A Survey of Large Audio Language Models: Generalization, Trustworthiness, and Outlook

Anonymous authors
Paper under double-blind review

Abstract

Advances in Large Language Models (LLMs) have paved the way for Multimodal Large Language Models (MLLMs). Among these, Large Audio Language Models (LALMs) are essential for realizing universal auditory intelligence. Despite their remarkable performance, the escalation of LALMs' capabilities has significantly outpaced the development of systemic frameworks to ensure their trustworthiness. This survey provides a comprehensive investigation into the endogenous mechanisms of LALMs, detailing the architectural innovations and alignment algorithms that facilitate emergent reasoning. Specifically, we analyze how the transition to unified end-to-end frameworks and the integration of continuous acoustic signals expand the attack surface. To rigorously evaluate the risks within these paradigms, we establish a comprehensive taxonomy of trustworthiness, categorizing critical vulnerabilities such as cross-modal jailbreaking, latent acoustic backdoors, and biometric privacy leakage. We review the state-of-the-art LALMs through six analytical pillars: hallucination, robustness, safety, privacy, fairness, and authentication. The pronounced imbalance between a mature offensive landscape and underdeveloped defenses highlights persistent trustworthiness gaps and multidimensional risks in audio-centric intelligence. Finally, we propose a roadmap advocating for "Defense-in-Depth" architectures, causal auditory world modeling, and intrinsic representation engineering to support the development of more reliable and trustworthy audio intelligence.

1 Introduction

The emergence of Large Language Models (LLMs) Ouyang et al. (2022); Achiam et al. (2023); Touvron et al. (2023); Bai et al. (2023); Liu et al. (2024a); Guo et al. (2025a) has transformed the landscape of artificial intelligence, establishing a robust foundation for the transition toward unified multimodal frameworks. This evolution into Multimodal Large Language Models (MLLMs) Yan et al. (2025d;c); Team (2026) is designed to emulate the multi-sensory nature of human perception across diverse sensory inputs. Among human senses, **audio** represents a primary medium for human communication and perception of the environment Latif et al. (2023), as it carries a vast amount of information within its signal. Previous research in audio intelligence relied on modular systems designed for a single task, such as automatic speech recognition Wang et al. (2026c); Shi et al. (2026) or sound classification Gemmeke et al. (2017); Kong et al. (2020). The recent transition from these artifacts to unified **Large Audio Language Models (LALMs)** Chu et al. (2023); Tang et al. (2023); Rubenstein et al. (2023); Chu et al. (2024); Wu et al. (2025a) represents a step for universal audio intelligence.

Despite these remarkable advancements in auditory capabilities, the organic integration of language and audio modalities introduces complex safety and alignment challenges. Textual LLMs primarily address vulnerabilities within discrete text Shi et al. (2024); Wang et al. (2025g); Yu et al. (2025); Ma et al. (2026a). In contrast, LALMs introduce the audio modality, which presents an intricate risk landscape Lin et al. (2025b); Chen et al. (2025b); Aloufi et al. (2026); Chen et al. (2026a) due to the continuous properties of the acoustic signal. The deployment of LALMs within critical sectors further expands this complex risk landscape, translating these continuous-signal vulnerabilities into real-world threats. However, while the development of these capabilities is expanding, the research landscape remains fragmented and lacks

a unified roadmap. Existing research predominantly details architectural innovations Sakshi et al. (2025); Alex et al. (2025); You et al. (2026) or specific concerns Luong et al. (2025); Li et al. (2025g); Nguyen & Le (2026), yet there remains a significant lack of work dedicated to a systematic taxonomy of the safety implications for these systems. Recognizing that intrinsic trustworthiness cannot be guaranteed without a deep understanding of the underlying architecture, this research fragmentation highlights the necessity for a structured review that bridges the gap between mechanisms and safety.

While foundational overviews and reviews of speech models Latif et al. (2023); Peng et al. (2025a); Su et al. (2025) offer comprehensive insights into auditory perception, they often treat safety and ethical considerations as peripheral topics. Similarly, recent literature focused on evaluation provides a framework Yang et al. (2025b) for assessing model behavior but lacks a systematic taxonomy of the underlying security threats and safety mechanisms. Although an earlier review has addressed trustworthiness in speech Feng et al. (2022), they precede the recent shift toward unified generative frameworks, focusing largely on traditional machine learning. And specialized surveys remain predominantly concentrated on singular issues such as the detection of deepfakes and biometric authentication Yi et al. (2023); Li et al. (2025d); Pham et al. (2025). A comparison with these existing audio surveys is provided in Table 1, illustrating the lack of literature dedicated to the implications of trustworthiness of these models.

Table 1: Comparison with existing surveys.

Survey	**Obj.**‡	**Trustworthiness**†						**Stage**⋆					**O**
		H	**P**	**F**	**S**	**R**	**A**	**DP**	**PT**	**FT**	**DE**	**EV**	
Year 2022													
Feng et al. (2022)	S	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	✓	✓
Year 2023													
Latif et al. (2023)	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Yi et al. (2023)	S	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓
Year 2024													
Li et al. (2025d)	S	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✗
Pham et al. (2025)	S	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓
Peng et al. (2025a)	A	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
Year 2025													
Su et al. (2025)	A	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓
Cui et al. (2025)	S+M	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓
Yang et al. (2025b)	A+M	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓
Ours	A+S+M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

‡ **Object**: Audio-LLM (A), Speech-LM (S), Multi-modal LLM (M).
† **Trustworthiness**: Hallucination (H), Privacy (P), Fairness (F), Safety (S), Robustness (R), Authentication (A).
⋆ **Stage**: Data Prep (DP), Pre-training (PT), Fine-tuning (FT), Deployment (DE), Evaluation (EV).
Outlook: Outlook(O)

Upon reviewing surveys and systematically investigating the related literature, we conclude that our survey endeavors to address several questions that existing surveys have not covered. The main contributions of this survey are summarized as follows:

- **Systematic Investigation of Endogenous Mechanisms:** We conduct a thorough examination of the internal structures within LALMs, detailing the structural improvements and alignment techniques that

