Novelty Assessment Report

Paper: AudioTrust: Benchmarking The Multifaceted Trustworthiness of Audio Large Language Models
PDF URL: https://openreview.net/pdf?id=E823AY0taq
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

The rapid development and widespread adoption of Audio Large Language Models (ALLMs) require a rigorous assessment of their trustworthiness. However, existing evaluation frameworks, primarily designed for text, are not equipped to handle the unique vulnerabilities introduced by audio's acoustic properties. We find that significant trustworthiness risks in ALLMs arise from non-semantic acoustic cues, such as timbre, accent, and background noise, which can be used to manipulate model behavior. To address this gap, we propose AudioTrust, the first framework for large-scale and systematic evaluation of ALLM trustworthiness concerning these audio-specific risks. AudioTrust spans six key dimensions: fairness, hallucination, safety, privacy, robustness, and authenticition. It is implemented through 26 distinct sub-tasks and a curated dataset of over 4,420 audio samples collected from real-world scenarios (e.g., daily conversations, emergency calls, and voice assistant interactions), purposefully constructed to probe the trustworthiness of ALLMs across multiple dimensions. Our comprehensive evaluation includes 18 distinct experimental configurations and employs human-validated automated pipelines to objectively and scalably quantify model outputs. Experimental results reveal the boundaries and limitations of 14 state-of-the-art (SOTA) open-source and closed-source ALLMs when confronted with diverse high-risk audio scenarios, thereby offering critical insights into the secure and trustworthy deployment of future audio models. Our platform and benchmark are publicly available at https://anonymous.4open.science/r/AudioTrust-8715/.

Core Task Landscape

This paper addresses: **Trustworthiness Evaluation of Audio Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Comprehensive Trustworthiness Assessment Frameworks**
- **Adversarial Robustness and Security Vulnerabilities**
- **Hallucination Detection and Mitigation**
- **Modality Bias and Conflict Resolution**
- **Fairness and Demographic Bias**
- **Privacy Risks and Attribute Inference**
- **Reliability and Uncertainty Quantification**
- **Instruction-Following and Task Specification**
- **Explainability and Faithfulness**
- **Speech Quality and Descriptive Evaluation**
- ... and 4 more categories

Complete Taxonomy Tree

- Trustworthiness Evaluation of Audio Large Language Models Survey Taxonomy
- Comprehensive Trustworthiness Assessment Frameworks
  - Multi-Dimensional Audio-Specific Risk Evaluation ★ (2 papers)
  - [0] AudioTrust: Benchmarking The Multifaceted Trustworthiness of Audio Large Language Models (Anon et al., 2026) View paper
  - [12] Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms (Chowdhury, 2025) View paper
  - General Capability and Performance Benchmarking (3 papers)
  - [1] AudioBench: A Universal Benchmark for Audio Large Language Models (Wang Bin, 2024) View paper
  - [2] Towards Holistic Evaluation of Large Audio-Language Models: A Comprehensive Survey (Yang, 2025) View paper
  - [39] A Survey on Evaluation of Multimodal Large Language Models (Huang Jiaxing, 2024) View paper
- Adversarial Robustness and Security Vulnerabilities
  - Audio Injection and Perturbation Attacks (4 papers)
  - [3] Evaluating Robustness of Large Audio Language Models to Audio Injection: An Empirical Study (Hou Guanyu, 2025) View paper
  - [6] Who can withstand chat-audio attacks? an evaluation benchmark for large language models (Yang Wan-qi, 2024) View paper
  - [7] Who Can Withstand Chat-Audio Attacks? An Evaluation Benchmark for Large Audio-Language Models (Wanqi Yang, 2024) View paper
  - [22] Attacker's Noise Can Manipulate Your Audio-based LLM in the Real World (Sadasivan, 2025) View paper
  - Speech-Audio Compositional and Multimodal Attacks (2 papers)
  - [20] Speech-Audio Compositional Attacks on Multimodal LLMs and Their Mitigation with SALMONN-Guard (Yudong Yang, 2025) View paper
  - [36] Beyond Text: Multimodal Jailbreaking of Vision-Language and Audio Models through Perceptually Simple Transformations (Kumar Divyanshu, 2025) View paper

