols are detailed in Section 5 and Appendix A.

3.3 Stage 3: Iterative refinement

The structured feedback from Stage 2 enables systematic quality improvement through collaborative human-AI correction.

Correction process Domain experts review flagged instances and implement corrections: rephrasing questions (**Q**), enhancing rationales (**A**), correcting reference answers (**R**), and fixing transcripts. AI tools assist in generating improved content, but human experts validate all changes to ensure quality.

Iterative loop Corrected data feeds back into Stage 1b to regenerate candidate answers, creating a continuous improvement cycle that enhances both annotations and benchmarks. Correction statistics are in Section 5 and Appendix A.

4 The ORCA model

Open-ended answer evaluation presents a fundamental challenge: Multiple human annotators often disagree on the correctness of the same answer, reflecting genuine ambiguity rather than annotation noise. Unlike traditional metrics that produce a single correctness score, ORCA models the full distribution of human judgments, capturing both expected correctness and the degree of uncertainty or disagreement among annotators.

4.1 Model architecture

ORCA is initialized from a pre-trained transformer-based LLM. Given an evaluation instance consisting of a question q, reference answer r, rationale a, audio transcript t, and candidate answer c, we concatenate these elements using separator tokens to obtain the final input: $x = [q; r; a; t; c]$, where semicolons denote separators. The concatenated input is tokenized and fed through the pre-trained LLM, producing contextualized representations. We extract the final hidden representation and apply a multilayer perceptron (MLP) to predict two values:

$$\log \alpha, \log \beta = \mathrm{MLP}(\mathbf{h}_{\text{final}}) \quad (1)$$

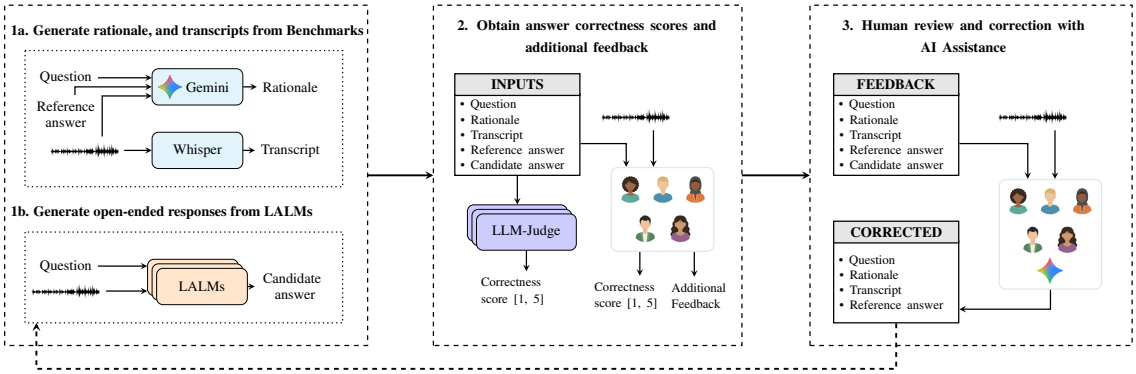

Figure 2: Annotation framework pipeline. Data preparation (Stage 1) generates rationales via Gemini, transcripts via Whisper, and candidate answers from LALMs. Annotation (Stage 2) collects correctness scores and structured feedback from humans and LLM-judges. Iterative refinement (Stage 3) implements human-AI corrections based on feedback, with corrected data re-entering the pipeline (Stage 1b).

where $\mathbf{h}_{\text{final}}$ denotes the final hidden representation. The model outputs $\log \alpha$ and $\log \beta$ to ensure that the Beta distribution parameters $\alpha = \exp(\log \alpha)$ and $\beta = \exp(\log \beta)$ remain positive. These parameters define a Beta distribution on correctness scores in the range $[0, 1]$.

4.2 Beta distribution for modeling judgments

The Beta distribution is a natural choice for modeling the distribution of answer correctness ratings for several reasons. First, it is defined on the interval $[0, 1]$, matching our normalized rating scale. Second, it is flexible enough to capture diverse rating patterns: high consensus (low variance) when most raters agree, high disagreement (high variance) when opinions diverge, and even U-shaped bimodal distributions at the extremes when raters are polarized between considering an answer completely correct or completely incorrect. Critically, this approach can model multiple judgments from any source – human annotators, LLM judges, or other evaluation methods, treating each rating as a sample of the underlying distribution.

The Beta distribution is parameterized by $\alpha > 0$ and $\beta > 0$, with probability density function:

$$\text{Beta}(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)} \tag{2}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function. The expected correctness score is given by:

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{3}$$

and the variance is:

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4}$$

Thus, the predicted parameters α and β provide both a point estimate (the mean μ) and an uncertainty estimate (the variance σ^2).

