AUDITA: A New Dataset to Audit Humans or AI is Better at Audio QA

Anonymous ACL submission

Abstract

Existing audio question answering benchmarks largely emphasize sound event classification or caption-grounded queries, often enabling models to succeed through short-duration cues, lexical priors, or dataset-specific biases rather than reasoning. Thus, we present AUDITA (**A**udio **U**nderstanding from **D**iverse **I**nternet **T**rivia **A**uthors), a large-scale, real-world benchmark to rigorously evaluate audio reasoning beyond surface-level acoustic recognition. AUDITA[1] comprises carefully curated, human-authored trivia questions grounded in real-world audio, explicitly designed to introduce adversarial distractors, long-range temporal dependencies, through probing queries that cannot be answered from isolated text or sound cues alone. A human average accuracy of 32.13% shows both the challenge of the task while demonstrating meaningful comprehension of the audio. In stark contrast, state-of-the-art audio question answering models perform poorly, with average accuracy below 7.97%. Beyond raw accuracy, we apply Item Response Theory (IRT) to estimate latent proficiency, question difficulty, and expose systematic deficiencies of the models and data.

1 Introduction

Question answering (QA) has a rich history in NLP, originating from early evaluation campaigns such as TREC QA (Voorhees and Tice, 1999) and expanding well beyond its textual roots. Large-scale benchmarks like SQuAD (Rajpurkar et al., 2016) catalyzed rapid progress in text-based QA, enabling large language models to achieve near- or superhuman performance on many tasks (Chowdhery et al., 2023; OpenAI, 2023). However, this impressive success has not uniformly translated across modalities. Audio question answering (Audio QA), which requires understanding and reasoning over complex auditory inputs, remains a significant challenge despite advances in speech recognition (Baevski et al., 2020), audio tagging (Kong et al., 2020; Gong et al., 2021; Hershey et al., 2017), and multimodal modeling (Guzhov et al., 2022; Elizalde et al., 2023).

Humans naturally perform intricate auditory reasoning tasks that go beyond simple acoustic recognition. Yet, many widely used audio datasets fall short of capturing this complexity. Existing benchmarks like VGGSound (Chen et al., 2020) largely focus on closed-set sound classification, e.g., labeling sounds as "engine" or "applause", emphasizing surface-level auditory features. Captioning datasets such as Clotho (Drossos et al., 2020) prioritize enumerating audible elements, rather than posing probing questions requiring deeper reasoning. As a result, strong model performance on these datasets often reflects proficiency in tagging or leveraging language biases rather than true auditory understanding.

Moreover, many existing Audio QA datasets rely on synthetic or templated audio scenes (Fayek and Johnson, 2020; Li et al., 2022) or predominantly feature spoken language (Kim et al., 2019). Such design choices enable models to exploit linguistic regularities, sometimes producing correct answers even with degraded or absent audio input (Agrawal et al., 2016; Jabri et al., 2016). The prevalence of templated or formulaic questions increases predictability, inflating model accuracy through language priors rather than genuine auditory reasoning (Section 2). Crucially, these datasets rarely include systematic human performance baselines, making it difficult to distinguish between model limitations and dataset artifacts- a concern highlighted in recent works advocating for Item Response Theory (IRT)-based evaluations to improve benchmark reliability (Lalor et al., 2019; Cardoso et al., 2022).

In response to these shortcomings, we present AUDITA (Section 3), a large-scale benchmark composed of human-authored, probing audio trivia

[1] Code and Data will be available after blind reviews.

questions sourced from real-world domains such as trivia competitions, archival media, and podcasts. Unlike prior work, AUDITA avoids fixed sound classes, synthetic scenes, and templated question formats. Instead, it focuses on naturally occurring, challenging questions that require attentive listening and auditory reasoning, with human resolvability as a core design principle.

The questions in AUDITA, carefully crafted to probe (Rogers et al., 2023), use latent auditory competencies beyond surface recognition or fact recall. Examples include identifying a film from an orchestral theme or recognizing a speaker from a voice excerpt. These tasks demand multi-cue integration and cross-modal reasoning, spanning diverse auditory phenomena such as music, language recognition, speaker identification, and environmental sounds. Importantly, AUDITA features natural linguistic variability and adversarial question design, explicitly avoiding formulaic patterns that could inflate model performance via language biases. We asked skilled humans to answer these questions, and their reliable performance, measured as best-per-category accuracy, establishes robust human upper bounds.

Under this metric (Section 4), humans' best-per-category accuracy is 69.17% on open-ended questions and 86.67% on multiple-choice questions, with near-perfect performance in categories like person identification and musical elements. In stark contrast, state-of-the-art audio and multimodal models perform near chance, with accuracies below 7.97%, demonstrating a substantial gap in auditory reasoning capabilities.

IRT analysis (Hambleton et al., 1991) further elucidates this gap (Section 5), revealing a clear separation between human and model ability, underscoring that Audio QA remains far from superhuman. This human-grounded psychometric evaluation also helps identify ambiguous or underspecified items. In contrast, models outperform humans primarily on narrow tasks involving isolated command recognition or highly structured synthetic sounds-regimes that poorly represent the demands of real-world Audio QA.

Our contributions are threefold: (1) **a novel benchmark** that stresses open-ended, multi-cue auditory reasoning over longer, diverse clips including music, speech, and environmental sounds; (2) **human performance baselines** revealing a wide and consistent gap between expert listeners and current models; and (3) **psychometric evaluation** using IRT to provide detailed insights into question difficulty, human-model agreement, and residual model weaknesses.

2 Why Audio QA Requires Better Questions

Recent advances in text-based question answering (QA) and large language model training have underscored that dataset quality is as important as model architecture. Poorly designed questions introduce ambiguity (Min et al., 2020), false presuppositions (Yu et al., 2023), unreliable supervision, and exploitable shortcuts, resulting in inflated model performance and unstable evaluations. Consequently, text QA research has increasingly emphasized careful question design, human calibration, and adversarial evaluation to ensure benchmarks truly test reasoning rather than dataset artifacts.

Many of these failure modes identified in text QA also appear in audio question answering (Audio QA) datasets. We uncover these issues specifically by applying Item Response Theory (IRT), which analyzes question difficulty (b) and discrimination (a) parameters to reveal how well questions differentiate between high- and low-ability models or annotators.

Table 7 and 8 summarizes common issues across several popular Audio QA datasets, illustrating these problems with concrete examples and IRT-based evidence. Ambiguity, a central challenge extensively studied by (Min et al., 2020), occurs when questions admit multiple plausible answers or interpretations. For example, in Clotho-Audio QA (Drossos et al., 2020), questions like "What animal is making the sound?" receive different yet valid responses such as "bird" or "dog", as both are in the corresponding clip reflecting overlapping audio sources or vague wording. This ambiguity results in low discrimination scores (e.g., $a = 0.35$ in VGGSound QA, Table 7, row 1) and low feasibility.

False presuppositions, well-documented in text QA (Yu et al., 2023), appear in Audio QA datasets such as FSD50K (Fonseca et al., 2022), where questions assume the presence of pattern in the sound clip (e.g., "What is the most notable thing about the pattern of the sound event that we hear?") that may not occur in the audio clip. This misalignment confuses annotators and models alike, leading to unreliable supervision (Table 8).

IRT diagnostics further highlight weak grounding and shortcut learning, particularly in caption-derived datasets like AudioCaps QA (Kim et al., 2019) and AudioSet QA (Gemmeke et al., 2017). Questions such as "Is someone laughing?" or "Is there music playing?" often exhibit low discrimination ($a \approx 0.28$) and can be answered correctly without audio by exploiting metadata or captions (Table 7, rows 3 and 4). These shortcuts undermine genuine auditory reasoning.

Underspecified answer keys also plague datasets; for instance, multiple lexical variants (e.g., "dog", "puppy", "hound") are treated as distinct answers in MUSIC-21 QA (Christodoulou et al., 2025) and Clotho-Audio QA, increasing noise and evaluation difficulty.

Lastly, overlapping audio sources common in real-world recordings exacerbate ambiguity, as seen in VGGSound QA and AudioCaps QA (Table 8, row 1). When multiple sounds occur simultaneously, it becomes difficult to isolate and identify the target audio event, complicating annotation and reducing question clarity.

Addressing these issues requires the design of human-authored, probing questions with clear, verifiable, audio-anchored answers to enable robust evaluation of auditory reasoning beyond superficial pattern recognition.