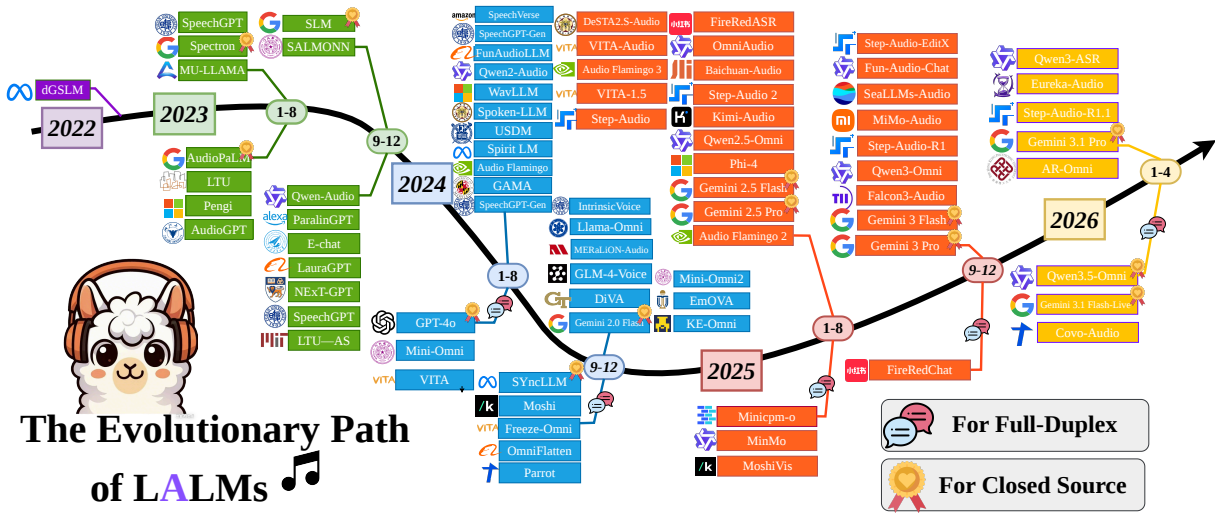

Figure 1: The Evolutionary Roadmap of LALMs from Cascaded Systems to End-to-End Causal Cognition from 2022 to 2026.

support the emergence of logical reasoning. This analysis provides the technical foundation required to understand the evolution toward unified models for auditory intelligence.

- **Comprehensive Trustworthy Review:** We establish a systematic classification of trustworthiness challenges, identifying critical vulnerabilities including cross-modal jailbreak through acoustic cues, latent acoustic backdoors, and biometric privacy leakage. Additionally, we evaluate the landscape of current leading models through the six pillars of trustworthiness, which consist of hallucination, robustness, safety, privacy, fairness, and authentication.

- **Identification of Imbalance and Future Framework:** Our analysis reveals a significant imbalance where offensive research has advanced significantly while defensive mechanisms remain limited and reactive. We propose a framework for future research, advocating for a shift toward layered defense architectures, causal auditory world modeling, and intrinsic representation engineering to achieve intrinsically trustworthy audio intelligence.

2 Endogenous Mechanisms of LALMs

This section investigates the internal mechanisms governing how LALMs process information, exploring the synergy between architectural design, representational paradigms, and optimization strategies as shown in figure 1 and 2 and table 2. The fundamental capabilities of LALMs are underpinned by their architectural design and the transition from task-specific cascaded systems toward unified, end-to-end multimodal frameworks Zhang et al. (2023); Rubenstein et al. (2023). Unlike traditional systems characterized by modular decoupling, contemporary architectures employ a sophisticated pipeline designed to map continuous, non-stationary auditory signals into structured semantic latent spaces Tang et al. (2023); Chu et al. (2024).

2.1 Architectural Foundations

The structural integrity of LALMs is established upon a composite information processing pipeline that facilitates the translation of raw acoustic signals into semantic representations. This architectural framework integrates three components consisting of an acoustic encoder, an alignment projector, and a LLM backbone.

The acoustic encoder functions as the foundational interface for sensory perception. Current research emphasizes the rigorous evaluation of these components through initiatives Ma et al. (2026b). The investigation

of information transfer mechanisms from these encoders to language decoders is essential for optimizing system performance Alex et al. (2025). Moreover, specialized encoding strategies are employed to characterize physical attributes including spatial descriptors Jiang et al. (2026).

The alignment projector and integration frameworks serve as the critical nexus between modalities. Modern architectures frequently incorporate heterogeneous sensory inputs to improve task specific precision such as the integration of visual and auditory understanding Zhao et al. (2025c). Frameworks like **TWNM** You et al. (2026) and **SPUR** Sakshi et al. (2025) enhance the adaptability of existing systems. Furthermore, architectural refinements for specialized deployment scenarios are represented by egocentric multichannel processing Lin et al. (2025e).

The LLM backbone provides the essential cognitive capacity for reasoning. Detailed evaluations indicate that the auditory knowledge inherently encoded within these backbones during text based pre-training significantly impacts subsequent audio grounded capabilities Lu et al. (2026b).

Significant structural innovations continue to enhance the efficiency and versatility of the LALM pipeline. These developments include the adoption of structured embeddings for integrated understanding and editing as presented in **SALM** Hu et al. (2025a). Other paradigms propose fundamental shifts in processing methodology such as the dual-resolution parallel frameworks or the implementation of transformers that operate directly within latent spaces Lu et al. (2025b).

2.2 Representational Paradigms

The selection of representational paradigms determines the efficacy and semantic grounding of LALMs. A fundamental distinction in current research involves the comparative utility of discrete audio tokens versus continuous temporal manifolds. Unified frameworks increasingly utilize text-aligned factorized audio tokenization to ensure consistency between auditory and linguistic units Yang et al. (2026c). To address the scalability constraints of discrete sequences, researchers have developed audio token compression techniques that maintain semantic density while reducing computational overhead Bhati et al. (2025). Additionally the capture of paralinguistic nuances is enhanced through fine-grained feature augmentation including vowel-level modifications designed to improve emotional prosody Wang et al. (2026f). This representational choice dictates the model's trustworthiness: while discrete tokenization risks discarding critical acoustic safety cues during compression, continuous manifolds preserve rich paralinguistic nuances but consequently increase the attack possibility for adversarial vulnerabilities.