We frame the training as a maximum likelihood estimation problem. Given a dataset $\mathcal{D}$ consisting of evaluation instances, where each instance i has multiple correctness ratings $\{y_{i,1}, y_{i,2}, \ldots, y_{i,N_i}\}$ from different annotators, we treat each rating as an independent sample from the underlying Beta distribution. Ratings are normalized from the original 1-5 scale to the range $[0, 1]$.

The training objective is to maximize the log-likelihood:

$$\mathcal{L}_{\text{Beta}}(\theta) = \sum_{i \in \mathcal{D}} \sum_{j=1}^{N_i} \log \text{Beta}(y_{i,j}; \alpha_i, \beta_i) \tag{5}$$

where α_i and β_i are the parameters predicted by the model for instance i, and θ represents all the trainable model parameters.

5 Experiments

5.1 Benchmarks and data collection

We conducted experiments using two audio question-answering benchmarks: $\text{MMAU}_{\text{v05.15.25}}$ (test-mini subset) containing 1,000 questions spanning speech, sound, and music modalities, and MMAR with 1,000 questions across speech, sound, music, and mixed-source modalities.

Stage 1: Candidate answer and context generation Following the data preparation framework described in Section 3, we augmented each question with rationales and transcripts (Stage 1a), and generated candidate answers from 15 state-of-the-art audio language models (Stage 1b). The 15

LALMs include Audio Flamingo 2 & 3, Audio Reasoner, DeSTA2 & DeSTA2.5-Audio, GAMA, Gemma-3n (2B, 4B), GLM-4-Voice, Kimi-Audio, Qwen2-Audio-7B & Qwen2-Audio-7B-Instruct, Qwen2.5-Omni-7B, and SALMONN (7B, 13B). This yielded a total of 30,000 question-candidate answer pairs (2,000 questions × 15 models) for evaluation.

5.2 Human annotation study

We collected human judgments on candidate answers across both benchmarks, following the process described in Section 3.

Stage 2: Initial annotations We built the annotation interface using the POTATO annotation tool (Pei et al., 2022), customized for our task requirements, and refined it through two pilot studies. Over a 4-week period, 37 annotators (graduate students, researchers, and professors known to us) evaluated candidate answers from 15 LALMs. Annotators received detailed instructions with illustrative examples. Out of 2,000 benchmark questions, 1,872 received at least one annotation (rating or feedback), leaving 128 questions unannotated. This coverage yielded 3,580 question-candidate answer pairs with correctness ratings (out of 30,000 possible pairs) and 1,993 feedback entries, totaling 11,721 annotations. Each rated pair received an average of 2.7 ratings.

Inter-annotator agreement on the 3,580 rated pairs, measured using Krippendorff's alpha (Krippendorff, 2019), yielded $\alpha = 0.76$, indicating substantial agreement. However, structured feedback revealed significant data quality issues: Annotators flagged problems with questions (**Q**), reference answers (**R**), or rationales (**A**). Furthermore, annotators reported when they were unable to judge due to ambiguity (**U**) or lack of expertise (**E**).

Stage 3: Review and correction Following the annotation feedback, six domain experts systematically reviewed all flagged instances over a two-week period, including the 128 questions that did not receive any annotations. Using the human-AI collaborative correction process described in Section 3, experts corrected problematic fields where necessary. Table 1 summarizes the number of corrections by field type and benchmark.

Field Corrected	MMAU	MMAR	Total
Question (Q)	268	134	402
Reference Answer (R)	30	43	73
Rationale (A)	166	150	316
Total	464	327	791

Table 1: Number of corrections by field and benchmark

Filtering unreliable annotations Based on the Stage 3 review, we filtered out annotations for 541 benchmark questions (out of 1,872 annotated). This invalidated 1,121 question-answer pairs (out of 3,580 rated pairs) and their associated 3,150 ratings (32% of total). Notably, these filtered ratings had substantially lower agreement ($\alpha = 0.59$), confirming they represented unreliable instances rather than genuine disagreement. The remaining valid ratings across 2,459 question-answer pairs showed improved agreement ($\alpha = 0.82$).

Despite the improved agreement, we observe substantial variance in ratings: 17.7% of question-answer pairs have rating variance greater than 1.0, indicating genuine interpretive ambiguity rather than annotation noise. This pattern – illustrated by the contrasting examples in Figure 1 – motivates our distributional modeling approach in ORCA.

The corrected data improved both our training data and the underlying benchmarks. Further details are in Appendix A.