3 Audio Questions for Triva Experts

We focus on a benchmark of human-authored audio questions designed by knowledgeable experts to test auditory and multimodal reasoning skills. This approach aligns with what Rogers et al. (2023) call the "probing" paradigm and Rodriguez and Boyd-Graber (2021) term the Manchester paradigm, where questions are crafted explicitly to evaluate human-level understanding. While probing-style questions are common in text-only QA datasets such as TriviaQA (Joshi et al., 2017), they remain underexplored in the audio domain.

We scrape questions from publicly available online audio tests, reflecting a broader practice across various communities- such as educators, trivia enthusiasts, and researchers- that release question sets to facilitate human study and benchmarking. This openness aligns with the Manchester paradigm's emphasis on carefully crafted questions designed to probe deep understanding.

However, unlike text-based questions, audio question answering (Audio QA) requires additional effort to segment continuous audio content into discrete, independently answerable questions that can be posed one at a time to either humans or machines. For instance, many examples have text instructions that apply to all of the following examples: e.g., "how many wheels to each of the following vehicles have", followed by the sound that each of the vehicles makes.

Total Questions	Unique Audio Clips	Avg. Ques. Length (words)	Avg. Audio Duration (s)	Audio Duration Range (s)
9,690	8,713	12.47	36.98	0.42 – 478.33

Table 1: Summary statistics describing the scale and structure of the AUDITA Benchmark Dataset.

Another difficulty is that many of the questions are pyramidal. Pyramidal questions are structured so that clues are presented sequentially, from broad to specific, allowing participants to answer as soon as they recognize the target answer. This format is well established in text-based trivia, with QuizBowl serving as a canonical example (Rodriguez et al., 2019).

While text-based pyramidal questions consist of clearly delineated sentences, audio pyramidal questions introduce additional complexity. They require segmenting continuous audio streams into discrete, ordered clues aligned with question prompts, posing unique challenges for both human and machine comprehension.

In total, AUDITA contains 9690 audio-question pairs, providing a valuable resource for studying audio question answering. Tables 1 report the dataset's scale and structure.

Audio Pyramidal Trivia Two publicly available repositories (Rodriguez et al., 2019) collect audio-based pyramidal trivia questions originally created for competitive human play. While inspired by the QuizBowl tradition, these datasets differ from standard text QuizBowl corpora in that each question is composed of multiple audio clips rather than a single textual prompt. The clips are ordered to provide progressively more distinctive evidence, mirroring the pyramidal structure used in text QuizBowl. For example, an artist-identification question may begin with obscure recordings and end with a widely recognized song. Most collections focus on music, while the SoundTrack set emphasizes film and television audio.

Quizmasters Website The Quizmasters website (Quizmasters, 2025) hosts curated collections of short audio clips organized by auditory skill

Statistic	Pavements	Audio-Packets	Quizmasters	External Sources
Questions (%)	6.94%	17.02%	42.70%	33.33%
Avg. Audio Duration (s)	63.42	65.25	41.81	13.75
Avg. Ques. Length	6.56	4.68	14.54	12.04

Table 2: Distribution of questions in the AUDITA Benchmark Dataset by source, with audio duration and question length statistics.

category, originally created for pub quiz-style challenges, but without accompanying questions. These collections are designed to probe specific listening abilities, such as recognizing transformed audio (e.g., reversed or filtered signals) or identifying less common musical material. We convert these collections into an audio question answering format by attaching human-authored questions aligned with each category's intended challenge. In addition to question–answer pairs, we retain clip-level metadata such as duration and sampling rate.

PAVEMENT The PAVEMENT dataset (Pavao, 2025) consists of audio-based pyramidal trivia questions written for human competition. Similar to text QuizBowl, questions are structured as sequences of increasingly revealing clues, but are realized through audio rather than text. Many questions include annotations specifying acceptable answers or clarifications, which we preserve. For evaluation, we segment each pyramidal question into individual audio–question pairs corresponding to its constituent clues.

To ensure correctness and consistency, we perform dataset cleaning and normalization, including verifying audio–question alignment and standardizing formatting. Details of the scraping, cleaning, and normalization procedures are provided in Appendix A. Table 2 presents the distribution of questions by source.

Data Preparation

AUDITA is prepared in three stages: extraction and alignment, normalization, and categorization to ensure consistent, human-readable audio–question–answer triples suitable for evaluation. We extract and align question prompts and answers with audio clips from diverse sources, correcting indexing where needed. Normalization cleans encoding artifacts, standardizes formatting (e.g., QuizBowl markup removal), and uses GPT-4o-mini for context-aware answer rewriting when necessary (e.g., raw text: *_Ž01cPathŏ0e9tiqueŽ01d Sonata_ [or _BeethovenŽ019s Piano Sonata No. 8_ in C minor, _Op. 13_ (accept any underlined part)* → *Pathetique Sonata or Beethoven's Piano Sonata No. 8 in C minor or Sonata No. 8 in C minor*). Further details on these procedures are provided in the Appendix A.

Categorization and consolidation. Because our dataset- and audio question answering benchmarks more broadly- span a wide range of auditory phenomena and reasoning skills, we organize questions into semantic categories. This categorization serves two complementary purposes. First, it enables fine-grained analysis of model performance, allowing us to identify domain-specific strengths and weaknesses (e.g., music recognition versus environmental sound reasoning). Second, because our human-centered QA task is designed for skilled participants, category information allows humans to select questions aligned with their expertise, supporting more effective allocation of human effort.

We categorize every question. Each item is initially assigned to one of six high-level categories with 26 total subcategories using GPT-4o-mini. For evaluation and presentation, we then collapse this hierarchy into six clearly defined, human-interpretable categories that are exposed to participants and used consistently throughout the paper: *Cultural Geography in Sound*, *Name The Music: Songs, Artists & Composers*, *Who's Who? Name That Persona*, *Elements of Musical Works*, *Pop Culture and Media*, and *Environmental and Acoustic Sound Recognition*. This mapping is one-to-one at the main-category level and is applied uniformly across all data sources (Table 3).

MCQ generation We create MCQ variants with one correct answer and three human-authored distractors that are acoustically and semantically plausible, sharing surface cues while differing in decisive auditory evidence. This limits shortcut strategies and ensures MCQ accuracy reflects audio reasoning rather than text bias.

Positioning relative to prior Audio QA benchmarks Prior Audio QA resources typically emphasize: (i) closed-set recognition and event labeling, (ii) caption- or metadata-derived QA, (iii) synthetic or templated audio scenes, and (iv) speech-centric QA reducing to ASR plus text QA. MMAU summarizes this landscape as spanning information extraction and reasoning tasks, noting systems remain far behind humans on a human-evaluated split (Sakshi et al., 2024). AUDITA targets a miss-

Statistic	Who's Who? Name That Persona	Cultural Geography in Sound	Name The Music: Songs, Artists & Composers	Pop Culture and Media	Elements of Musical Works	Environmental and Acoustic Sound Recognition
% Questions	7.66%	4.56%	26.60%	19.13%	7.80%	34.24%
Avg. Audio Duration (s)	21.25	64.14	41.63	68.95	58.70	10.43
Avg. Question Length (words)	13.41	8.21	9.86	13.94	8.56	25.27

Table 3: Breakdown of questions in the AUDITA Benchmark Dataset by category, with average audio durations and question lengths.

ing regime: *probing audio trivia* requiring *audio-to-referent grounding* (linking audio to specific real-world referents like songs or films) and *long-range cue integration*, beyond short acoustic cues or caption-level descriptions.

3.1 Dataset Composition

AUDITA contains 9690 questions: 6460 human-authored questions designed for human evaluation and 3230 questions from external benchmarks. The human-authored portion consists of 2322 pyramidal-style questions (673 from *Pavements* and 1649 from *Audio-Packets*) and 4138 trivia-style questions from *Quizmasters*. All human-authored questions are closed-ended with discrete, verifiable answers.

The external portion comprises 2907 questions from OpenAudio QA (90%) and 323 questions from ClothoAudio QA (10%) (Gong et al., 2023b; Lipping et al., 2022). By question type, the external set contains 1,205 closed-ended questions (882 from OpenAudio QA and 323 from ClothoAQA) and 2025 open-ended questions (all from OpenAQA) that require semantic evaluation rather than exact string matching.