- Hallucination Detection and Mitigation
  - Object and Event Hallucination Assessment (3 papers)
  - [11] Understanding Sounds, Missing the Questions: The Challenge of Object Hallucination in Large Audio-Language Models (Huang Wei Ping, 2024) View paper
  - [26] Assessing Audio Hallucination in Large Multimodal Models (Namesa Kirishima, 2024) View paper
  - [50] Can Large Audio-Language Models Truly Hear? Tackling Hallucinations with Multi-Task Assessment and Stepwise Audio Reasoning (Chun-Yi Kuan, 2024) View paper
  - Hallucination Mitigation Through Training and Reasoning (2 papers)
  - [16] Teaching Audio-Aware Large Language Models What Does Not Hear: Mitigating Hallucinations through Synthesized Negative Samples (Lee, 2025) View paper
  - [34] From Alignment to Advancement: Bootstrapping Audio-Language Alignment with Synthetic Data (Lee, 2025) View paper
- Modality Bias and Conflict Resolution (2 papers)
  - [4] When audio and text disagree: Revealing text bias in large audio-language models (Wang Cheng, 2025) View paper
  - [29] A Multimodal Emotion Analysis System for Psychological Counseling Using Image, Audio, and Large Language Model Inference (Shuang Sun, 2025) View paper
- Fairness and Demographic Bias
  - Paralinguistic and Emotional Variation Bias (2 papers)
  - [24] Investigating Safety Vulnerabilities of Large Audio-Language Models Under Speaker Emotional Variations (Feng Bo Han, 2025) View paper
  - [32] TrustNavGPT: Modeling Uncertainty to Improve Trustworthiness of Audio-Guided LLM-Based Robot Navigation (Xingpeng Sun, 2024) View paper
  - Demographic and Accent-Based Bias (1 papers)
  - [41] Evaluating Bias in Spoken Dialogue LLMs for Real-World Decisions and Recommendations (Wu Yihao, 2025) View paper
- Privacy Risks and Attribute Inference (1 papers)
  - [9] The man behind the sound: Demystifying audio private attribute profiling via multimodal large language model agents (Wang Lixu, 2025) View paper
- Reliability and Uncertainty Quantification
  - Knowledge Boundary Recognition and Refusal Mechanisms (1 papers)
  - [18] Towards Reliable Large Audio Language Model (Ma, 2025) View paper
  - Confidence-Based Reliability and Self-Improvement (2 papers)
  - [37] A Novel Framework for Uncertainty Quantification via Proper Scores for Classification and Beyond (Gruber, 2025) View paper
  - [40] Self-Improving Intelligence: Learning From Unlabeled Data Across Domains (Duane, 2025) View paper
- Instruction-Following and Task Specification
  - Instruction-Following Capability Assessment (1 papers)
  - [14] IFEval-Audio: Benchmarking Instruction-Following Capability in Audio-based Large Language Models (Gao, 2025) View paper
  - Prompt Sensitivity and Attention-Based Task Specification (1 papers)
  - [8] AHAMask: Reliable Task Specification for Large Audio Language Models without Instructions (Guo, 2025) View paper
  - Selection Bias and Answer Ordering Effects (1 papers)
  - [25] Hearing the Order: Investigating Selection Bias in Large Audio-Language Models (Lin Yu-xiang, 2025) View paper
- Explainability and Faithfulness (2 papers)
  - [17] Investigating Faithfulness in Large Audio Language Models (Lovenya Jain, 2025) View paper
  - [33] Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey (Dang, 2024) View paper
- Speech Quality and Descriptive Evaluation (2 papers)
  - [19] Audio large language models can be descriptive speech quality evaluators (Chen Chen, 2025) View paper
  - [27] Reliability and Suitability of Evaluation of Speech by LLM Using Voice Mode (Yugo Tagami, 2025) View paper
- Domain-Specific Applications and Trustworthiness
  - Healthcare and Mental Health Applications (4 papers)
  - [5] Interface Matters: Exploring Human Trust in Health Information from Large Language Models via Text, Speech, and Embodiment (Xin Sun, 2025) View paper
  - [15] Reinforcing Trustworthiness in Multimodal Emotional Support Systems (Huy M. Le, 2025) View paper
  - [21] GPT-4o and multimodal large language models as companions for mental wellbeing (Surendrabikram Thapa, 2024) View paper
  - [44] SpeechT-RAG: Reliable Depression Detection in LLMs with Retrieval-Augmented Generation Using Speech Timing Information (Xiangyu Zhang, 2025) View paper
  - Specialized Populations and Domains (3 papers)
  - [23] The role of large language models in musicology: Are we ready to trust the machines? (Ramoneda, 2024) View paper
  - [45] Can large audio language models understand child stuttering speech? speech summarization, and source separation (Chibuzor Okocha, 2025) View paper
  - [46] Fine-Tuning Large Language Models for Saudi Arabic Voice Agents (Mahmoud Abdelhadi Mahmoud Safia, 2024) View paper
- Cross-Modal and Multimodal Integration (4 papers)
  - [13] Bridging Ears and Eyes: Analyzing Audio and Visual Large Language Models to Humans in Visible Sound Recognition and Reducing Their Sensory Gap via Cross â¦ (X Jiang, 2025) View paper
  - [35] Larger Encoders, Smaller Regressors: Exploring Label Dimensionality Reduction and Multimodal Large Language Models as Feature Extractors for Predicting Social Perception (IvÃ¡n MartÃn-FernÃ¡ndez, 2024) View paper
  - [43] Two Heads Are Better Than One: Audio-Visual Speech Error Correction with Dual Hypotheses (Kim, 2025) View paper
  - [49] Multimodal large language model enhancement network for multimodal sentiment analysis (Huanyu Zhu, 2025) View paper
- Architectural and Methodological Foundations (5 papers)
  - [10] Towards trustworthy and reliable language models (R. P. Zhao, 2025) View paper
  - [30] Toward Edge General Intelligence with Multiple-Large Language Model (Multi-LLM): Architecture, Trust, and Orchestration (Luo Haoxiang, 2025) View paper
  - [31] Rethinking Hallucinations: A Cognitive-Inspired Taxonomy and Comprehensive Survey in Large Language Models, Large Vision-Language Models, and Multimodal â¦ (SJ Xia, 2025) View paper
  - [42] Response to M. Trengove & coll regarding â□□Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicineâ□□ (Stefan Harrer, 2023) View paper
  - [48] Protect: Towards Robust Guardrailing Stack for Trustworthy Enterprise LLM Systems (Pareek, 2025) View paper