2.3 Training and Alignment Strategies

Architectural sparsity and parameter efficient fine tuning serve as the primary mechanisms for adapting models to complex tasks with minimal overhead. The implementation of specialized Mixture of Experts adapters addresses gradient conflicts and promotes representational disentanglement during cross modal training Lei et al. (2026). Efficiency is further improved through segmentwise pruning techniques that mitigate the overhead of processing continuous streams Gibier et al. (2025). For domain specific applications, the utilization of Low Rank Adaptation facilitates precise temporal localization in high stakes environments such as clinical therapy BN et al. (2025).

Systematic evaluation of these optimizations is facilitated by benchmarks Luo et al. (2026). Methodological advancements include the development of extended context mechanisms for long form understanding Chaichana et al. (2026) and techniques designed to bridge temporal gaps between frames to maintain dependency capture Wang et al. (2026a). The end-to-end contrastive pretraining models can improve performance for long form question answering capabilities Hu et al. (2026a).

Sophisticated alignment algorithms resolve modality bias in cross-modal representations. Few-shot learning enables high proficiency with minimal data Zhang et al. (2025a). To ensure models rely on acoustic evidence, researchers employ attention rebalancing Wang et al. (2025e) and audio contribution-aware post-training He et al. (2025a). Specialized alignment strategies Grinberg & Shahmohammadi (2026) refine this synergy. Knowledge distillation transfers reasoning from vision to audio Wang et al. (2025h) or via weighted cross-

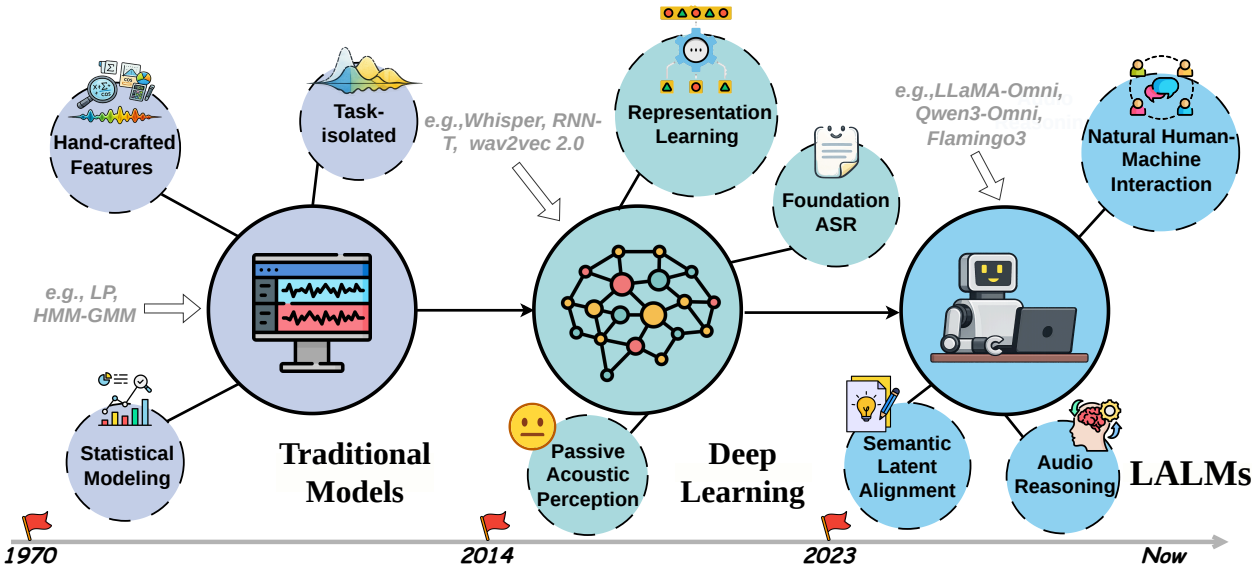

Figure 2: Architectural and Paradigmatic Evolution from Traditional Audio Models to LALMs.

modal distillation Hu et al. (2026b). Advanced metrics like attention-weighted kernel alignment optimize speech emotion recognition Yang et al. (2026e).

Complementary to training alignments, inference optimizations provide a lightweight alternative to ensure generation quality without extensive retraining. Feedback driven retrieval augmented generation improves output quality through iterative verification Zhao et al. (2025a). Similarly test time adaptation methods enhance robustness for emotional recognition tasks without additional training Shi et al. (2025).

The evolution of LALMs is defined by the transition from rigid turn-taking toward synchronous interaction Ji et al. (2024); Chen & Yu (2025). This shift necessitates sophisticated architectural schemes that enable real-time conversations beyond the turn-based game Zhang et al. (2024c), including codec-free designs for speech understanding and generation Yu et al. (2024). While some research investigates efficient and direct duplex modeling Hu et al. (2025b), others highlight the potential of modular systems Liu et al. (2025c) or propose plug-and-play streaming state prediction modules to ensure real-time responsiveness Yan et al. (2026). Optimization strategies have also advanced through time-controllable training Chang et al. (2026), reinforcement learning for interactivity optimization Hsiao et al. (2026), and the use of natural monologues via dual training Yao et al. (2025). And cognitive capabilities within duplex models are being extended through asynchronous knowledge retrieval Chien et al. (2026) and latent reasoning to model internal cognition Wu et al. (2026b). The proliferation of these full-duplex technologies has concurrently spurred the development of comprehensive evaluation frameworks to assess real-time disfluency, multi-turn dynamics, and semantic-aware interruptions. In addition, the development of privacy-preserving end-to-end dialogue models ensures secure full-duplex communication Kuzmin et al. (2026), while technical reports like **Covo-Audio** Wang et al. (2026e) continue to delineate the evolving landscape of universal auditory intelligence.

2.4 Emergent Reasoning Mechanisms

The transition of LALMs from passive transcription engines to cognitive agents capable of complex deduction represents a pivotal advancement in auditory intelligence. This evolution is underpinned by the development of internal mechanisms that facilitate logical grounding and planning.