Across the dataset, human evaluation collected 1517 human guesses-individual answer attempts by participants- providing a reliable set of judgments to benchmark model performance. In the next section, we describe our evaluation framework, which leverages these human guesses to jointly model question properties and participant abilities.

4 How hard is AUDITA for Humans and Computers

We evaluate the proposed dataset using both state-of-the-art audio-language models and human participants in order to characterize its difficulty, adversarial nature, and suitability for evaluating genuine auditory reasoning. We collect both free-form text and multiple-choice (MCQ) responses from humans and models to better understand question ambiguity, scale difficulty, and keep the task engaging. Free-form answers reveal the variety in how people interpret questions- for example, synonyms or partial answers-which helps identify ambiguous or tricky items. MCQs, by contrast, simplify the task and make it easier to scale evaluations, while allowing us to test how well participants can choose the correct answer among plausible distractors. Using both formats lets us capture richer data and more precisely measure reasoning abilities. Both models and humans are evaluated under the same input and evaluation conditions.

4.1 Models

We evaluate 16 open-source audio–multimodal models with mid-scale language backbones (approximately 4B–13B parameters), grouped into three capability classes. **Omnimodal models** (6) support unified understanding across text, audio, vision, and video, with both text and speech generation. **Audio–language models** (4) focus on audio understanding with text-only outputs, while **speech-capable models** (6) emphasize speech recognition and generation, including both speech-first and modular architectures.

This taxonomy reflects the diversity of AUDITA questions: some require speech or lyric understanding, others test purely acoustic reasoning over music or environmental sounds, and some benefit from broader multimodal context. Evaluating across capability groups enables analysis of whether failures are systematic or capability-specific, analogous to how humans identify songs via lyrics or melody. All models are evaluated using publicly released checkpoints with recommended inference settings and no task-specific fine-tuning. Full model details are provided in Appendix B.

4.2 Answer Formats

We evaluate two answer settings:

Free-Response Question Answering. In the free-response setting, models generate open-ended textual answers. Generated responses are evaluated using the PEDANT (Li et al., 2024) framework, which determines semantic equivalence between

model outputs and reference answers through structured normalization and equivalence rules, rather than relying on exact string matching.

Multiple-Choice Question Answering (MCQ). In the MCQ setting, models are given a fixed set of candidate answers and must select the correct option. We convert model outputs to a choice via either explicit option selection or scoring each candidate independently, depending on model capabilities. Accuracy is reported as the fraction of correctly answered questions.

4.3 Human Evaluation

To contextualize model performance, we also evaluate human accuracy. Human participants listen to the same audio clips and answer the same questions under controlled conditions. Participants are not given access to transcripts or external information and are instructed to rely solely on the audio content.

Human responses are collected for both the multiple-choice and free-response settings. Accuracy is computed using the same evaluation criteria applied to models for both MCQ and free-response.

To benchmark against prior work, we include minimally processed questions from OpenAQA (Gong et al., 2023b) and ClothoAQA (Lipping et al., 2022), applying filtering to remove unanswerable items.

4.4 Input Representation

To fairly evaluate both humans and computational models on our audio question answering task, we ensure consistent presentation of audio and questions. Each example consists of an audio clip paired with a natural-language question, presented verbatim without templating or normalization. For human participants, audio clips are provided in a standardized playback format.

For computational models, audio inputs are uniformly preprocessed: audio is converted to mono and resampled to the sample rate expected by each model. Depending on the model's input requirements, we supply either raw waveform audio or alternative audio representations such as log-mel spectrograms or codec tokens, following each model's prescribed preprocessing pipeline. This consistency supports fair comparison across diverse model architectures and implementations.

5 Empirical Analysis of AUDITA

We analyze human and model responses to the fixed audio–question pairs of AUDITA following the protocol in Section 4. While we report aggregate accuracy for comparability with prior work, accuracy alone is known to obscure important structure in QA benchmarks.

Prior psychometric work shows that questions vary substantially in difficulty and informativeness, and that some items contribute little to evaluating reasoning ability despite equal weighting in accuracy (Lalor et al., 2016). Related studies demonstrate that apparent gains on leaderboards may be driven by easy or artifact-laden items rather than genuine improvements in reasoning or perception (Rodriguez and Boyd-Graber, 2021; Rodriguez et al., 2021). Motivated by these findings, we apply Item Response Theory (IRT) to characterize item difficulty and responder ability in Audio QA.

Table 4 reports accuracy alongside IRT ability estimates for humans and models. Humans substantially outperform all models under both metrics. System rankings are broadly consistent across accuracy and ability.

To localize this gap, Figure 1 presents category-level accuracy alongside mean item difficulty. Categories involving music recognition and environmental sound inference combine high difficulty with low model accuracy, while humans remain reliable. This alignment suggests that model failures arise from intrinsic auditory challenges rather than annotation noise or scoring artifacts.

Accuracy alone cannot distinguish uniformly difficult questions from systematic reasoning failures. Figure 2 shows the IRT ability distributions for humans and models on a shared scale. Models cluster in a narrower, substantially lower ability range than humans, revealing limited robustness as question difficulty increases.

We analyze item-level performance in Figure 3, showing both humans and models do well on easy items, but the gap widens as difficulty rises, with humans outperforming models on hard questions. Questions tend to be easier when they contain clear, distinctive audio cues, familiar sounds, or require minimal temporal integration. Such questions have lower IRT difficulty (b) values, reflecting their relative simplicity for both humans and models. Table 5 compares performance across prior Audio QA datasets and AUDITA. While models achieve some-

System	Accuracy (%)(Text)	Ability θ (Text)	Ability SD (Text)	MCQ Accuracy (%)	Ability θ (MCQ)	Ability SD σ(MCQ)
Humans	32.13	0.05	0.26	60.16	0.08	0.25
Models	7.97	−2.91	0.55	12.98	-2.45	0.54
Model	Text Accuracy (%)	Text Ability (θ)	Text Rank	MCQ Accuracy (%)	MCQ Ability (θ)	MCQ Rank
Qwen2.5-Omni (Xu et al., 2025a)	10.01	-2.16	1	21.021	-1.76	2
AudioGPT (Huang et al., 2024)	8.98	-2.31	2	23.49	-1.61	1
OpenOmni (Luo et al., 2025)	7.99	-2.45	3	16.01	-2.06	4
Audio-Flamingo (Kong et al., 2024)	7.77	-2.49	4	13.99	-2.11	6
Phi-4-Multimodal (Abouelenin et al., 2025)	7.49	-2.54	5	15.73	-2.09	5
Qwen3-Omni (Xu et al., 2025b)	6.87	-2.62	6	18.89	-2.04	3
LTU-AS (Gong et al., 2023a)	6.69	-2.62	7	13.81	-2.11	7
Qwen-2 Audio (Chu et al., 2024)	6.53	-2.70	8	13.72	-2.31	8
Baichuan-Omni-1.5 (Li et al., 2025)	6.49	-2.71	9	13.63	-2.43	9
VITA-1.5 (Fu et al., 2025)	5.76	-2.81	10	12.59	-2.71	10
Mini-Omni2 (Xie and Wu, 2024)	4.85	-3.03	11	8.74	-2.78	12
SpeechGPT (Zhang et al., 2023)	3.99	-3.28	12	12.38	-2.71	11
SALMONN-2 (Tang et al., 2025)	2.77	-3.60	13	5.68	-3.31	15
SALMONN-2+ (Tang et al., 2025)	2.62	-3.61	14	6.28	-2.89	14
MU-LLaMA (Liu et al., 2024)	2.19	-3.61	15	7.48	-2.81	13
SALMONN (Tang et al., 2024)	1.99	-4.00	16	4.18	-3.40	16

Table 4: This table presents accuracy and IRT-based ability estimates with standard deviations for humans and models on text and MCQ audio question answering tasks. Accuracy measures raw correctness, while IRT ability (θ) quantifies latent skill considering item difficulty. The data highlights the gap between human and model performance, and variability in skill across systems.