5.3 Evaluation protocol

To assess the robustness and generalization capability of ORCA and baseline LLM-judges, we design two complementary evaluation scenarios using the 2,459 valid question-answer pairs with stratified train/dev/test splits in an 8:1:1 ratio. For each scenario, we report performance metrics averaged across five splits with standard deviations.

Scenario 1: Unseen questions We evaluate generalization to novel questions not seen during training. We create five independent splits with different random seeds, each maintaining no overlap between training and test questions. To ensure balanced representation, we employ two-level stratification: primary stratification by audio modality (speech, sound, music) and secondary stratification by category (e.g., semantics, temporal reasoning, counting, perception, etc.).

Scenario 2: Unseen LALM responses We evaluate generalization to responses from previously unseen audio language models. We systematically hold out two LALMs in each split, creat-

ing five independent splits that collectively cover 10 different held-out LALMs.

5.4 Evaluation metrics

We assess model performance using four complementary metrics. For ranking quality, we compute Spearman's rank correlation coefficient (ρ) and Kendall's tau (τ) between predicted and human average correctness scores. For point estimate accuracy, we report mean absolute error for expected correctness (MAE_{μ}). Uniquely, since ORCA predicts the full Beta distribution of human judgments, we also evaluate variance prediction accuracy using mean absolute error for variance (MAE_{σ^2}).

5.5 ORCA training

We initialize ORCA from several pre-trained language models to assess robustness across different model families and scales. We experiment with OLMo-2 (1B, 7B) (Walsh et al., 2025), selected for its fully open-source nature; Gemma 3 (270M, 1B, 4B, 12B) (Gemma Team, 2025a); and Llama 3.2 (1B) (Llama Team, 2024). The model architecture consists of the pre-trained transformer encoder followed by a single-layer MLP that predicts log α and log β parameters for the Beta distribution.

ORCA is trained using the maximum likelihood objective described in Section 4, optimizing over all individual ratings. Training takes approximately 15 minutes per run on a single 24GB or 48GB GPU.

Post-processing ORCA's Beta distribution naturally avoids hard zeros/ones, unlike humans and LLM-judges, creating noisy predictions at the extremes. However, clear zeros/ones typically have low variance for both humans and ORCA, so we devise a simple *clamping* scheme: we set the ORCA score to hard zero/one if the original score is within 0.125 of zero/one and the predicted variance falls below a threshold optimized on the development set to maximize $\rho + \tau - \mathrm{MAE}_{\mu}$.

5.6 LLM-judge baselines

We evaluated several LLM-based answer-correctness judges as baseline systems. Our baseline selection covers both general-purpose models and evaluation-specialized systems.

Offline open-weight LLMs We evaluated four open-weight language models representing state-of-the-art reasoning and instruction-following capabilities: Llama 3.1-8B (Llama Team, 2024), Qwen2.5-7B (Qwen Team, 2025), and two variants (4B, 12B) from the Gemma 3 family (Gemma Team, 2025a).

Prometheus We include Prometheus 2-7B (Kim et al., 2024), an open-weight LLM explicitly fine-tuned for evaluation tasks, which rates responses based on rubrics and reference answers.

API-based LLM-judge We also evaluate Gemini-2.5-Flash (Gemini Team, 2025), a proprietary API-based model.

To ensure robustness, we evaluated all judges under four prompting conditions of increasing contextual richness: (1) reference and candidate answers only; (2) with question; (3) with rationale and transcript; (4) rationale only (for speech questions). Prompt templates are in Appendix B.

Aggregation We also evaluate aggregated predictions from multiple LLM-judges using two approaches: (1) simple averaging of individual judge scores, and (2) learned weighted aggregation with score calibration, where weights are optimized on the training set and applied to calibrated scores using the development set.

6 Results and analysis

6.1 Comparison of ORCA with LLM-as-a-judge

We train and evaluate several ORCA configurations on five unseen-question-based test sets from Section 5.3, comparing against individual LLM-judges, as well as their aggregated counterparts. The results are shown in Table 2.

ORCA models outperform LLM-judges of similar and often much larger sizes, (with the best judge being Gemini-2.5-Flash). ORCA's advantage is particularly evident in MAE_{μ} and MAE_{σ^2}, where even smaller ORCA models achieve lower errors than even aggregated LLM-judges. Clamping low-variance predictions to hard 0/1 values (Section 5.5) further improves correlation scores (especially Kendall's τ), addressing ORCA's tendency to produce soft probabilities rather than discrete judgments.