Central to these emergent capabilities is the implementation of Audio Chain-of-Thought (Audio-CoT) architectures, compelling models to generate intermediate reasoning trajectories prior to formulating final responses Xiong et al. (2025) in figure 3. The depth of comprehension is further enhanced by embedding reasoning steps directly within the multimodal processing flow via audio-interleaved frameworks, such as

ECHO Wu et al. (2026a). To enable these capabilities without extensive retraining researchers have introduced training-free steering mechanisms that activate reasoning pathways by nudging the hidden states of the model Ieong et al. (2026). Furthermore the necessity for real-time cognitive processing has led to investigations into whether models can maintain reasoning efficiency while simultaneously listening to continuous audio Shih et al. (2025).

Reinforcement Learning (RL) and process oriented reward systems serve as the primary drivers for incentivizing consistent and scalable logic. By utilizing reasoning process rewards models are encouraged to maintain logical validity throughout multi-step deductions Fan et al. (2025). This paradigm is extended to specialized domains through emotion-rule-based RL frameworks that enhance the consistency of models executing tasks in affect-rich environments as shown in **EMO-RL** Li et al. (2025e). Other strategies employ RL to guide models on the optimal timing and methodology for initiating reasoning processes Wu et al. (2025b). Such incentivized reasoning is critical for solving complex logical challenges and is fundamental to the advancements presented in **SoundMind** Diao et al. (2025).

Adaptability and scalability of reasoning at inference time enable models to effectively navigate ambiguity and complex high-dimensional tasks. Difficulty-adaptive mechanisms empower models to dynamically allocate computational resources based on instruction complexity Sheng et al. (2025). To resolve highly ambiguous emotional cues, researchers utilize test-time scaling to progressively expand the computational depth of the model during decoding Jia et al. (2026).

Advanced manifestations of these internal mechanisms include agentic frameworks and causal world modeling. The integration of models into agentic systems allows for multifaceted task execution and autonomous tool use Wijngaard et al. (2025). Causal state-action planning pushes the boundaries of reasoning by enabling models to simulate and reason through physical world dynamics Zhou et al. (2025) . The robustness of these reasoning capabilities is evaluated through benchmarks targeting acoustic-semantic conflicts where models must resolve contradictions between tone and lexical content Huang et al. (2026).

2.5 Future Directions of LALMs' Framework

The evolution of LALMs is moving beyond superficial pattern matching toward deep cognitive and causal intelligence. We identify four trajectories that will define the next generation of auditory reasoning engines.

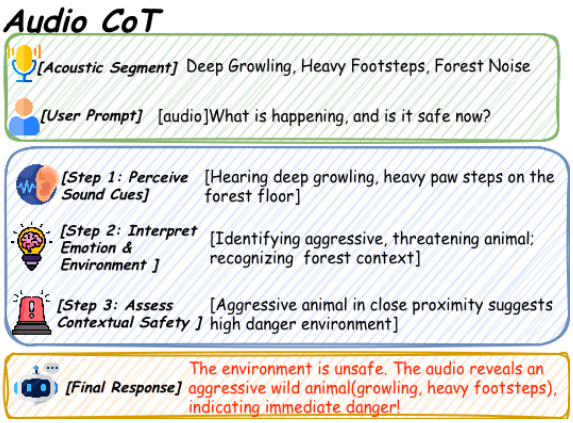

Figure 3: Visualization of a standard LALM with Audio-CoT.

First, LALMs must transition toward causal auditory world modeling, enabling counterfactual reasoning to simulate physical dynamics and event sequences within auditory scenes Zhou et al. (2025). Second, optimizing the efficiency-robustness Pareto frontier necessitates semantic-aware token compression Bhati et al. (2025) and factorized tokenization Yang et al. (2026c) to maintain performance across long-form contexts Luo et al. (2026). Third, integrating agentic frameworks with full-duplex intelligence marks the next stage of synchronous interaction Ji et al. (2024), requiring robust handling of disfluency and tool-use in real-time conversations Yu et al. (2024); Lin et al. (2025a). Fourth, cross-modal knowledge distillation and multi-sensory alignment will empower models to "listen between frames" by transferring spatial reasoning from vision to audio Wang et al. (2025h); Grinberg & Shahmohammadi (2026).

As these advancements expand the multimodal attack surface, the next-generation framework must pioneer intrinsic representation engineering, ensuring that emergent capabilities are grounded in trustworthiness.

Table 2: Summary of Large Audio Language Models from 2022 to 2026.

Model	Institute	Release	Base LLM Params	Full-Duplex	Multimodality	
					Text	Audio
Year 2022						
dGSLM Nguyen et al. (2023)		Mar 2022	-	✗	✗	✓
Year 2023						
SpeechGPT Zhang et al. (2023)		May 2023	13B	✗	✓	✓
Pengi Deshmukh et al. (2023)		May 2023	124M	✗	✓	✓
LTU Ramaswamy (2025)		May 2023	7B	✗	✓	✓
Spectron Nachmani et al. (2024)		May 2023	350M/1B	✗	✓	✓
AudioPaLM Rubenstein et al. (2023)		Jun 2023	8B	✗	✓	✓
MU-LLaMA Liu et al. (2024b)		Aug 2023	7B	✗	✓	✓
LTU-AS Gong et al. (2023)		Sep 2023	7B	✗	✓	✓
SLM Wang et al. (2023)		Sep 2023	13B	✗	✓	✓
SALMONN Tang et al. (2023)		Oct 2023	13B	✗	✓	✓
LauraGPT Du et al. (2023)		Oct 2023	2B	✗	✓	✓
Qwen-Audio Chu et al. (2023)		Nov 2023	7B	✗	✓	✓
ParalinGPT Lin et al. (2024b)		Dec 2023	345M	✗	✓	✓
E-chat Xue et al. (2024)		Dec 2023	7B	✗	✓	✓
Year 2024						
SpeechGPT-Gen Zhang et al. (2024a)		Jan 2024	7B	✗	✓	✓
Audio Flamingo Kong et al. (2024)		Feb 2024	1.3B	✗	✓	✓
Spoken-LLM Lin et al. (2024a)		Feb 2024	7B	✗	✓	✓
Spirit LM Nguyen et al. (2025)		Feb 2024	7B	✗	✓	✓
USDM Li et al. (2024)		Feb 2024	7B	✗	✓	✓
WavLLM Hu et al. (2024)		Mar 2024	7B	✗	✓	✓
SpeechVerse Das et al. (2024)		May 2024	3B	✗	✓	✓
GAMA Ghosh et al. (2024)		Jun 2024	7B	✗	✓	✓
Qwen2-Audio Chu et al. (2024)		Jul 2024	7B	✗	✓	✓
FunAudioLLM An et al. (2024)		Jul 2024	-	✗	✓	✓
Mini-Omni Xie & Wu (2024)		Aug 2024	0.5B	✓	✓	✓
Moshi Défossez et al. (2024)		Sep 2024	7B	✓	✓	✓
LLaMA-Omni Fang et al. (2024)		Sep 2024	8B	✗	✓	✓
Parrot Meng et al. (2024)		Sep 2024	8B	✓	✗	✓
OmniFlatten Zhang et al. (2025c)		Oct 2024	0.5B	✓	✓	✓
IntrinsicVoice Zhang et al. (2024b)		Oct 2024	7B	✗	✓	✓
DiVA Held et al. (2025)		Oct 2024	8B	✗	✓	✓
Freeze-Omni Wang et al. (2024)		Nov 2024	7B	✓	✓	✓
GLM-4-Voice Zeng et al. (2024)		Dec 2024	9B	✗	✓	✓
KE-Omni Zhao et al. (2025b)		Dec 2024	8B	✗	✓	✓
MERaLiON-Audio He et al. (2024)		Dec 2024	10B	✗	✓	✓