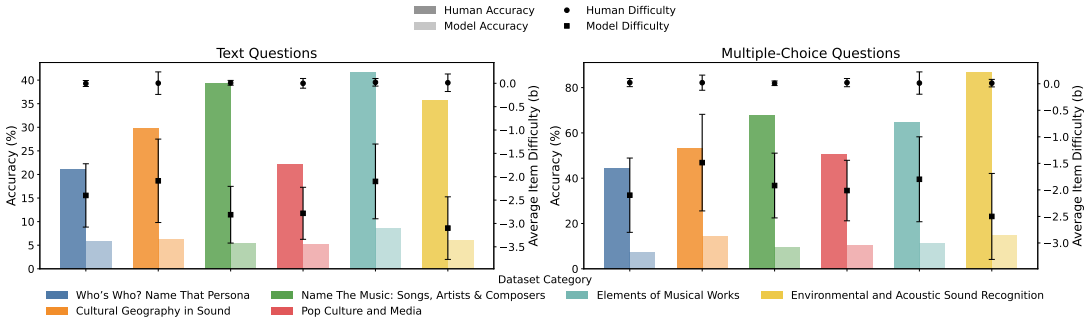

Figure 1: Category-level accuracy and average item difficulty for humans and models on text and multiple-choice questions. Bars show accuracy; points show mean IRT difficulty ± one standard deviation. Higher difficulty correlates with lower accuracy for both, but models' performance declines more sharply, revealing systematic category-specific gaps.

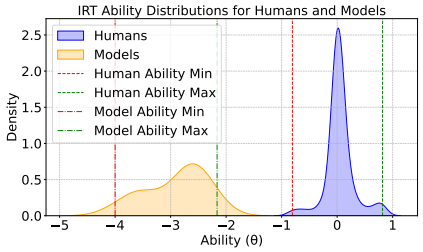

Figure 2: Distributions of IRT ability parameters (θ) for humans (blue) and models (orange) are shown on a shared scale. Kernel density estimates highlight that humans cluster at higher θ, with dashed lines marking each group's ability range. This reflects humans' superior performance and a clear latent ability gap between humans and models despite model variability.

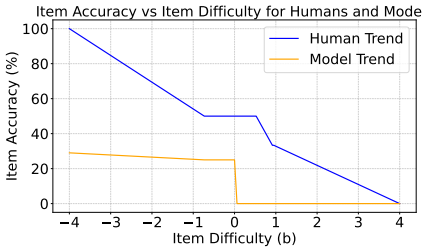

Figure 3: Item accuracy plotted against item difficulty (b) for humans (blue) and models (orange). Humans generally show higher accuracy than models across difficulties; however, some human accuracy values drop to zero due to sparsity of responses on certain items. The trend lines, highlight the widening performance gap between humans and models.

what higher accuracy on some earlier benchmarks (like External Datasets), their estimated abilities (θ) cluster tightly with lower variability, suggesting those datasets contain fewer difficult questions that truly test reasoning. In contrast, AUDITA presents a wider range of question difficulties and larger gaps between human and model performance, highlighting more challenging items that better expose model weaknesses. Table 9 presents top discriminator questions that best separate human and model performance. Together, accuracy and IRT reveal that current Audio QA models lag significantly be-

Text Questions				
Dataset	Modality	Acc (%)	Mean θ	SD σ
Pavements	Human	35.28	0.09	1.74
Pavements	Model	4.26	-2.81	0.60
Audio-Packets	Human	34.24	0.08	0.61
Audio-Packets	Model	8.89	-2.03	0.67
Quizmasters	Human	21.34	0.03	0.91
Quizmasters	Model	1.58	-3.88	0.62
External Datasets	Human	25.79	0.05	0.51
External Datasets	Model	13.49	-1.45	0.51
MCQ Questions				
Dataset	Modality	Acc (%)	Mean θ	SD σ
Pavements	Human	53.90	0.11	1.75
Pavements	Model	5.89	-2.33	0.61
Audio-Packets	Human	61.21	0.13	0.62
Audio-Packets	Model	11.76	-1.83	0.65
Quizmasters	Human	50.31	0.10	0.93
Quizmasters	Model	3.04	-2.70	0.62
External Datasets	Human	74.30	0.15	0.49
External Datasets	Model	20.59	-1.11	0.50

Table 5: Comparison of accuracy and mean ability (θ) across datasets, modalities (human vs. model), and question types (Text vs. MCQ) shows humans consistently outperform models. Pavements and Audio-Packets yield moderate human accuracy (≈30–60%) with positive abilities, while models score lower with negative abilities. Quizmasters is more challenging, especially for models. External datasets show the highest human accuracy, notably on MCQs. These results reveal clear human-model gaps and varying task difficulty by dataset and question type.

hind humans on difficult, real-world tasks—exactly the challenge this benchmark targets.

Further analysis reveals variation across question categories. Environmental sounds and complex music are difficult for models, though humans perform well, showing models struggle with nuanced audio. Models often confuse similar clips or miss long-term cues. They do better on multiple-choice questions but gaps remain. IRT shows models fail more on hardest items, exposing limits in audio reasoning. Human responses are consistent, validating the benchmark, while models show higher variance, indicating room for improvement.

Interpreting human accuracy AUDITA consists of open-ended audio trivia with a very large effective answer space, making chance performance in free response effectively zero. In multiple-choice, chance accuracy is $1/K$ (e.g., 25% for $K = 4$), yet humans substantially exceed this baseline under identical conditions. Open-answer scoring is intentionally strict—many items admit plausible near-misses (e.g., confusing franchise installments or covers vs. originals)—so even moderate free-response accuracy reflects meaningful auditory reasoning rather than guessing, consistent with human baselines reported by MMAU (Sakshi et al., 2024).

6 Related Work

Audio Question Answering (AQA) is an emerging field with relatively few datasets, many of which impose constraints limiting evaluation of genuine auditory reasoning. Early datasets like CLEAR (Lin et al., 2021) and DAQA (Fayek and Johnson, 2020) use synthetic audio or fixed vocabularies with templated questions, restricting linguistic diversity and encouraging shortcut learning. More naturalistic datasets such as ClothoAQA (Drossos et al., 2020) and MusicAVQA (Li et al., 2022) still rely heavily on caption-derived or formulaic questions, allowing models to exploit text priors and metadata instead of true audio understanding.

Existing benchmarks rarely provide human performance baselines, complicating assessment of difficulty and model shortcomings (Lalor et al., 2019). Recent studies demonstrate that state-of-the-art audio–language models remain vulnerable to misleading textual cues and perform poorly on open-ended, information-seeking questions requiring multi-cue integration and real-world knowledge (Wang et al., 2025a). Unlike textual and visual QA, adversarial human-authored Audio QA datasets are scarce. Appendix C provides detailed discussion and examples of limitations in existing Audio QA datasets and where models show superhuman performance on narrow audio tasks (Table 12). AUDITA fills this gap by providing a large-scale, human-authored, adversarial benchmark with naturally occurring trivia questions that require genuine auditory reasoning beyond surface recognition or caption matching.

7 Conclusion

We present AUDITA, a benchmark of human-authored, probing audio questions from real-world sources that captures the complexity of genuine auditory reasoning. Designed for human solvability, it provides a clear baseline for evaluating model understanding beyond synthetic or templated tasks. Our evaluations show a clear gap between humans and models, revealing limits in processing acoustic cues, context, and time. IRT analysis highlights key differences in reasoning. AUDITA helps pinpoint model weaknesses and dataset challenges, guiding future work toward better audio understanding.

8 Limitations

Our evaluation targets out-of-the-box performance of locally runnable, open-checkpoint models in a mid-scale regime, and we do not claim to cover the full space of proprietary or cloud-only systems. Larger closed models may improve absolute accuracy, but the scale question is discussed directly in Appendix D.1, including why scale alone is unlikely to close the gap we observe on a difficult, human-authored benchmark. We also do not run full scaling sweeps across cloud-scale systems because evaluating 9,690 items with audio inputs would impose substantial cost.

We intentionally evaluate end-to-end model behavior without external tool augmentation. In particular, we do not benchmark pipelines that add ASR plus retrieval, music fingerprinting, database lookups, or web search. These systems are relevant in practice and may reduce errors on referent-linking items, but they measure a different capability than the audio-grounded reasoning we aim to isolate. Relatedly, we evaluate all systems in an audio+question to text-answer setting for consistency, even when a model can generate speech, which may understate the value of speech-first interaction designs in real deployments.

Our psychometric analysis depends on the breadth and quality of human responses. While IRT helps separate item difficulty from responder ability, it can still be affected by annotator variability and by items where the audio is noisy, overlapping, or underspecified. Human participants are also not uniformly "trivia experts," so aggregate human accuracy should be interpreted as a baseline rather than a ceiling. Finally, although AUDITA is deliberately sourced from real-world domains, it is still shaped by the distribution of publicly available audio trivia and by English-centric question writing, which may limit generalization to other languages, accents, and niche audio domains.