Overall, even the larger ORCA models (especially OLMo2-7B and Gemma3-12B) require only a single forward pass compared to numerous decoding steps of LLM-judges, as the judges

Model	Spearman ρ	Kendall τ	MAE_{μ}	MAE_{σ^2}
ORCA Models				
OLMo2-1B	0.8877 ± 0.0103	0.7418 ± 0.0139	0.1004 ± 0.0119	0.0223 ± 0.0049
OLMo2-1B (clamped)	0.8953 ± 0.0071	0.7842 ± 0.0143	0.0943 ± 0.0112	0.0223 ± 0.0049
OLMo2-7B	0.8973 ± 0.0116	0.7631 ± 0.0152	0.0861 ± 0.0047	**0.0179** ± 0.0015
OLMo2-7B (clamped)	0.8992 ± 0.0100	0.7900 ± 0.0145	**0.0827** ± 0.0036	**0.0179** ± 0.0015
Gemma3-270M (clamped)	0.8444 ± 0.0166	0.7209 ± 0.0199	0.1210 ± 0.0039	0.0229 ± 0.0008
Gemma3-1B (clamped)	0.8890 ± 0.0053	0.7704 ± 0.0089	0.0967 ± 0.0063	**0.0197** ± 0.0018
Gemma3-4B (clamped)	0.8981 ± 0.0160	0.7853 ± 0.0169	0.0945 ± 0.0104	0.0218 ± 0.0037
Gemma3-12B (clamped)	**0.9103** ± 0.0086	**0.8085** ± 0.0108	**0.0840** ± 0.0065	0.0199 ± 0.0028
Llama3.2-1B (clamped)	0.8926 ± 0.0091	0.7809 ± 0.0112	0.0933 ± 0.0064	0.0226 ± 0.0042
LLM Judges				
Gemma3-12B	0.8927 ± 0.0040	0.7989 ± 0.0072	0.1109 ± 0.0051	—
Gemma3-4B	0.8149 ± 0.0046	0.7095 ± 0.0045	0.1471 ± 0.0051	—
Llama3.1-8B	0.8435 ± 0.0102	0.7325 ± 0.0118	0.1279 ± 0.0060	—
Prometheus2-7B	0.7439 ± 0.0143	0.6362 ± 0.0127	0.1941 ± 0.0066	—
Qwen2.5-7B	0.8553 ± 0.0068	0.7514 ± 0.0077	0.1205 ± 0.0025	—
Gemini-2.5-Flash	**0.8998** ± 0.0066	**0.8070** ± 0.0073	0.0911 ± 0.0048	—
Average judge (-Gemini)	0.8902 ± 0.0065	0.7614 ± 0.0099	0.1172 ± 0.0026	0.0296 ± 0.0014
Average judge (+Gemini)	0.8972 ± 0.0064	0.7704 ± 0.0089	0.1078 ± 0.0017	0.0283 ± 0.0011
Judge Fusion (-Gemini)	0.8993 ± 0.0054	0.7682 ± 0.0087	0.1079 ± 0.0031	—

Table 2: Overall comparison of ORCA models and LLM-judges across different metrics, model sizes, and configurations. Values shown as mean ± standard deviation aggregated over five seeded annotation unseen question-based test sets described in Section 5.3.

also produce reasoning for the given rating. Importantly, only ORCA provides explicit uncertainty quantification: the low MAE_{σ^2} indicates that ORCA accurately predicts human annotator disagreement for each instance.

6.2 Comparison on held-out LALMs

We evaluate whether ORCA can generalize to LALMs (response styles) not seen during training. Using the OLMo2-7B-clamped configuration, we train on data from 13 LALMs and test on the remaining 2 held-out LALMs for each of five splits (10 unseen LALMs total). Figure 3 compares ORCA against Gemini-2.5-Flash and an offline LLM-judge fusion baseline trained on the same splits. ORCA consistently outperforms the LLM-judge fusion across all withheld LALMs. Compared to Gemini-2.5-Flash, ORCA achieves marginally lower correlation scores but matches or beats Gemini on MAE_{μ}, consistent with findings in Table 2. One notable exception is AudioReasoner, which generates significantly longer responses than other LALMs. ORCA struggles to generalize to this unseen response style, highlighting the importance of training data diversity.

6.3 Input ablations: question, rationale

We ablate the usage of rationales, transcriptions, and the original questions as inputs to both ORCA and LLM-judge models. This evaluation is once again done over the five unseen-question-based test splits, where for each ablation we report the mean achieved score and the standard deviation. The results are shown in Figure 4.

The *default* input configuration uses the rationale, but no transcript. We observe that adding the transcript tends to marginally worsen the MAE performance for both ORCA and LLM-judges, and so does removing the rationale. However, a more significant performance gap is seen when the question is removed from the input, where LLM-judge performance falls dramatically, and the same trends can be observed also for the correlation metrics. Further input ablation analysis for LLM-judges is provided in Appendix C.