- Emerging Applications and Technical Utilities (3 papers)
  - [28] NLOS Identification and Ranging Trustworthiness for Indoor Positioning With LLM-Based UWBâ□□IMU Fusion (Hongchao Yang, 2025) View paper
  - [38] Comprehensive Investigation of Algorithmic Models and Prospects in Artificial Intelligence Generated Content (Bofeng Peng, 2025) View paper
  - [47] Leveraging Audio LLMs for Data Fault Detection in Audio Datasets (Guoying Chen, 2025) View paper

Narrative

Core task: trustworthiness evaluation of audio large language models. As audio-capable large language models become increasingly deployed, the field has organized around a multi-dimensional view of trustworthiness that spans adversarial robustness, hallucination detection, fairness, privacy, reliability, and explainability. The taxonomy reflects this breadth through branches addressing comprehensive assessment frameworks that evaluate models across multiple risk dimensions simultaneously (e.g., AudioTrust[0], Avtrustbench[12]), alongside specialized branches targeting individual concerns such as adversarial attacks (Audio Injection Robustness[3], Chat Audio Attacks[6]), modality conflicts (Audio Text Disagreement[4]), demographic bias (Spoken Dialogue Bias[41]), and privacy risks (Audio Private Profiling[9]). Additional branches cover instruction-following fidelity (IFEval Audio[14]), faithfulness and explainability (Faithfulness Audio Language[17]), and domain-specific applications ranging from mental health support (Mental Wellbeing Companions[21]) to speech quality assessment (Descriptive Speech Quality[19]).

A particularly active line of work focuses on holistic benchmarking that integrates perception, reasoning, and safety dimensions, with AudioBench[1] and Holistic Audio Language Evaluation[2] establishing multi-faceted evaluation protocols. In contrast, many studies drill into specific vulnerabilities: adversarial robustness research explores injection attacks and jailbreaking (Chat Audio Attacks Benchmark[7], Multimodal Jailbreaking[36]), while hallucination studies examine object-level errors (Object Hallucination Audio[11], Audio Hallucination Assessment[26]) and mitigation strategies (AHAMask[8]). AudioTrust[0] sits within the comprehensive assessment branch, emphasizing multi-dimensional risk evaluation that spans adversarial, fairness, privacy, and reliability concerns in a unified framework. Compared to narrower benchmarks like Audio Injection Robustness[3] or domain-focused evaluations such as Interface Trust Health[5], AudioTrust[0] adopts a broader lens, aiming to capture the interplay among diverse trustworthiness facets rather than isolating individual threat models or application contexts.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms

Authors: Chowdhury, Sanjoy, Nag, Sayan, Sanjoy Chowdhury, et al. (19 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

With the rapid advancement of Multi-modal Large Language Models (MLLMs), several diagnostic benchmarks have recently been developed to assess these models' multi-modal reasoning proficiency. However, these benchmarks are restricted to assessing primarily the visual aspect and do not examine the holistic audio-visual (AV) understanding. Moreover, currently, there are no benchmarks that investigate the capabilities of AVLLMs to calibrate their responses when presented with perturbed inputs. To thi...

Relationship Analysis

Both papers belong to the Multi-Dimensional Audio-Specific Risk Evaluation category, assessing trustworthiness of audio large language models across multiple dimensions using comprehensive benchmarks. They overlap in evaluating robustness, safety, and hallucination risks arising from acoustic properties, with both employing multi-task evaluation frameworks and automated assessment pipelines. However, AudioTrust focuses on six dimensions (fairness, hallucination, safety, privacy, robustness, authentication) with 26 sub-tasks and 4,420 audio samples emphasizing non-semantic acoustic cues like timbre and accent, while AVTrustBench emphasizes three dimensions (adversarial attack, compositional reasoning, modality-specific dependency) with 600K samples and introduces a novel preference optimization training strategy (CAVPref) to improve model reliability.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: AudioTrust benchmark framework for evaluating ALLM trustworthiness

Description: The authors introduce AudioTrust, the first comprehensive benchmark designed to systematically evaluate the trustworthiness of Audio Large Language Models across six critical dimensions: fairness, hallucination, safety, privacy, robustness, and authentication. The framework addresses unique vulnerabilities introduced by audio's acoustic properties that existing text-based evaluation frameworks cannot capture.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Safebench: A safety evaluation framework for multimodal large language models

URL: View paper

Brief Assessment

Safebench[61] focuses on multimodal large language models (MLLMs) with emphasis on vision and text modalities, not specifically audio large language models (ALLMs). While both are safety evaluation frameworks, they target different model types and modalities.

---

2. Thinking with Sound: Audio Chain-of-Thought Enables Multimodal Reasoning in Large Audio-Language Models

URL: View paper

Brief Assessment

Audio Chain of Thought[64] focuses on audio reasoning robustness through chain-of-thought techniques and acoustic tool integration, not on comprehensive trustworthiness evaluation across fairness, hallucination, safety, privacy, robustness, and authentication dimensions.

---

3. AudioBench: A Universal Benchmark for Audio Large Language Models

URL: View paper

Brief Assessment

AudioBench[1] focuses on evaluating audio large language models across general capabilities (speech understanding, audio scene understanding, voice understanding) rather than trustworthiness dimensions. The candidate does not address fairness, hallucination, safety, privacy, robustness, or authentication evaluation.

---

4. JALMBench: Benchmarking Jailbreak Vulnerabilities in Audio Language Models

URL: View paper

Brief Assessment

JALMBench[62] focuses specifically on jailbreak attacks against audio language models, while the original paper proposes a comprehensive trustworthiness evaluation across six dimensions (fairness, hallucination, safety, privacy, robustness, authentication). These are distinct evaluation frameworks with different scopes and objectives.

---

5. Audio Jailbreak: An Open Comprehensive Benchmark for Jailbreaking Large Audio-Language Models

URL: View paper

Brief Assessment

Audio Jailbreak Benchmark[65] focuses specifically on jailbreak attacks against large audio-language models, while AudioTrust provides a comprehensive multi-dimensional trustworthiness evaluation framework spanning fairness, hallucination, safety, privacy, robustness, and authentication. The candidate addresses only one dimension (safety via jailbreak attacks) of the original's six-dimensional framework.