AUDITA contains short audio clips (averaging 37 seconds) from publicly available trivia sources, following practices established by multimedia QA benchmarks such as TVQA (Lei et al., 2018). We will release stable metadata and acquisition scripts to support reproducibility.

9 Ethical Considerations

Human evaluation was conducted under Institutional Review Board (IRB) approval and informed consent. We collected participant email addresses solely to deliver remuneration. Emails are stored securely, are not used for analysis, and are not linked to response data in the released benchmark or reported results. Aside from compensation logistics, we do not collect additional personally identifying information, and we analyze results in anonymized form.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, and Dongdong Chen and. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *CoRR*, abs/2503.01743.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1955–1960. The Association for Computational Linguistics.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. Computer Vision Foundation / IEEE Computer Society.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Lucas F. F. Cardoso, José de S. Ribeiro, Vitor Cirilo Araujo Santos, Raíssa Lorena Silva da Silva, Marcelle Pereira Mota, Ricardo B. C. Prudêncio, and Ronnie C. O. Alves. 2022. Explanation-by-example based on item response theory. In *Intelligent Systems*

- *11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 - December 1, 2022, Proceedings, Part I*, volume 13653 of *Lecture Notes in Computer Science*, pages 283–297. Springer.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 721–725. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Anna-Maria Christodoulou, Kyrre Glette, Olivier Lartillot, and Alexander Refsum Jensenius. 2025. Musiqal: A dataset for music question-answering through audio-video fusion. *Trans. Int. Soc. Music. Inf. Retr.*, 8(1).

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. *CoRR*, abs/2407.10759.

Santiago Cuervo and Ricard Marxer. 2024. Scaling properties of speech language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 351–361. Association for Computational Linguistics.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 736–740. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Haytham M. Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2283–2294.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: an open dataset of human-labeled sound events. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:829–852.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. VITA-1.5: towards gpt-4o level real-time vision and speech interaction. *CoRR*, abs/2501.01957.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: audio spectrogram transformer. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 571–575. ISCA.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James R. Glass. 2023a. Joint audio and speech understanding. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 976–980. IEEE.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory*, volume 2. Sage.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 131–135. IEEE.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications*

of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 23802–23804. AAAI Press.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 727–739. Springer.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132. Association for Computational Linguistics.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2880–2894.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 648–657. The Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4248–4258. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: localized, compositional video question answering. *CoRR*, abs/1809.01696.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19086–19096. IEEE.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. Adversarial VQA: A new benchmark for evaluating the robustness of VQA models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2022–2031. IEEE.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, and Lingfeng Ming and. 2025. Baichuan-omni-1.5 technical report. *CoRR*, abs/2501.15368.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan L. Boyd-Graber. 2024. PEDANTS: cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 9373–9398. Association for Computational Linguistics.

Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. 2021. The CLEAR benchmark: Continual learning on real-world imagery. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. *arXiv preprint arXiv:2204.09634*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 286–290. IEEE.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Yangyi Chen, Hamid Alinejad-Rokny, and Fei Huang. 2025. Openomni: Large language models pivot zero-shot omnimodal alignment across language with real-time self-aware emotional speech synthesis. *CoRR*, abs/2501.04561.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5783–5797. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Victor Pavao. 2025. Pavement-Music-Tournaments. *Github*. Online; Available at: `https://github.com/vcpavao/Pavement-Music-Tournaments/`.

Quizmasters. 2025. The Quizmasters Audio Database. Online; Available at: `https://www.quizmasters.biz/DB/Audio/AMenu.html`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan L. Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4486–4503. Association for Computational Linguistics.

Pedro Rodriguez and Jordan L. Boyd-Graber. 2021. Evaluation paradigms in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9630–9642. Association for Computational Linguistics.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *CoRR*, abs/1904.04792.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10):197:1–197:45.

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

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20346–20359.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn 2: Captioning-enhanced audio-visual large language models. *CoRR*, abs/2506.15220.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ellen M. Voorhees and Dawn M. Tice. 1999. The TREC-8 question answering track evaluation. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025a. When audio and text disagree: Revealing text bias in large audio-language models. *CoRR*, abs/2508.15407.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025b. When audio and text disagree: Revealing text bias in large audio-language models. *Preprint*, arXiv:2508.15407.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190.

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. 2025a. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

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. 2025b. Qwen3-omni technical report. *CoRR*, abs/2509.17765.

Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, Sakshi Singh, Vaibhavi Lokegaonkar, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, Gunhee Kim, Jun Du, Rafael Valle, and Bryan Catanzaro. 2025. Multi-domain audio question answering toward acoustic content reasoning in the DCASE 2025 challenge. *CoRR*, abs/2505.07365.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. AVQA: A dataset for audio-visual question answering on videos. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3480–3491. ACM.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10457–10480. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773. Association for Computational Linguistics.

A Data Collection and Processing Details

Quizmasters Website The Quizmasters website[2] publishes standalone audio clips grouped under categorical umbrellas, but without associated questions. These clips range from 3 to 40 seconds in length and are organized into categories that either apply audio transformations (e.g., reversal) or present challenging identification tasks involving less popular material. Since clips within a category share common properties, we assign handwritten questions appropriate to the category format. For example, a clip from the National Anthems category may be paired with the question "What country is this national anthem from?" Some collections also include a corresponding reveal clip containing the original, untransformed audio; because these clips do not exhibit the intended transformation, we instead assign identification-style questions such as "Name the artist who recorded this song." In addition to question–answer pairs, we store clip-level metadata including sampling rate, duration, and other audio attributes.

PAVEMENT Processing Each dataset entry includes a *notes* field that preserves additional information from the original QuizBowl questions, such as alternative acceptable answers or clarifying remarks. These annotations are retained as part of the source material. In the PAVEMENT subset, QuizBowl questions are split into individual clues for model evaluation.

Some questions refer to shared attributes across clues, such as "What is the common number in the titles of these songs?" or "What is the common profession mentioned in the titles of these songs?" While some tournaments, such as SoundTrack, provide clip descriptions in their GitHub repositories, this practice is inconsistent and absent in PAVEMENT.

Our processing of PAVEMENT focused on data integrity. Earlier scraped versions contained mismatches between audio clips and questions, ordering errors, and formatting issues. We therefore re-scraped the dataset to ensure correct question–answer alignment and applied normalization steps to make outputs human- and model-ready, including resolving Unicode issues, inconsistent notation, and formatting irregularities.

Data Preparation

We prepare AUDITA in three stages: (i) extraction and alignment, (ii) normalization for evaluation, and (iii) categorization for analysis. The goal is not to change the underlying questions, but to make the resulting audio–question–answer triples consistent, human-readable, and robust to evaluation artifacts.

[2] https://www.thequizmasters.biz/

Extraction and alignment. For GitHub-hosted QuizBowl-style sources, we extract question prompts and answerlines from the provided materials and align them with audio files using source-specific directory conventions and indexing. For *Pavements*, we correct earlier indexing mismatches by enforcing consistent clip identifiers and alignment rules between prompts, answerlines, and audio files. Table 3 reports the resulting category distribution.

External benchmark preparation. To provide a concrete comparison point to prior audio question answering resources, we include questions from OpenAQA and ClothoAQA. OpenAQA is largely generated from captions and metadata as part of LTU's OpenAQA-5M pipeline, while ClothoAQA is crowdsourced, which offers a small human-written counterpoint to caption-derived questions (Gong et al., 2023b; Lipping et al., 2022). We do not rewrite the benchmark questions or answers to make them more human-friendly. Instead, we apply only minimal preprocessing needed for evaluation consistency.

For OpenAQA, we run the filtering utilities released with LTU to remove unanswerable or hallucinated question–answer pairs (Gong et al., 2023b). In our snapshot, this removes 18.65% of candidate items. We then sample proportionally across OpenAQA's constituent source datasets to preserve its original mixture.

Normalization for evaluation. Raw answerlines and prompts contain encoding noise and formatting conventions that can create spurious evaluation failures for both humans and models. We apply multi-stage cleaning that targets: *(1) Character normalization:* mapping diacritics and special symbols to ASCII equivalents and removing Unicode artifacts introduced by heterogeneous source encodings. *(2) Formatting normalization:* stripping QuizBowl markup, removing prompt instructions and bracketed editorial artifacts, and rewriting alternative acceptable answers into a consistent "A or B" form. *(3) Context-aware cleanup:* for cases where rule-based edits are insufficient, we use GPT-4o-mini to rewrite answers into a standardized, human-readable form while preserving semantic content and listed alternatives. When cleanup is uncertain, we prefer conservative edits that preserve the original label over aggressive rewriting.