6.4 Training on LLM-judge data

We additionally analyze the usage of LLM-judge data for training ORCA in five different scenarios: (1) training on the available human data, (2) training on LLM-judge ratings on the same question ids as in the first scenario, (3) training on all available LLM-judge data, (4) augmenting the

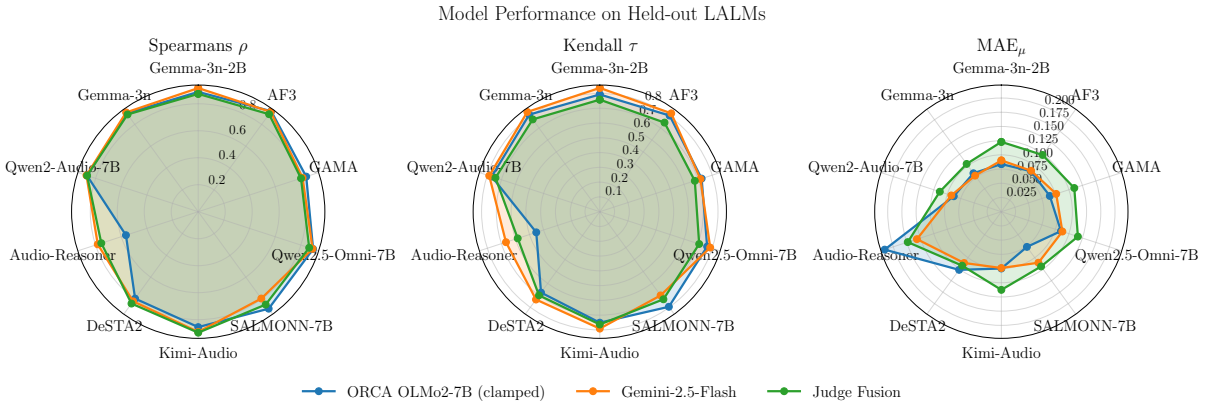

Figure 3: Comparison of LLM-judge (Gemini, offline LLM-judge fusion) to clamped ORCA OLMo-7B when trained and evaluated on the model hold-out sets.

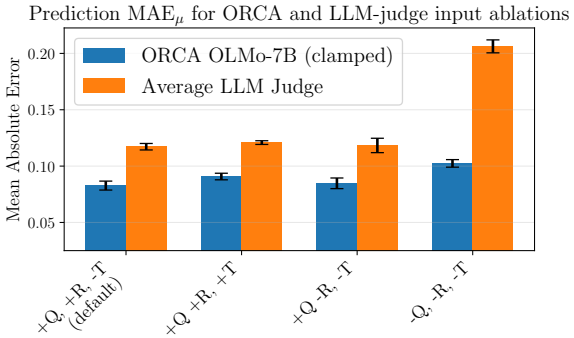

Figure 4: Ablation study on ORCA and avg. LLM-judge inputs. We report MAE of the predicted scores with respect to the average human rating. Q, R, T denote the original question, the rationale, and the transcript, respectively.

Training data	Spearman ρ	MAE$_\mu$
(1) Human data	**0.899** ± 0.010	**0.083** ± 0.004
(2) LLM-judge	0.886 ± 0.003	0.108 ± 0.006
(3) LLM-judge (all data)	0.889 ± 0.010	0.103 ± 0.014
(4) Human + LLM-judge	0.893 ± 0.008	0.100 ± 0.005
(5) Two-stage	**0.902** ± 0.008	**0.085** ± 0.011

Table 3: Performance of clamped ORCA OLMo2-7B across different human/LLM-judge training data usage scenarios. Metrics are aggregated over five unseen question-based test splits.

human-annotated instances with additional LLM-judge scores, (5) pretraining on the full LLM-judge data and fine-tuning on the human ratings. The training is again run using OLMo2-7B, and the (clamped) prediction results are aggregated on the five seeded unseen-question data splits, and the results are shown in Table 3.

This experiment highlights the importance of quality human annotations, and how the measured disagreement between LLM-judges can impact the correlation and MAE, as the noisier LLM-judge ratings bring slight performance degradation even in the fourth scenario. However, the two-stage fifth scenario shows that there is potential in leveraging the lower-agreement LLM-judge data for pretraining, showing potential for reducing the amount of human annotations required to train a reliable ORCA model.

7 Conclusions

We presented ORCA, a framework for evaluating open-ended audio QA responses that models the variability in human judgments using Beta distributions, capturing both expected correctness and annotator uncertainty. Our three-stage annotation framework simultaneously generates high-quality training data and improves the quality of the underlying benchmarks, yielding 11,721 annotations across 3,580 question-answer pairs with substantial agreement (Krippendorff's $\alpha = 0.82$), representing a comprehensive resource for studying open-ended audio QA evaluation.