---

6. Ahelm: A holistic evaluation of audio-language models

URL: View paper

Brief Assessment

Ahelm[60] focuses on holistic evaluation of audio-language models across 10 aspects (audio perception, knowledge, reasoning, emotion detection, bias, fairness, multilinguality, robustness, toxicity, and safety), whereas AudioTrust specifically targets trustworthiness evaluation across six dimensions (fairness, hallucination, safety, privacy, robustness, and authentication) with emphasis on audio-specific vulnerabilities from acoustic properties. The frameworks have different scopes and evaluation methodologies.

---

7. When audio and text disagree: Revealing text bias in large audio-language models

URL: View paper

Brief Assessment

Audio Text Disagreement[4] focuses on text bias when audio and text modalities conflict in large audio-language models, not on comprehensive trustworthiness evaluation across fairness, hallucination, safety, privacy, robustness, and authentication dimensions.

---

8. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms

URL: View paper

Brief Assessment

Avtrustbench[12] focuses on audio-visual LLMs (AVLLMs) and evaluates adversarial robustness, compositional reasoning, and modality dependency. The original paper targets audio-only LLMs (ALLMs) with different dimensions (fairness, hallucination, safety, privacy, robustness, authentication).

---

9. Air-bench: Benchmarking large audio-language models via generative comprehension

URL: View paper

Brief Assessment

Air-Bench[63] focuses on evaluating generative comprehension abilities across speech, natural sounds, and music through foundation and chat benchmarks. It does not address trustworthiness dimensions like fairness, hallucination, safety, privacy, robustness, or authentication that are central to AudioTrust.

---

10. MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models

URL: View paper

Brief Assessment

MuChoMusic[66] focuses on evaluating music understanding capabilities in audio-language models through multiple-choice questions, not on trustworthiness dimensions like fairness, hallucination, safety, privacy, robustness, and authentication that AudioTrust addresses.

Contribution 2: Curated dataset of over 4,420 audio samples across 26 sub-tasks

Description: The authors construct a large-scale dataset comprising over 4,420 audio samples spanning 26 distinct sub-tasks and 18 experimental configurations. The samples are purposefully collected from real-world scenarios such as daily conversations, emergency calls, and voice assistant interactions to probe trustworthiness across multiple dimensions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. An invasive species model and dataset for bioacoustic monitoring of common brushtail possum

URL: View paper

Brief Assessment

Possum Bioacoustic Monitoring[56] focuses on a specialized invasive species detection dataset (3,500 possum vocalizations), not a multi-dimensional trustworthiness evaluation framework spanning fairness, safety, privacy, and robustness across 26 sub-tasks.

---

2. COVID-19 sounds: a large-scale audio dataset for digital respiratory screening

URL: View paper

Brief Assessment

COVID Sounds Dataset[58] focuses on respiratory health screening with breathing, cough, and voice recordings for COVID-19 detection, not a multi-dimensional trustworthiness evaluation framework across fairness, safety, privacy, and other dimensions as in the original paper.

---

3. Acoustic-Based Industrial Diagnostics: A Scalable Noise-Robust Multiclass Framework for Anomaly Detection

URL: View paper

Brief Assessment

Acoustic Industrial Diagnostics[57] focuses on industrial machine anomaly detection using the MIMII dataset, not on evaluating model trustworthiness across conversational tasks like emergency calls or voice assistant interactions.

---

4. Windy events detection in big bioacoustics datasets using a pre-trained Convolutional Neural Network

URL: View paper

Brief Assessment

Windy Events Detection[51] focuses on detecting wind noise in bioacoustic recordings using a pre-trained CNN (YAMNet), not on constructing a multi-dimensional trustworthiness evaluation dataset for audio language models.

---

5. Integrating Vehicle Acoustic Data for Enhanced Urban Traffic Management: A Study on Speed Classification in Suzhou

URL: View paper

Brief Assessment

Vehicle Acoustic Traffic[54] focuses on a vehicular acoustic dataset for speed classification in urban traffic management, not a multi-task trustworthiness evaluation dataset spanning fairness, hallucination, safety, privacy, robustness, and authentication dimensions.

---

6. Sounds of the deep: How input representation, model choice, and dataset size influence underwater sound classification performance

URL: View paper

Brief Assessment

Underwater Sound Classification[53] focuses on marine acoustic classification using passive acoustic data from three offshore sites in Scotland, not on evaluating model trustworthiness across multiple dimensions as in the original paper.