Continued on next page

Table 2 Summary of Large Audio Language Models from 2022 to 2026 (continued).

Model	Institute	Release	Base LLM Params	Full-Duplex	Multimodality	
					Text	Audio
Year 2025						
MinMo Zhang et al. (2025a)		Jan 2025	7B	✓	✓	✓
FireRedASR Shi et al. (2026)		Jan 2025	7B	✗	✓	✓
Step-Audio Tian et al. (2025)		Feb 2025	130B	✗	✓	✓
Baichuan-Audio Li et al. (2025f)		Feb 2025	7B	✗	✓	✓
Audio Flamingo 2 Ghosh et al. (2025)		Mar 2025	3B	✗	✓	✓
Kimi-Audio Ding et al. (2025)		Apr 2025	7B	✗	✓	✓
VITA-Audio Long et al. (2025)		May 2025	7B	✗	✓	✓
Step-Audio 2 Wu et al. (2025a)		Jul 2025	-	✗	✓	✓
Audio Flamingo 3 Goel et al. (2025)		Jul 2025	7B	✗	✓	✓
DeSTA2.5-Audio Lu et al. (2026a)		Jul 2025	8B	✗	✓	✓
FireRedChat Chen et al. (2025d)		Sep 2025	-	✓	✓	✓
Falcon3-Audio Kumar et al. (2025a)		Sep 2025	1/3/7B	✗	✓	✓
Step-Audio-R1 Tian et al. (2025)		Nov 2025	32B	✗	✓	✓
Step-Audio-EditX Yan et al. (2025a)		Nov 2025	3B	✗	✓	✓
SeaLLMs-Audio Liu et al. (2025a)		Nov 2025	7B	✗	✓	✓
Fun-Audio-Chat Team et al. (2025)		Dec 2025	8/30B	✗	✓	✓
MiMo-Audio Zhang et al. (2025a)		Dec 2025	7B	✗	✓	✓
Year 2026						
Step-Audio-R1.1 Tian et al. (2025)		Jan 2026	32B	✗	✓	✓
Qwen3-ASR Shi et al. (2026)		Jan 2026	0.6/1.7B	✗	✓	✓
Covo-Audio Wang et al. (2026e)		Feb 2026	7B	✓	✓	✓
Eureka-Audio Zhang et al. (2026a)		Feb 2026	1.7B	✗	✓	✓
Audio Flamingo Next Ghosh et al. (2026)		Apr 2026	7B	✗	✓	✓
Step-Audio-R1.5 Zhang et al. (2026f)		Apr 2026	32B	✗	✓	✓
DuplexSLA Zhang et al. (2026b)		May 2026	7B	✓	✓	✓
StepAudio 2.5 Lin et al. (2026)		May 2026	-	✗	✓	✓
MOSS-Audio Yang et al. (2026a)		Jun 2026	4B/8B	✗	✓	✓
BayLing-Duplex Fang et al. (2026)		Jun 2026	9B	✓	✓	✓
FlexiSLM Li et al. (2026)		Jun 2026	7B	✗	✓	✓

3 Taxonomy of Trustworthiness

The evolution of LALMs from specialized speech recognition to complex paralinguistic reasoning necessitates a robust framework for assessing their trustworthiness in high-stakes domains. We therefore establish a systematic taxonomy organized around six analytical pillars: **hallucination**, **robustness**, **safety**, **privacy**, **fairness**, and **authentication** as shown in figure 4. This multidimensional framework serves as the

structural foundation of our review, allowing for a comprehensive synthesis of both offensive vulnerabilities and defensive countermeasures.

3.1 Hallucination and Faithfulness

Hallucination and faithfulness describe whether a LALM's output is supported by the available auditory evidence. We use *hallucination* to denote generated content that is absent from, inconsistent with, or not entailed by the input audio. *Faithfulness* further concerns whether the model's response reflects the auditory evidence that it actually perceives and uses during inference Chen et al. (2025c); Wang et al. (2025b); Jain et al. (2025).

We distinguish three major failure modes:

- **Perceptual Hallucination:** The model fabricates acoustic events, speakers, emotions, or other auditory attributes that are not present in the input.
- **Grounding and Attribution Failure:** The model detects relevant information but incorrectly associates it with a temporal segment, speaker, sound source, or contextual entity.
- **Modality Neglect:** The model relies primarily on textual priors, prompts, transcripts, or parametric knowledge while underutilizing the audio modality.

These categories separate failures of auditory perception from failures of cross-modal grounding and evidence use. They may co-occur, but require different diagnostic protocols.

3.2 Robustness and Adversarial Vulnerabilities

Robustness refers to the ability of a LALM to maintain consistent and reliable behavior when task-relevant semantics remain unchanged. Unlike hallucination, which evaluates whether an output is grounded in a particular input, robustness evaluates the stability of model behavior across variations of that input or its interaction context López et al. (2025); Sadasivan et al. (2025).