ORCA models achieve 0.91 Spearman correlation with average human judgments, matching or outperforming LLM-judge baselines including Gemini-2.5-Flash, while also achieving consistently lower mean absolute rating prediction error. ORCA also only requires a single forward pass through the model, and maintains full reproducibility by leveraging open-weight models.

ORCA will be released as an installable package–including our trained models, annotation framework, and curated dataset–with ongoing support for additional audio benchmarks to facilitate further research in this area and offer a lightweight, reliable, and human judgment-aligned alternative to other audio LLM evaluation frameworks.

References

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. 2025. Benchmarking and Confidence Evaluation of LALMs For Temporal Reasoning. In *Interspeech 2025*, pages 2068–2072.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Ding Ding, Zeqian Ju, and Yichong Leng. 2025. Kimi-Audio Technical Report. Technical report, MoonshotAI.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gemma Team. 2025a. Gemma 3 Technical Report.

Gemma Team. 2025b. Gemma 3n.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S. Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. In *Proceedings of the 42 nd International Conference on Machine Learning*. PMLR.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, Miami, Florida, USA. Association for Computational Linguistics.

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. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. In *NeurIPS*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, Siddhi Patil, Soham Deshmukh, Lasha Koroshinadze, Yao Liu, Leibny Paola Garcia Perera, Eleni Zanou, Themos Stafylakis, Joon Son Chung, David Harwath, Chao Zhang, Dinesh Manocha, Alicia Lozano-Diez, Santosh Kesiraju, Sreyan Ghosh, and Ramani Duraiswami. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari,

Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Llama Team. 2024. The Llama 3 Herd of Models.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025a. DeSTA2: Developing Instruction-Following Speech Language Model Without Speech Instruction-Tuning Data. ArXiv:2409.20007 [eess].

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, Yi-Cheng Lin, Yu-Xiang Lin, Chi-An Fu, Chun-Yi Kuan, Wenze Ren, Xuanjun Chen, Wei-Ping Huang, En-Pei Hu, Tzu-Quan Lin, Yuan-Kuei Wu, Kuan-Po Huang, Hsiao-Ying Huang, Huang-Cheng Chou, Kai-Wei Chang, Cheng-Han Chiang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025b. DeSTA2.5-Audio: Toward General-Purpose Large Audio Language Model with Self-Generated Cross-Modal Alignment. ArXiv:2507.02768 [eess].

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

Grigor Nalbandyan, Rima Shahbazyan, and Evelina Bakhturina. 2025. SCORE: Systematic COnsistency and robustness evaluation for 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 3: Industry Track)*, pages 470–484, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 Technical Report.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. In *The Twelfth International Conference on Learning Representations*.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun

Ma, and Chao Zhang. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. 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)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

Wen Wu, Wenlin Chen, Chao Zhang, and Phil Woodland. 2024. Modelling variability in human annotator simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1139–1157, Bangkok, Thailand. Association for Computational Linguistics.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report.

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

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot.

A Human annotations

Annotation Guidelines

Correctness Rating Scale (1-5):

1. Not correct: The answer is irrelevant or too long.
2. Slightly correct: Contains a few relevant keywords, but is either too long or too short to be satisfactory.
3. Moderately correct: At least 50% accurate but missing key information.
4. Almost correct: Close to the reference answer, but includes unnecessary details.
5. Completely correct: The candidate and reference answers have the exact same semantic meaning, and the candidate response is short and precise.

Structured Feedback Codes:

- **Q:** Incomplete question: The options mentioned in the question were not provided.
- **A:** Had to use audio: The rationale and the description were insufficient for judgement.
- **R:** Reference answer incorrect/incomplete: The reference answer provided was flawed.
- **U:** Unable to judge (audio): Still could not judge even after listening to the audio; the question was ambiguous.
- **E:** Unable to judge (expertise): Lacked the specific expertise required to make a judgment.
- Free-form textual feedback.

The following Table 4 presents the statistics for the additional feedback we have received from the human annotators.

Feedback Code	Count	%
Data Quality Issues (unique questions, n=1,872)		
(Q) Incomplete question	275	14.7
(A) Insufficient rationale	571	30.5
(R) Incorrect reference answer	309	16.5
Annotator Limitations (tuples, n=1,993)		
(U) Unable to judge (ambiguous)	158	7.9
(E) Unable to judge (expertise)	515	25.8
Free-form feedback	184	9.2
Total feedback codes	**2,620**	
Tuples with feedback	**1,993**	

Table 4: Distribution of structured feedback codes. Top section shows number of unique questions with data quality issues (out of 1,872 annotated questions). Bottom section shows number of tuples where annotators reported difficulty (out of 1,993 tuples with feedback). Multiple codes could be assigned per tuple.