---

7. Listener Acoustic Personalization Challengeâ□□LAP24: Head-Related Transfer Function Dataset Harmonization

URL: View paper

Brief Assessment

HRTF Dataset Harmonization[59] focuses on harmonizing head-related transfer function datasets for spatial audio personalization, not on constructing a multi-task trustworthiness evaluation dataset for audio language models.

---

8. THAI Speech Emotion Recognition (THAI-SER) corpus

URL: View paper

Brief Assessment

THAI SER[55] focuses on Thai speech emotion recognition with 27,854 utterances across 5 emotions, not a multi-dimensional trustworthiness evaluation framework spanning 26 sub-tasks across fairness, hallucination, safety, privacy, robustness, and authentication.

---

9. A zero-shot model for diagnosing unknown composite faults in train bearings based on label feature vector generated fault features

URL: View paper

Brief Assessment

Zero Shot Composite Faults[52] focuses on train bearing fault diagnosis using acoustic datasets, not trustworthiness evaluation of audio language models across diverse real-world scenarios.

---

10. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms

URL: View paper

Brief Assessment

Avtrustbench[12] comprises 600k samples for audio-visual evaluation, not audio-only. The scale and modality focus differ fundamentally from the original paper's 4,420 audio samples across 26 sub-tasks for ALLM trustworthiness.

Contribution 3: Human-validated automated evaluation pipeline for scalable assessment

Description: The authors develop an automated evaluation pipeline that employs model-based evaluators (GPT-4o and Qwen3) with human validation to ensure rigorous and reproducible assessment. The pipeline achieves over 97% agreement rate with human experts and enables scalable quantification of model outputs across diverse high-risk audio scenarios.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. MEMERAG: A Multilingual End-to-End Meta-Evaluation Benchmark for Retrieval Augmented Generation

URL: View paper

Brief Assessment

MEMERAG[72] focuses on meta-evaluation of RAG systems using human annotations for faithfulness and relevance, not on automated evaluation pipelines with human validation for audio model assessment. The domains and evaluation targets differ fundamentally.

---

2. Large Language Model-Powered Automated Assessment: A Systematic Review

URL: View paper

Brief Assessment

Automated Assessment Review[69] focuses on LLM-powered automated assessment in educational contexts (reading comprehension, language education), not on evaluating audio large language models across trustworthiness dimensions as in the original paper.

---

3. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

4. IQA-EVAL: Automatic Evaluation of Human-Model Interactive Question Answering

URL: View paper

Brief Assessment

IQA EVAL[74] focuses on interactive question answering evaluation using LLM-based agents to simulate human behaviors, not on evaluating audio large language models across trustworthiness dimensions with acoustic-specific risks.

---

5. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization

URL: View paper

Brief Assessment

PandaLM[68] focuses on evaluating instruction-tuned LLMs for hyperparameter optimization, not on assessing trustworthiness of audio large language models across diverse high-risk scenarios. The evaluation contexts and objectives differ fundamentally.

---

6. Chartcap: Mitigating hallucination of dense chart captioning

URL: View paper

Brief Assessment

Chartcap[76] focuses on chart captioning evaluation using visual consistency metrics and cycle consistency-based verification, not on general audio model trustworthiness assessment pipelines.

---

7. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation

URL: View paper

Brief Assessment

Videoautoarena[73] focuses on automated evaluation for video analysis tasks using LMMs, not audio-specific trustworthiness assessment. The technical domains and evaluation targets differ fundamentally.

---

8. STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment

URL: View paper

Brief Assessment

STAIR AIG[71] focuses on human-in-the-loop review of AI-generated assessment items for critical thinking, not on automated evaluation pipelines for model outputs. The candidate addresses educational item generation and review, while the original paper evaluates audio large language model trustworthiness.

---

9. Alignbench: Benchmarking chinese alignment of large language models

URL: View paper

Brief Assessment

AlignBench[70] focuses on evaluating Chinese language alignment of LLMs using text-based queries and responses, not audio-specific evaluation pipelines for audio large language models across diverse high-risk audio scenarios.

---

10. Human-Calibrated Automated Testing and Validation of Generative Language Models

URL: View paper

Prior Art Analysis

Human Calibrated Testing[75] demonstrates that similar automated evaluation pipelines with human validation existed prior to the original paper. The candidate paper presents a comprehensive framework called HCAT (Human-Calibrated Automated Testing) that employs model-based evaluators with human validation to ensure rigorous assessment. Like the original paper, HCAT uses automated metrics combined with human calibration to achieve high agreement rates with human experts and enables scalable quantification of model outputs. Both papers describe two-stage evaluation processes where automated metrics are calibrated against human judgments, achieving similar high agreement rates (97-98% in the candidate vs. over 97% in the original).