B Evaluated Models: Architectures, Training Objectives, and Assumptions

We evaluate a diverse set of state-of-the-art open checkpoint models runnable locally, covering mid-scale language backbones (approximately 6B to 17B parameters in the language component) and a variety of audio front ends. These include models trained primarily on audio–text alignment objectives, large multimodal foundation models with audio inputs, and audio-specialized models adapted for question answering. To address the varied nature of questions—some requiring speech understanding such as lyrics or quoted lines—we categorize models into three groups: (i) audio-only understanding, (ii) speech-aware models, and (iii) unified audio+speech models, reporting results separately for each subgroup. For all models, we use publicly released checkpoints and follow recommended inference settings when available, with no fine-tuning on our dataset.

We evaluate a diverse collection of state-of-the-art open-checkpoint models designed for audio and multimodal language understanding, spanning mid-scale language backbones ranging from approximately 6 billion to 17 billion parameters. These models differ widely in their training objectives, architectural designs, and audio processing strategies. Among them are models primarily trained for audio–text alignment, such as Qwen2.5-Omni (Xu et al., 2025a) and AudioGPT (Huang et al., 2024), which focus on aligning acoustic features with language representations to handle speech and audio understanding tasks. In addition, we consider large multimodal foundation models that incorporate audio inputs alongside other modalities, including OpenOmni (Luo et al., 2025), Audio-Flamingo (Kong et al., 2024), and Phi-4-Multimodal (Abouelenin et al., 2025). These models leverage powerful language backbones integrated with audio encoders, enabling flexible multimodal reasoning across diverse input types. Specialized audio language models such as SALMONN-2 and SALMONN-2+ (Tang et al., 2025) emphasize improved audio–language alignment and training methods to enhance reasoning over complex auditory inputs, while models like MU-LLaMA (Liu et al., 2024) and the original SALMONN (Tang et al., 2024) also contribute to the landscape of advanced multimodal comprehension. Other models, including Qwen3-Omni (Xu et al., 2025b), LTU-AS (Gong et al.,

Family	Typical Construction	Common Shortcut and What AUDITA Stresses	Typical Question Style & Example
Sound labeling	Closed label sets, short clips, event tags	Salient cue detection, weak long context. AUDITA uses longer clips and open answer space.	*Example:* Classify short environmental sounds (e.g., engine, applause).
Caption-derived QA	Questions derived from captions or metadata (e.g., OpenAQA (Gong et al., 2023b))	Lexical priors and caption artifacts. AUDITA uses human-authored trivia not derived from captions.	Caption-consistent attribute query. *Example:* Identify the sound source described by the caption.
Synthetic or templated	Generated scenes, fixed templates, constrained language	Template regularities and limited linguistic diversity. AUDITA uses natural, non-templated questions.	Programmatic logic over fixed attributes. *Example:* Count occurrences of a specified event.
Speech-centric QA	Spoken content dominates, transcript-like supervision	Can collapse to ASR plus text QA. AUDITA includes speech but also music and non-speech audio.	Spoken content identification, often transcript-based questions.
AUDITA	Human-written questions with real-world referents	Requires audio entity linking and multi-cue integration. This is the core target regime of AUDITA.	Probing trivia grounded in real referents. *Example:* Identify a film theme, speaker, or cultural artifact from audio cues.

Table 6: Positioning of AUDITA relative to common AQA benchmark designs, including typical dataset construction, common shortcuts that models exploit, and representative question styles and examples.

Dataset	Example Question	Answer (Gold)	IRT Difficulty / Discrimination	Issue / Notes
VGGSound QA (Chen et al., 2020)	"What type of animal is making the high-pitched and sharp sound described in the audio?"	*The high-pitched and sharp sound is most likely a bark produced by a dog.*	Low difficulty, low discrimination ($b = -2.1$, $a = 0.35$)	Simple audio classification task. Some clips have overlapping sounds causing ambiguity. Occasional metadata leakage possible, making some questions answerable without listening.
ClothoAQA (Drossos et al., 2020)	"Is this outdoors?"	*Yes*	Moderate difficulty, low discrimination ($b = -0.6$, $a = 0.42$)	Binary classification style question, with limited complexity. Contextual metadata sometimes gives away the answer. Not all questions require detailed auditory reasoning.
AudioCaps QA (Kim et al., 2019)	"Create a brief audio description, create labels, caption next."	*Labels: Vehicle; Tire squeal; Car; Race car, auto racing. Audio caption: Race car engines speed by, changing gears and screeching.*	Varied difficulty, often low discrimination ($b \approx 0.1$, $a = 0.28$)	Open-ended captioning questions. Subjective answers complicate evaluation and lack precise metrics. Do not constitute discrete QA.
"No listening needed" Questions (e.g. AudioCaps QA (Kim et al., 2019))	Examples: "What is the most likely reason for someone to strike a metal trailer with a wooden rod?" or "What is the significance of thunder in mythology?"	*There could be a variety of reasons, such as trying to get someone's attention, testing the durability of the trailer, or making a musical sound.* and *In many cultures, thunder represents the authority of gods and goddesses, and it is often associated with power, strength, and fertility.*	Very low difficulty, near-zero discrimination ($b = -3.1$, $a \approx 0$) and ($b = -3.4$, $a \approx 0$)	Answerable without listening, often appearing in scraped or poorly filtered datasets. Undermines auditory reasoning benchmarks. Not typical of curated datasets but important to highlight.

Table 7: Representative question examples from popular audio QA datasets, annotated with illustrative ranges of psychometric (IRT) difficulty and discrimination, and highlighting issues such as reliance on metadata, low reasoning complexity, and "no listening needed" questions. This underscores the need for carefully curated, human-authored datasets that robustly evaluate auditory reasoning.

2023a), Baichuan-Omni-1.5 (Li et al., 2025), and VITA-1.5 (Fu et al., 2025), further expand the diversity of architectures and training approaches assessed.

For each model, we report performance on both text-based and multiple-choice (MCQ) question formats from our dataset, including accuracy percentages and item response theory (IRT) estimated ability scores (θ), which provide a latent measure of model proficiency relative to question difficulty. Models are ranked within each task format to facilitate comparative evaluation. All evaluations utilize publicly released checkpoints with recommended inference settings, without any task-specific fine-tuning, ensuring an unbiased benchmarking environment. This comprehensive evaluation enables us to analyze strengths and limitations across modalities, task types, and audio reasoning capabilities, thereby offering insights into current progress and challenges in the field of audio question answering.

C Related Work

Audio question answering (Audio QA) Audio Question Answering remains a nascent field with relatively few datasets, many of which impose design constraints that limit their ability to evaluate genuine auditory reasoning. Early efforts such as CLEAR (Lin et al., 2021) construct synthetic acoustic scenes by layering individual musical notes from the GoodSounds database. Questions are programmatically generated from logical templates and target specific attributes, for example, "How many times does the C note occur in this clip?" While this approach enables precise semantic control, it restricts linguistic diversity and limits reasoning complexity.