The following histogram Fig. 5 shows the distribution of annotations for each question-candidate answer pair.

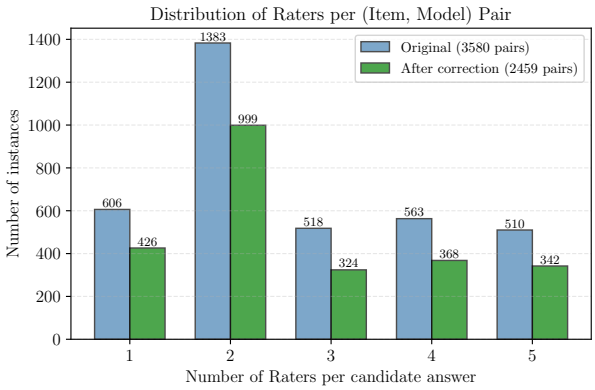

Figure 5: Histogram of human ratings for answer correctness before and after correction (Stage 3).

LALM	Num. answers	Num. human ratings	Avg. human rating	Avg. answer length
Audio Flamingo 2 (Ghosh et al., 2025)	103	275	2.13 ± 1.55	12.2
Audio Flamingo 3 (Goel et al., 2025)	197	544	2.90 ± 1.77	13.0
Audio-Reasoner (Xie et al., 2025)	211	560	2.46 ± 1.55	64.9
DeSTA2 (Lu et al., 2025a)	109	328	2.21 ± 1.62	56.4
DeSTA25-Audio (Lu et al., 2025b)	223	612	2.42 ± 1.63	47.2
GAMA (Ghosh et al., 2024)	103	180	2.25 ± 1.61	15.7
Gemma-3n-2B (Gemma Team, 2025b)	89	247	2.43 ± 1.75	5.2
Gemma-3n-4B (Gemma Team, 2025b)	230	641	2.81 ± 1.84	4.3
GLM-4-Voice (Zeng et al., 2024)	96	266	1.81 ± 1.42	23.9
Kimi-Audio (Ding et al., 2025)	236	669	2.91 ± 1.80	12.3
Qwen2-Audio-7B (Chu et al., 2024)	106	285	2.63 ± 1.82	6.1
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	217	594	2.79 ± 1.78	14.3
Qwen2.5-Omni-7B (Xu et al., 2025)	212	606	3.12 ± 1.76	22.6
SALMONN-7B (Tang et al., 2024)	104	292	2.49 ± 1.72	17.4
SALMONN-13B (Tang et al., 2024)	223	615	2.57 ± 1.78	10.9

Table 5: Summary statistics for 15 LALMs evaluated. Shows number of unique answers annotated, total human ratings collected after filtering (Stage 3), mean correctness rating (1-5 scale), and average response length in words. Models are sorted alphabetically.

B Prompts for LLMs

B.1 Prompts for Generating Rationale

Gemini-2.5-Flash generated rationales grounding answers in audio content, given the audio, question, and expected answer. Stage 3 used custom human expert prompts.

B.2 Prompts for LLM-Judges

This appendix details the prompt formulations for our LLM-Judges evaluation. We present the full prompt (Fig. 7) with supplementary information; other variants remove specific fields. Table 6 shows results across prompt conditions. Performance drops with less context, but removing only the transcript yields best results, suggesting rationale is most informative.

Additionally, we evaluated how wording and structure affect LLM performance by creating several variants using the basic prompt (reference and candidate answers) and the complete prompt (adding supplementary information and the question). Although multiple versions were tested, we include only one representative alternative (Fig. 8), modifying the instruction length, emphasis, and scoring descriptions. Table 7 shows the results for two models. In the case of Qwen2.5-7B, the base prompt shows slightly higher variability in correlation and MAE, indicating that its evaluations are sensitive to the inclusion of contextual information and phrasing. Prometheus2-7B, exhibits lower variability across prompt variants, suggesting more stable predictions regardless of prompt wording or contextual details

System Prompt:

You are an expert at analyzing audio, describing it, and writing clear explanations and rationales for answers given the question with regard to the audio.

User Prompt Template:

Listen to the provided audio carefully and provide a concise grounding rationale/audio description for the following question-answer pair. Try to write at least a few sentences without going into unnecessary detail.

Question: [question]
Answer: [answer]
Audio: [audio file URI via Gemini Files API]

Output:
A concise rationale explaining how the audio supports the given answer to the question.