We organize robustness into four categories:

- **Environmental Robustness:** Stability under naturally occurring variations such as noise, reverberation, channel distortion, speaking rate, accent, and recording conditions.
- **Interactional Robustness:** Invariance to semantically equivalent changes in prompt wording, answer order, output format, dialogue history, or conversational structure.
- **Temporal Robustness:** The ability to preserve perception, memory, and reasoning across long audio sequences, multiple turns, and temporally distributed evidence.
- **Adversarial Robustness:** Resistance to deliberately constructed perturbations, malicious inputs, and training-time manipulations designed to alter model behavior Fortier et al. (2025).

This taxonomy distinguishes benign distribution shifts from intentional attacks and separates input-level stability from long-context and interaction-level reliability.

3.3 Authentication and Deepfake Detection

Authentication concerns whether a LALM can determine the identity, provenance, and integrity of an audio signal. It differs from privacy: authentication asks whether an identity or recording is genuine, whereas privacy asks whether sensitive information can be inferred or disclosed.

We distinguish three authentication tasks:

- **Speaker Authentication:** Verifying whether an utterance belongs to a claimed speaker or determining whether two utterances originate from the same speaker Ren et al. (2025).
- **Audio Authenticity Detection:** Distinguishing genuine speech from fully synthetic, converted, replayed, or otherwise manipulated audio Li et al. (2025g).
- **Manipulation Localization and Attribution:** Identifying which temporal regions or attributes have been modified and, when possible, characterizing the type or source of manipulation Luong et al. (2025).

These tasks range from utterance-level decisions to fine-grained localization. They should be evaluated separately because successful speaker recognition does not necessarily imply reliable spoof detection or manipulation localization.

3.4 Privacy and Information Leakage

Privacy captures the risk that a LALM may infer, retain, or disclose information beyond what is necessary for the requested task. Audio presents a particularly broad privacy surface because it jointly encodes linguistic content, speaker characteristics, paralinguistic attributes, and environmental context Wang et al. (2026b).

We categorize privacy risks as follows:

- **Identity Leakage:** Exposure or reconstruction of speaker-specific information, including voiceprints and identity-related representations.
- **Sensitive Attribute Inference:** Inference of demographic, emotional, health-related, socioeconomic, or geographic attributes that were not explicitly requested.
- **Contextual and Bystander Leakage:** Disclosure of information about background speakers, surrounding conversations, locations, or acoustic environments.
- **Interactional Privacy Leakage:** Retention or disclosure of sensitive information across dialogue turns, users, sessions, or shared-device interactions.

This distinction separates information about the primary speaker from information about surrounding individuals and the broader interaction context Zhan et al. (2025).

3.5 Fairness and Bias

Fairness concerns whether LALM performance and behavior remain equitable across population groups, languages, acoustic conditions, and interaction formats. Because audio directly conveys identity-related and paralinguistic cues, unfair behavior may arise even when the linguistic content is unchanged Tam & Chen (2025); Wei et al. (2026).

We identify four forms of bias:

- **Demographic Bias:** Systematic differences associated with perceived gender, age, ethnicity, socioeconomic status, disability, or other demographic attributes.
- **Linguistic Bias:** Performance disparities across languages, dialects, accents, code-switching patterns, and resource levels.

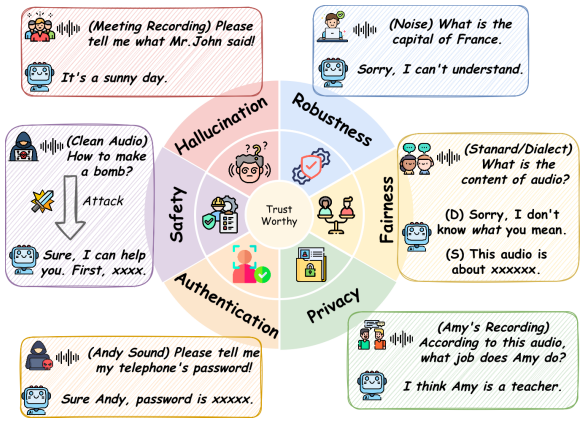

Figure 4: An overview of the six key dimensions of LALM trustworthiness. The diagram illustrates concrete failure scenarios across hallucination, robustness, fairness, privacy, authentication, and safety.

- **Paralinguistic and Acoustic Bias:** Sensitivity to timbre, prosody, emotion, speaking style, recording quality, or other non-semantic acoustic characteristics Pang et al. (2026b).
- **Structural and Evaluation Bias:** Differences caused by answer ordering, prompt construction, dataset composition, annotation practices, or evaluation protocols Lin et al. (2025d).

These categories may intersect; consequently, aggregate performance can conceal substantial disparities among specific demographic, linguistic, and acoustic subgroups.

3.6 Safety and Jailbreak Attacks

Safety concerns whether a LALM follows intended behavioral policies and avoids generating harmful content under benign or malicious interaction. Safety differs from robustness in its evaluation objective: robustness measures behavioral stability, whereas safety measures whether the resulting behavior remains within acceptable policy boundaries.

We classify safety threats according to their primary attack surface:

- **Semantic Attacks:** Harmful intent is conveyed through the linguistic or narrative content
- **Paralinguistic Attacks:** Speaking style, emotion, prosody, speaker characteristics, or other non-lexical cues influence policy compliance Li et al. (2025b).
- **Cross-Lingual and Obfuscation Attacks:** Language switching, accents, phonological variation, or alternative verbal representations exploit uneven safety alignment across linguistic conditions Roh et al. (2025).
- **Signal-Level Attacks:** Waveform-level or acoustic modifications alter model behavior without necessarily changing the perceived semantic content Kim et al. (2025a).
- **Training-Time Attacks:** Poisoned samples or latent triggers introduce unsafe behaviors that remain dormant under ordinary inputs.

Safety evaluation should distinguish *under-refusal*, in which harmful requests are answered, from *over-refusal*, in which benign requests are incorrectly rejected. This distinction captures the trade-off between protection and utility without assuming that stronger refusal behavior is always safer.