Evidence

Evidence 1 - **Rationale**: Both papers describe comprehensive automated evaluation pipelines that are calibrated with human judgments. The candidate's HCAT framework explicitly includes automated metrics aligned with human evaluations, predating the original paper's similar approach. - **Original**: we deploy a large-scale automated evaluation pipeline to ensure rigorous and reproducible assessment. the reliability of our automated metrics and results is verified by human experts (over 97% agreement rate). - **Candidate**: hcat integrates a) automated test generation using stratified sampling, b) embedding-based metrics for explainable assessment of functionality, risk and safety attributes, and c) a two-stage calibration approach that aligns machine-generated evaluations with human judgments through probability calib...

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] AudioTrust: Benchmarking The Multifaceted Trustworthiness of Audio Large Language Models View paper
- [1] AudioBench: A Universal Benchmark for Audio Large Language Models View paper
- [2] Towards Holistic Evaluation of Large Audio-Language Models: A Comprehensive Survey View paper
- [3] Evaluating Robustness of Large Audio Language Models to Audio Injection: An Empirical Study View paper
- [4] When audio and text disagree: Revealing text bias in large audio-language models View paper
- [5] Interface Matters: Exploring Human Trust in Health Information from Large Language Models via Text, Speech, and Embodiment View paper
- [6] Who can withstand chat-audio attacks? an evaluation benchmark for large language models View paper
- [7] Who Can Withstand Chat-Audio Attacks? An Evaluation Benchmark for Large Audio-Language Models View paper
- [8] AHAMask: Reliable Task Specification for Large Audio Language Models without Instructions View paper
- [9] The man behind the sound: Demystifying audio private attribute profiling via multimodal large language model agents View paper
- [10] Towards trustworthy and reliable language models View paper
- [11] Understanding Sounds, Missing the Questions: The Challenge of Object Hallucination in Large Audio-Language Models View paper
- [12] Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms View paper
- [13] Bridging Ears and Eyes: Analyzing Audio and Visual Large Language Models to Humans in Visible Sound Recognition and Reducing Their Sensory Gap via Cross â View paper
- [14] IFEval-Audio: Benchmarking Instruction-Following Capability in Audio-based Large Language Models View paper
- [15] Reinforcing Trustworthiness in Multimodal Emotional Support Systems View paper
- [16] Teaching Audio-Aware Large Language Models What Does Not Hear: Mitigating Hallucinations through Synthesized Negative Samples View paper
- [17] Investigating Faithfulness in Large Audio Language Models View paper
- [18] Towards Reliable Large Audio Language Model View paper