Figure 6: Prompt used for generating rationales with Gemini-2.0-Flash. The model receives the audio file along with the question-answer pair.

Complete Prompt (Including rationale and transcript)

You are an expert evaluator specialized in assessing answer quality. Your task is to **compare the candidate answer to the expected answer** and **rate how well it matches**, considering the question and any supplementary information provided.

Evaluation Criteria
Key considerations:
- Your primary goal is to determine **how closely the candidate answer matches the expected answer**.
- Use the **supplementary information** (e.g., audio transcription, description, rationale) to assess whether the candidate answer is similar to the expected answer in context.
- Evaluate the candidate's understanding of the **intent of the question**, and whether the answer is **factually accurate**, **complete**, and **aligned with the expected answer**.

Scoring:
- **5 - Exact Match**: The candidate answer fully matches the expected answer in meaning and detail. It is accurate, complete, and demonstrates clear understanding.
- **4 - Close Match**: The candidate answer is mostly correct and similar in meaning to the expected answer but may omit minor details or slightly differ in phrasing.
- **3 - Partial Match**: The candidate answer captures some relevant aspects of the expected answer but lacks key details or contains noticeable inaccuracies.
- **2 - Minimal Match**: The candidate answer is only loosely related to the expected answer and shows limited understanding or relevance.
- **1 - No Match**: The candidate answer is entirely incorrect, irrelevant, or contradicts the expected answer and supplementary information.

Input:
- **Question**: [question]
- **Expected Answer**: [gt_answer]
- **Candidate Answer**: [candidate_answer]
- **Supplementary Information**: [rationale] + [trasncript]

Output:
Provide your evaluation in the following format:
Explanation: [Explain your reasoning. Compare the candidate answer directly with the expected answer, referencing supplementary information if needed to justify correctness or incorrectness.]
Score: [Numerical score from 1–5]

Figure 7: Full prompt formulation for the LLM-judges evaluation. Other variants were generated by removing selected information(transcript, full supplementary information and question).

C Additional Results

Complete Prompt (Modified example)

You are an expert LLM judge. Decide how closely the candidate answer aligns with the expected one, and provide a quality rating. Use any supplementary information to check contextual relevance, and ensure the response is accurate, complete, and reflects the question's intent.

Scoring Scale: Rate the candidate's answer from 1 to 5, where 5 indicates a complete match fully consistent in meaning, facts, and intent; and 1 indicates incorrect, irrelevant, or contradicts the expected answer or context.

Input:
- **Question**: [question]
- **Expected Answer**: [gt_answer]
- **Candidate Answer**: [candidate_answer]
- **Supplementary Information**: [rationale] + [trasncript]

Output:
Output your evaluation exactly in this format:
Explanation: [Short statement explaining your reasoning.]
Score: [1|2|3|4|5]

Figure 8: Modified full LLM-judge prompt example. Base variants omit the transcript, rationale, and the question.

Prompt Condition	Qwen2.5-7B		Llama3.1-8B		Gemma3-4B		Gemma3-12B		Prometheus2-7B		Gemini-2.5-Flash	
	ρ	MAE	ρ	MAE	ρ	MAE	ρ	MAE	ρ	MAE	ρ	MAE
With Context	**0.814**	0.139	0.775	0.152	0.756	0.186	0.832	0.140	0.692	0.198	**0.865**	0.108
– Without Transcription	**0.814**	**0.138**	**0.784**	**0.151**	**0.762**	**0.182**	**0.836**	**0.139**	**0.698**	**0.195**	0.864	**0.107**
Without Context	0.796	0.148	0.737	0.171	0.760	0.169	0.828	0.140	0.640	0.214	0.848	0.117
– Without Question	0.660	0.204	0.565	0.236	0.538	0.253	0.675	0.186	0.526	0.246	0.803	0.138

Table 6: Effects of adding context to prompts for different LLM judges. Evaluation is performed on the full dataset. Higher Spearman correlation (ρ) indicates better agreement with human ratings, while lower MAE reflects smaller prediction errors.

Prompt	Qwen2.5-7B		Prometheus2-7B	
	Correlation	MAE	Correlation	MAE
Base	0.6214 ± 0.0360	0.2188 ± 0.0141	0.5347 ± 0.0296	0.2474 ± 0.0103
Full	0.8106 ± 0.0031	0.1400 ± 0.0025	0.6988 ± 0.0283	0.1926 ± 0.0175

Table 7: LLM-judges results for Qwen2.5-7B and Prometheus2-7B over prompt variants, shown as mean ± standard deviation. Base prompt only includes reference and candidate answers, while full prompt includes context and question.