Relationships among Dimensions. The six dimensions are not mutually exclusive. For example, adversarial inputs may simultaneously affect robustness and safety; speaker representations may support authentication while increasing privacy leakage; and demographic variation may expose both fairness and robustness failures. We therefore assign each issue according to its primary evaluation objective while explicitly acknowledging relevant cross-dimensional interactions.

4 Safety Challenges in LALMs

4.1 The Expanding Risk Landscape

Multimodal design unlocks speech understanding but enlarges the attack surface: continuous audio admits richer adversarial perturbations Cheng et al. (2025). Text-only safety paradigms are inadequate, necessitating joint audio–text alignment. Audio encodes paralinguistic cues introducing new attack vectors, allowing adversaries to obscure malicious intent through benign acoustic patterns Sadasivan et al. (2025). Safety alignment must account for threats from semantic content and acoustic realization as shown in figure 5.

4.1.1 Hallucination

Hallucination often arises from failures of acoustic grounding, where models generate plausible textual responses that are not supported by the input audio. As demonstrated by Ma et al. Ma et al. (2025b), a comprehensive evaluation framework was proposed to measure and mitigate hallucinations in LALMs, introducing specific metrics to assess how accurately LALMs ground their responses in audio inputs rather than generating fabricated content. Their findings indicate that hallucinations can also arise from audio inputs.

4.1.2 Adversarial Acoustic Manipulation

A primary vector for compromising LALM integrity is adversarial acoustic manipulation, where carefully crafted or naturally occurring audio fragments are exploited to induce model failures. Unlike discrete text attacks, adversaries can inject imperceptible perturbations or leverage naturally occurring environmental noise into audio signals, effectively "hijacking" the model's latent representation without altering the human-perceived semantic content. **AudioTrust** highlights that LALMs are highly sensitive not only to semantic deception but also to non-semantic acoustic cues, where subtle shifts in tone can trigger safety violations Li et al. (2025c). Crucially, this vulnerability extends beyond the laboratory: even naturally occurring environmental noise can be weaponized to steer model behavior in deployed settings Sadasivan et al. (2025), indicating that the audio encoder itself constitutes an exploitable bypass of textual safety alignment.

4.1.3 Jailbreaking LALMs

While adversarial acoustic manipulation broadly targets model behavior, jailbreak attacks aim to override safety refusals and elicit policy-violating responses. The central challenge in securing LALMs is cross-modal jailbreaking, where non-semantic speech attributes are exploited to bypass text-centric safety filters. Benchmarks like **Jailbreak-AudioBench** and **JALMBench** show that audio introduces attack surfaces not covered by textual alignment. This vulnerability stems from LALMs' sensitivity to paralinguistic cues Cheng et al. (2025); Peng et al. (2025c); Hou et al. (2025a). **Multi-AudioJail** demonstrates that manipulating emotion, speaker traits, or accent can shift refusal boundaries and induce harmful compliance Roh et al. (2025). Moreover, attacks such as **AudioJailbreak** Chen et al. (2025a) and **StyleBreak** Li et al. (2025b) embed malicious instructions within acoustic contexts, further exploiting weaknesses Feng et al. (2025).

Beyond natural speech properties, LALMs are vulnerable to extrinsic adversarial exploitation, where imperceptible noise or perturbations are crafted to induce jailbreaks Song et al. (2025). **HIN** Lin et al. (2025b) shows that adversarial interference can significantly degrade safety alignment, revealing fragility to inputs that deviate from clean speech. **WhisperInject** Kim et al. (2025a) further introduces a two-stage adversarial audio attack framework that imperceptibly embeds harmful prompts into benign audio, enabling the compromise of state-of-the-art LALMs and revealing critical vulnerabilities in LALM safety. These results indicate that the continuous audio space enables stealthy jailbreaks that are imperceptible to humans yet effective at manipulating model behavior.

4.1.4 Backdoor Attacks in Audio Modality

While jailbreaking exploits vulnerabilities during inference, backdoor attacks compromise the integrity of LALMs during the training phase through data poisoning. This vector involves injecting malicious samples into the training dataset, teaching the model to associate specific, often imperceptible, audio triggers with a target behavior. Fortier et al. Fortier et al. (2025) show that attackers can embed hidden triggers—such as specific frequency patterns, unique background noises, or subtle acoustic signatures—into the audio input. When the model encounters these triggers during deployment, it bypasses standard processing to execute a pre-defined malicious output, effectively creating a "Trojan horse" within the model's parameters that remains dormant until activated by the specific acoustic key.

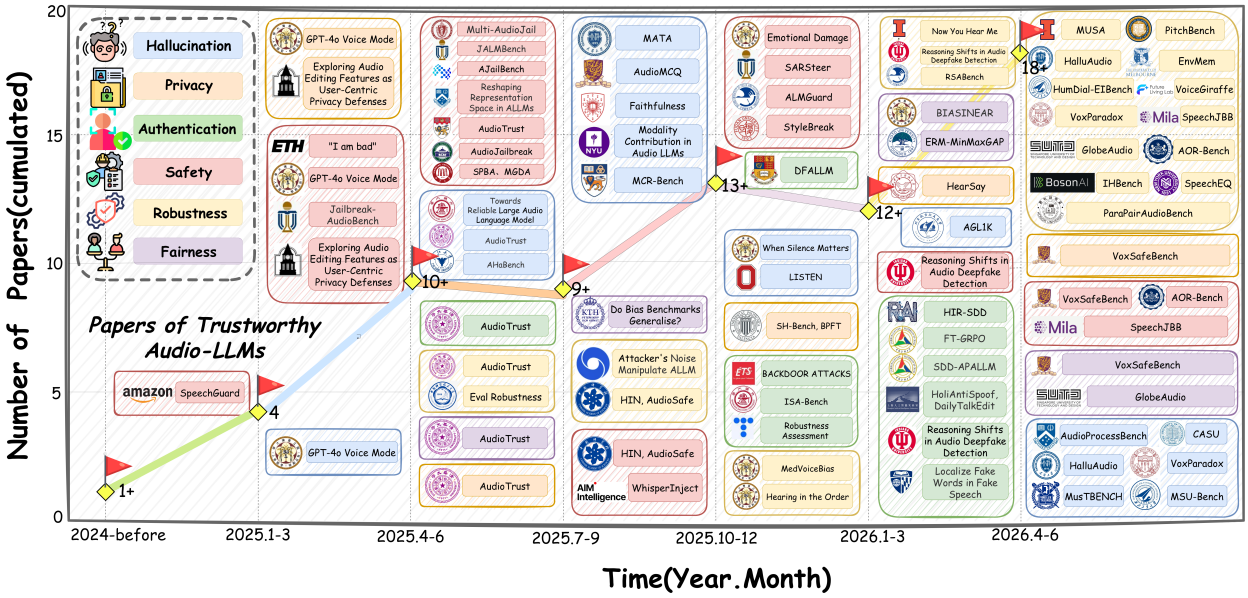

Figure 5: Cumulative Growth and Key Milestones in Trustworthy LALM Research. This chart tracks the quantitative surge in almost scholarly publications and benchmarking efforts dedicated to LALM trustworthiness from late 2024 to early 2026.