Similarly, DAQA (Fayek and Johnson, 2020) composes variable-length audio clips from a closed vocabulary of 32 sound classes (e.g., "dog bark," "car horn") and asks questions such as "Does the sound of a car horn occur more than twice in this clip?" Although DAQA allows limited temporal

Issue	Dataset(s)	Example Question & Answer	Notes
Ambiguity	Clotho-AQA (Drossos et al., 2020), VGGSound QA (Chen et al., 2020), AudioCaps QA (Kim et al., 2019), FSD50K QA (Fonseca et al., 2022)	Q: "What animal is making the sound?" A: "bird" / "dog" (multiple audible) Q: "Where is the sound coming from?" A: "indoors" / "outside" (disagreement among participants) Q: "What is the dominant sound in the clip?" A: "siren" / "car horn" (subjective) Q: "Is there a sound of laughter?" A: "yes" / "no" (faintness ambiguity) Q: "Is there a vehicle sound?" A: "yes" / "no" (ambiguous engine/horn sounds)	Multiple overlapping sounds cause unclear targets; vague or underspecified question wording leads to inconsistent answers across annotators and models.
Weak Grounding / Shortcut	AudioCaps QA (Kim et al., 2019), AudioSet QA (Gemmeke et al., 2017)	Q: "Is someone laughing?" A: "yes" Q: "Is there music playing?" A: "yes"	Questions answerable from captions or metadata alone without listening; templated language encourages shortcut learning.
Underspecified Answer Keys	MUSIC-21 QA (Christodoulou et al., 2025), Clotho-AQA (Drossos et al., 2020)	Q: "What instrument is playing?" A: "piano" Q: "What animal can be heard?" A: "dog" / "puppy" / "hound"	Multiple valid lexical variants treated as separate answers, increasing noise.
False Presuppositions	Speech Commands QA (Warden, 2018)	Q: "Is the command 'stop' present?" A: "no"	Questions assume presence of commands that may not exist, confusing annotators and models.
Overlapping Audio	VGGSound QA (Chen et al., 2020), AudioCaps QA (Kim et al., 2019)	Q: "What animal is making the sound?" A: "dog"	Overlapping sound sources cause ambiguity, complicating correct labeling and answering.
Synthetic Question Bias	AudioSet QA (Gemmeke et al., 2017)	Q: "Is there music playing?" A: "yes"	Automated templated generation reduces linguistic diversity and causes models to exploit shortcuts.

Table 8: Summary of common issues in audio QA datasets, including improved, concrete ambiguous question examples from real datasets.

reasoning, its answer space is restricted to yes/no or counts, constraining the evaluation of richer auditory inference.

More naturalistic datasets attempt to move beyond synthetic audio. ClothoAQA (Drossos et al., 2020) relies on crowd workers to write questions about environmental sound recordings originally collected for captioning. Questions such as "What animal makes the sound in this clip?" or "Is the sound recorded indoors or outdoors?" better resemble real-world queries. However, this process introduces strong linguistic priors: models trained on ClothoAQA perform competitively even when audio is removed, indicating reliance on textual cues rather than acoustic understanding. MusicAVQA (Li et al., 2022) exhibits similar limitations, using templates like "What instrument is playing?" or "Is the tempo fast or slow?" These formats further limit linguistic variability and encourage shortcut learning, echoing issues observed in textual and visual QA (Section C).

Audio Question Answering Benchmarks Existing AQA benchmarks provide important testbeds for audio–language modeling but do not systematically expose failures in human-relevant auditory reasoning. CLEAR and ClothoAQA primarily reduce to controlled attribute queries or implicit classification tasks, making it difficult to distinguish true reasoning from surface-level pattern matching. While foundational, these benchmarks are limited in their ability to reveal nuanced model weaknesses in realistic, probing scenarios.

More recently, the DCASE 2025 Audio Question Answering challenge (Yang et al., 2025) introduced multi-domain QA subsets spanning bioacoustics, temporal soundscapes, and complex real-world audio. Although this effort broadens domain coverage, it remains centered on multiple-choice evaluation and lacks an adversarial, human-authored component designed to surface model brittleness.

Complementary evidence from Wang et al. (2025a) shows that large audio–language models are highly sensitive to misleading or conflicting textual cues paired with audio, revealing robustness gaps in current evaluations. Together, these findings motivate benchmarks that deliberately incorporate adversarial examples and human-authored questions to stress robust audio–language reasoning.

Adversarial Evaluation in QA and Multimodal Tasks Outside of audio QA specifically, adversarial evaluation has been successfully applied in other QA domains. In multimodal QA such as Visual Question Answering, human-in-the-loop adversarial data collection (e.g., Adversarial VQA) has been shown to produce questions that systematically expose model weaknesses by allowing annotators to target model failure modes through iterative feedback. Studies on adversarial QA in text also reveal that without adversarial examples, models can achieve high accuracy by exploiting dataset biases rather than robust reasoning. These findings underscore the value of adversarially collected questions for diagnosing model behavior, but comparable efforts are scarce for purely auditory content.

To address these gaps, we construct a human-

authored adversarial audio QA dataset grounded in real audio recordings and designed to be easy for humans but challenging for current models. Unlike prior AQA benchmarks, our dataset emphasizes semantic richness, adversarial focus, and natural realism, enabling evaluations that reveal model brittleness that structured or synthetic datasets fail to surface. **Multimodal Question Answering** One of the first large-scale multimodal datasets for question answering was the VQA dataset (Antol et al., 2015), which used image data to give models the context to answer a natural language question. The authors had crowd workers write questions about images from the COCO dataset and synthetically generated scenes to produce the data. Shortly after VQA was released in 2015, Johnson et al. (2017a) of the CLEVR dataset created a VQA task that uses fully synthetic scenes to produce a comprehensive visual reasoning test (Johnson et al., 2017b). One of the key aspects of this dataset was that questions were generated using a functional program, which would inspire future Visual Datasets.

One of the main issues within the VQA task was the presence of heavy priors within the data, where emergent statistical patterns would undermine the goal of reasoning over the image and the text. Many groups have been making efforts to improve upon this. The VQA v2 dataset balances the VQA dataset by introducing complements to each data point, where the new image is similar to the original but produces a different answer to the corresponding question (Goyal et al., 2017). Later, efforts were made to control the distribution of answers by (Agrawal et al., 2018) of the VQA-CP, who changed the splits of the VQA and VQA v2 datasets to alter the priors of the answer distribution. While remedies have been made to the original VQA dataset to fix its issues with heavy bias, other datasets have been introduced to overcome these pitfalls. One such dataset is GQA, which uses scene graphs based on images from COCO and Flickr to build questions automatically (Hudson and Manning, 2019). From the scene graph, questions and answers are built from a functional program similar to what was used in CLEVR, which further allowed the authors to smooth the answer distribution for various groups of questions.

Outside of entirely image/text-based multimodal QA datasets, several examples of datasets explore different mediums and combine already popular ones. For example, the MultimodalQA creates multimodal questions by composing single modality questions about Wikipedia tables and the entities linked within them (such as images or other objects) (Talmor et al., 2021). To compensate for the algorithmic generation of questions, the authors use crowd workers to rephrase the question into a more natural alternative and have other workers verify the question's validity. Several video-based datasets have also been released following the pattern of looking at combinations of modalities. For example, Yang et al. (2022) used videos from the VGG-sound dataset and expert annotators to write questions about each video for AVQA dataset. Similarly, Li et al. (2022) released the Music-AVQA dataset by collecting YouTube videos of music performances and crowd workers produce questions that followed a predefined template.

While many of these datasets utilize datasets from adjacent tasks of similar modality, only a subset of those available are web-curated. An advantage of our dataset in this field is that humans have already written the questions we collect outside of the context of our research, which minimizes much of the bias observed from directly using crowd workers to produce data for a benchmark.

Audio Question Answering – AQA datasets have been generally sparse over the past 6 years. Major contributions were either synthetically generated like DAQA (Fayek and Johnson, 2020) and CLEAR (Lin et al., 2021) or reliant on crowd workers for annotations like ClothoAQA (Drossos et al., 2020). One of the earliest examples of a contemporary AQA dataset is the CLEAR dataset, which shares many similarities with the CLEVR dataset. In particular, the CLEAR datasets combines individual musical notes from the GoodSounds database to generate an acoustic scene for a model to analyze. Like CLEVR, CLEAR's questions are constructed from templates represented as a logical tree with a functional program associated with them (Lin et al., 2021).

Using a similar methodology of combining smaller events, the DAQA dataset constructs audio-question pairs by stitching together several audio events. The main difference between the two is that DAQA uses events of variable length at various frequencies, so they can ask questions about how often a specific event occurs within a clip. While this dataset can help test a surface level of reasoning, the answer space is very small, with only 32 classes, with many answers being yes or no. Later, the Clotho AQA dataset was made, which takes advantage of the Clotho audio captioning dataset

and uses crowd workers to produce new questions. Regarding crowd work, the resulting dataset may often contain heavy priors as quality control has been difficult. There have been several attempts to treat this issue in the VQA space (Anderson et al., 2017); however, this hasn't been extended to the world of audio yet. Notably, this issue also skews results from this dataset, as many of the experiments conducted simply answering the question yield higher performance than using the audio and the question.