- [19] Audio large language models can be descriptive speech quality evaluators View paper
- [20] Speech-Audio Compositional Attacks on Multimodal LLMs and Their Mitigation with SALMONN-Guard View paper
- [21] GPT-4o and multimodal large language models as companions for mental wellbeing View paper
- [22] Attacker's Noise Can Manipulate Your Audio-based LLM in the Real World View paper
- [23] The role of large language models in musicology: Are we ready to trust the machines? View paper
- [24] Investigating Safety Vulnerabilities of Large Audio-Language Models Under Speaker Emotional Variations View paper
- [25] Hearing the Order: Investigating Selection Bias in Large Audio-Language Models View paper
- [26] Assessing Audio Hallucination in Large Multimodal Models View paper
- [27] Reliability and Suitability of Evaluation of Speech by LLM Using Voice Mode View paper
- [28] NLOS Identification and Ranging Trustworthiness for Indoor Positioning With LLM-Based UWBâ□□IMU Fusion View paper
- [29] A Multimodal Emotion Analysis System for Psychological Counseling Using Image, Audio, and Large Language Model Inference View paper
- [30] Toward Edge General Intelligence with Multiple-Large Language Model (Multi-LLM): Architecture, Trust, and Orchestration View paper
- [31] Rethinking Hallucinations: A Cognitive-Inspired Taxonomy and Comprehensive Survey in Large Language Models, Large Vision-Language Models, and Multimodal â□¦ View paper
- [32] TrustNavGPT: Modeling Uncertainty to Improve Trustworthiness of Audio-Guided LLM-Based Robot Navigation View paper
- [33] Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey View paper
- [34] From Alignment to Advancement: Bootstrapping Audio-Language Alignment with Synthetic Data View paper
- [35] Larger Encoders, Smaller Regressors: Exploring Label Dimensionality Reduction and Multimodal Large Language Models as Feature Extractors for Predicting Social Perception View paper
- [36] Beyond Text: Multimodal Jailbreaking of Vision-Language and Audio Models through Perceptually Simple Transformations View paper
- [37] A Novel Framework for Uncertainty Quantification via Proper Scores for Classification and Beyond View paper
- [38] Comprehensive Investigation of Algorithmic Models and Prospects in Artificial Intelligence Generated Content View paper
- [39] A Survey on Evaluation of Multimodal Large Language Models View paper
- [40] Self-Improving Intelligence: Learning From Unlabeled Data Across Domains View paper
- [41] Evaluating Bias in Spoken Dialogue LLMs for Real-World Decisions and Recommendations View paper
- [42] Response to M. Trengove & coll regarding â□□Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicineâ□□ View paper
- [43] Two Heads Are Better Than One: Audio-Visual Speech Error Correction with Dual Hypotheses View paper
- [44] SpeechT-RAG: Reliable Depression Detection in LLMs with Retrieval-Augmented Generation Using Speech Timing Information View paper
- [45] Can large audio language models understand child stuttering speech? speech summarization, and source separation View paper
- [46] Fine-Tuning Large Language Models for Saudi Arabic Voice Agents View paper
- [47] Leveraging Audio LLMs for Data Fault Detection in Audio Datasets View paper
- [48] Protect: Towards Robust Guardrailing Stack for Trustworthy Enterprise LLM Systems View paper
- [49] Multimodal large language model enhancement network for multimodal sentiment analysis View paper
- [50] Can Large Audio-Language Models Truly Hear? Tackling Hallucinations with Multi-Task Assessment and Stepwise Audio Reasoning View paper
- [51] Windy events detection in big bioacoustics datasets using a pre-trained Convolutional Neural Network View paper
- [52] A zero-shot model for diagnosing unknown composite faults in train bearings based on label feature vector generated fault features View paper
- [53] Sounds of the deep: How input representation, model choice, and dataset size influence underwater sound classification performance View paper
- [54] Integrating Vehicle Acoustic Data for Enhanced Urban Traffic Management: A Study on Speed Classification in Suzhou View paper
- [55] THAI Speech Emotion Recognition (THAI-SER) corpus View paper
- [56] An invasive species model and dataset for bioacoustic monitoring of common brushtail possum View paper
- [57] Acoustic-Based Industrial Diagnostics: A Scalable Noise-Robust Multiclass Framework for Anomaly Detection View paper
- [58] COVID-19 sounds: a large-scale audio dataset for digital respiratory screening View paper
- [59] Listener Acoustic Personalization Challengeâ□□LAP24: Head-Related Transfer Function Dataset Harmonization View paper
- [60] Ahelm: A holistic evaluation of audio-language models View paper
- [61] Safebench: A safety evaluation framework for multimodal large language models View paper
- [62] JALMBench: Benchmarking Jailbreak Vulnerabilities in Audio Language Models View paper
- [63] Air-bench: Benchmarking large audio-language models via generative comprehension View paper
- [64] Thinking with Sound: Audio Chain-of-Thought Enables Multimodal Reasoning in Large Audio-Language Models View paper
- [65] Audio Jailbreak: An Open Comprehensive Benchmark for Jailbreaking Large Audio-Language Models View paper
- [66] MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models View paper
- [67] HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models View paper
- [68] PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization View paper
- [69] Large Language Model-Powered Automated Assessment: A Systematic Review View paper
- [70] Alignbench: Benchmarking chinese alignment of large language models View paper
- [71] STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment View paper
- [72] MEMERAG: A Multilingual End-to-End Meta-Evaluation Benchmark for Retrieval Augmented Generation View paper
- [73] Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation View paper
- [74] IQA-EVAL: Automatic Evaluation of Human-Model Interactive Question Answering View paper
- [75] Human-Calibrated Automated Testing and Validation of Generative Language Models View paper
- [76] Chartcap: Mitigating hallucination of dense chart captioning View paper