Adversarial Dataset Creation – As models grow increasingly complex, it becomes significantly more difficult to understand precisely why a model makes particular decisions during its inferences. A consequence of this trend is that understanding the weaknesses of a model turns into a hard task, especially when it comes to black-box models. The goal of adversarial dataset generation is to explore how robust models are to noisy data and to look at ways models underperform compared to humans. Thus, an important quality of adversarial questions is that they are hard for computers but relatively easy for humans; if it is very difficult for both groups, then it simply shows the dataset may be too tough. For VQA, a group of researchers made Adversarial VQA (AVQA), which uses a human-in-the-loop approach to generate questions that challenge models. An important aspect of their question creation interface is that a SOTA VQA model is present to provide answer feedback, so people can tweak questions until the model is tricked (Sheng et al., 2021). After the questions have been written, the model is retrained and given to the workers to write new, more difficult questions. In an interesting case of coincidence, within the same year, Li et al. (2021) produced a dataset called AdVQA using a similar technique. Although their annotation processes differ, AdVQA doesn't use model retraining. An interesting example of intentionally designing a challenging dataset is TriviaQA, a reading comprehension dataset that grounds its queries in trivia questions (Joshi et al., 2017). Since Trivia questions are designed to test the ability of humans, they are also a great way to benchmark models, as they are written to find out the best of a group of people.

D Result

D.1 Model Scale and Validity of Conclusions

Many audio-language models couple a pretrained audio encoder to a pretrained text LLM by projecting audio features into the LLM token space, so the language backbone is a key determinant of downstream instruction-following and knowledge-heavy QA performance (Gong et al., 2023b). Evidence from recent scaling analyses of speech-text and audio-centric language models also supports the general trend that larger backbones and stronger decoders improve aggregate performance across understanding and reasoning tasks, although gains vary by task and domain (Cuervo and Marxer, 2024).

At the same time, scale alone does not necessarily resolve failures caused by weak audio grounding and over-reliance on textual priors, which can manifest as confident but incorrect answers when audio evidence is insufficient or ignored (Wang et al., 2025b). This is consistent with MMAU, which reports that even strong proprietary systems remain far below human performance on its human-evaluated test-mini split and are only moderately separated from strong open models on the benchmark evaluation (Sakshi et al., 2024). For example, MMAU reports human performance of about 82.23 on test-mini, while Gemini Pro v1.5 achieves 52.97 on the test split and strong open models such as Qwen2-Audio-Instruct are comparable in several settings (Sakshi et al., 2024). These gains are meaningful, but they are not of a magnitude that would plausibly convert near-chance behavior on a difficult, human-authored benchmark into reliable audio reasoning.

Accordingly, while our evaluation focuses on open models in the mid-scale regime, our qualitative conclusions about failure modes and dataset difficulty are unlikely to be artifacts of model scale alone. Exhaustively evaluating cloud-scale proprietary models across all 9,690 questions would also impose substantial cost, which limits full scaling sweeps in this study.

Top 10 Best Discriminator Questions (IRT)		
#	**Question**	**Answer**
1	Name the title character of these movies.	*Sherlock Holmes*
2	What is the name of the person who is speaking in this clip?	*Harry Styles*
3	Name the character that inspired this music.	*Batman*
4	Name the lead artist.	*Halsey*
5	Give the common word found in the names of these lead artists.	*A$AP*
6	Name the artist.	*beabadoobee*
7	What country is this national anthem from?	*Bolivia*
8	What TV show is this clip from?	*A Team*
9	What TV show is this clip from?	*Happy days*
10	Name the city where these movies are wholly or mostly set	*Paris*

Top 10 Worst Discriminator Questions (IRT)		
#	**Question**	**Answer**
1	What is the language spoken in this clip?	*Chinese*
2	What is the language spoken in this clip?	*German*
3	What is the next line of lyrics that occurs after the song in the clip ends?	*Two and Two Were Four*
4	What is the next line of lyrics that occurs after the song in the clip ends?	*On the Pages in Between*
5	What country is this national anthem from?	*Iceland*
6	Name the character.	*Siegfried*
7	Name the mythical figure who is singing in these excerpts.	*Hades*
8	Name the male lead of these movies	*Charlie Chaplin*
9	What type of role do these characters have in common?	*Trouser role or Pants role or Breeches role or Male roles played by women*
10	What is the name of the person who is speaking in this clip?	*Bruce Forsyth*

Table 9: Examples of the top 10 best and worst discriminator questions by Item Response Theory (IRT), including their answers. High discrimination indicates questions that effectively differentiate between high- and low-ability respondents, while low discrimination indicates poor differentiation power.

Model	LLM Core	Inputs	Outputs	Rationale
Omnimodal models (6)				
Qwen2.5-Omni	7B	text, audio, image, video	text, speech	Thinker–Talker architecture with explicit reasoning–speech decoupling and unified multimodal support.
Qwen3-Omni	30B-A3B	text, audio, image, video	text, speech	MoE-based omni model (30B total, $\sim$3B active) optimized for scalable multimodal reasoning and streaming speech.
OpenOmni	7B	text, audio, image	text, speech	Language-pivot alignment with progressive modality training and preference-based speech tuning.
VITA-1.5	7B	text, audio, image, video	text, speech	End-to-end omni model using a three-stage training pipeline without cascaded ASR/TTS.
Mini-Omni2	0.5B	text, audio, image	text, speech	Lightweight end-to-end omni assistant with parallel text–audio decoding and command-based duplex interruption.
Baichuan-Omni-1.5	7B	text, audio, image, video	text, speech	Unified decoder with explicit audio token modeling and staged multimodal training, optimized for Chinese–English bilingual use.
Audio-language models (4)				
Audio-Flamingo	1.3B	text, audio	text	Flamingo-style gated cross-attention, sliding-window audio features, ICL/RAG and multi-turn dialogue.
Qwen2-Audio	7B	text, audio	text	Whisper-large-v3–initialized audio encoder into Qwen-7B, unified prompting, SFT+DPO alignment.
LTU-AS	7B	audio, text	text	Frozen Whisper perception + TLTR time/layer aggregation, continuous audio tokens + transcript, LLaMA-7B w/ LoRA.
MU-LLaMA	7B	text, audio	text	Frozen MERT music encoder + adapter injection into LLaMA-2 7B, MusicQA supervision for QA and captioning.

Table 10: Evaluated models organized by capability grouping (Part 1 of 2). All models evaluated in audio-question to text-answer setting.

Model	LLM Core	Inputs	Outputs	Rationale
Speech-capable models (6)				
Phi-4-Multimodal	3.8B	text, audio, image, video	text	Mixture-of-LoRAs with frozen language backbone, modality-specific adapters, 128K context.
SpeechGPT	13B	text, speech	text, speech	Discrete speech units expanded into LLaMA vocabulary, three-stage training with Chain-of-Modality instruction tuning.
AudioGPT	modular	text, speech, audio, music, image	text, audio, video	Modular orchestration system using ChatGPT to coordinate 16+ foundation models via ASR/TTS interface.
SALMONN	13B	text, speech, audio, music	text	Dual encoder (Whisper + BEATs), window-level Q-Former, activation tuning for emergent abilities.
video-SALMONN 2	7B	text, audio, video	text	Frozen backbone with audio branch, MrDPO for caption optimization, atomic event-based quality metrics.
video-SALMONN 2+	7B	text, audio, video	text	Caption-enhanced training via MrDPO-generated data, SOTA on Video-MME/WorldSense/AVUT benchmarks.

Table 11: Evaluated models organized by capability grouping (Part 2 of 2). All models evaluated in audio-question to text-answer setting.

Dataset / Paper	**Example or Description**	**Why Models Excel**
VGGSound (Chen et al., 2020)	Models detect synthetic or repeated alarm/beep sounds perfectly. These sounds have highly structured, repetitive waveforms easy for pattern matching.	Models trained on millions of audio clips memorize these patterns and detect them with near-perfect accuracy, often better than non-expert humans.
Speech Command Recognition (benchmark dataset) (Warden, 2018)	Recognizing isolated spoken command keywords like "stop," "go," or "yes" — models achieve $>99\%$ accuracy, often exceeding average human recognition.	Limited vocabulary and clean synthetic data make these tasks trivial for models.
AudioSet Tagging (Gemmeke et al., 2017)	Models detect environmental sounds like sirens, horns, or machine noises with very high precision, sometimes outperforming humans in noisy clips.	Large training data and strong feature extraction enable models to spot subtle acoustic cues missed by humans.

Table 12: Examples of tasks where models demonstrate superhuman or near-superhuman performance in audio question answering or classification.