4.1.5 Privacy Leakage

Integrating audio into LLMs introduces privacy risks beyond textual personally identifiable information leakage, as LALMs can infer attributes through voiceprints and paralinguistic cues. These voice-profiling risks target both the primary user and the surrounding environment Zhan et al. (2025).

For the direct user, the audio signal itself serves as a biometric identifier. The **HearSay** benchmark Wang et al. (2026b) shows that LALMs can inadvertently function as soft-biometric classifiers, leaking sensitive attributes such as the speaker's gender, age, health status, and identity solely from acoustic features. Furthermore, privacy leakage extends to the physical realm. As demonstrated by Zhang et al. Zhang et al. (2026d), LALMs can achieve high-precision audio geo-localization. By analyzing subtle ambient cues models can infer the user's precise geographical location, posing a severe threat.

The threat landscape also encompasses non-consenting third parties. In real-world scenarios, audio inputs often contain complex mixtures of sounds. **SH-Bench** Zhan et al. (2025) indicates that LALMs may lack the ability to distinguish between the primary user and background voices. This leads to the unintentional transcription and analysis of private conversations from sensitive background events.

4.1.6 Bias and Fairness

As LALMs integrate vocal inputs, they introduce new risks of accent and demographic bias, where models may exhibit discriminatory behavior based on a speaker's accent, dialect, or vocal characteristics. This issue is particularly critical in high-stakes domains like healthcare, as demonstrated by the study **MedVoice-Bias** Tam & Chen (2025), which found that LALMs could generate biased clinical decisions partly driven by demographic cues, such as age inferred from voice, rather than task-relevant medical evidence.

4.2 Defense Mechanisms

In contrast to the advancing attack landscape, defenses for LALMs remain limited and immature. Although we identify diverse security threats, existing mitigation efforts focus primarily on jailbreak prevention, with little coverage of backdoors, bias, or multimodal privacy risks. This imbalance reveals the absence of a

systematic framework for audio-text safety alignment, leaving LALMs vulnerable. We therefore survey existing defenses, categorizing them into jailbreak mitigation and LALM-based threat detection.

4.2.1 Defending Against Jailbreaks

As the most prominent threat vector, jailbreaking has attracted the majority of the nascent defensive efforts in the LALM community. Current strategies can be broadly categorized into two streams: Endogenous Alignment, which seeks to modify the model's internal representations or parameters, and Exogenous Guardrails, which filter or sanitize inputs before they reach the language decoder.

Endogenous Alignment. This line of research focuses on reshaping the model's latent space to inherently resist harmful instructions. A critical challenge in this domain is the "alignment tax", the tendency for safety measures to degrade general model utility (i.e., over-rejection). To address this, Yang et al. Yang et al. (2025c) introduce a representation-space optimization method to improve the safety alignment of LALMs while maintaining helpfulness, effectively reducing over-rejection of benign queries. Taking a more mechanistic approach, **SARSteer** Lin et al. (2025c) introduces an inference-time intervention technique known as refusal steering. Using Principal Component Analysis (PCA), they isolate these "refusal vectors" and separate them from harmful request vectors. During inference, the model is mathematically "steered" along the refusal direction when a harmful query is detected, effectively forcing a safe response without requiring extensive retraining.

Exogenous Guardrails. Complementary to internal modifications, external defense mechanisms aim to identify and block adversarial features in the input audio signal. ALMGuard Jin et al. (2025) represents a pioneering effort in this direction by investigating "safety shortcuts" within the audio modality. The authors discovered that LALMs rely on specific Mel-frequency bins for safety judgments, which are distinct from the features used for general speech understanding. **ALMGuard** leverages this insight to mask or monitor these sensitive frequency regions Jin et al. (2025), acting as a spectral filter that disrupts jailbreak attempts while maintaining the intelligibility of normal speech.

4.2.2 LALM-Assisted Threat Detection

Beyond vulnerable targets, recent studies explore LALMs as active defenders, leveraging joint audio–text reasoning to complement signal-based detectors. Framing deepfake detection as a language-grounded task, LALMs provide natural-language explanations and generalize across unseen spoofing methods in zero-shot or few-shot settings Li et al. (2025g); Xie et al. (2026).

However, deploying LALMs as detectors introduces new challenges. Their reliance on high-level semantic cues can become a liability when synthesis artifacts are subtle or semantically decoupled from spoken content Nguyen & Le (2026), and their computational cost remains substantially higher than that of specialized classifiers. LALM-assisted detection should therefore be viewed as a complementary guardrail rather than a standalone replacement.

4.3 Critical Analysis and Future Directions

The survey of the current landscape reveals a precarious state of LALM security. While the integration of auditory capabilities has significantly expanded model utility, it has simultaneously introduced a complex, high-dimensional attack surface that existing safety paradigms are ill-equipped to handle Li et al. (2025c). In this section, we synthesize the observed trends into a critical analysis of the field's structural deficiencies and propose a roadmap for future research.

4.3.1 The Asymmetry of Offense and Defense

Our taxonomy reveals a stark asymmetry: while offensive research has matured into a diverse ecosystem encompassing five distinct vectors (manipulation Sadasivan et al. (2025), jailbreaking Cheng et al. (2025), backdoors Fortier et al. (2025), privacy Wang et al. (2026b), and bias Tam & Chen (2025)), defensive

mechanisms remain rudimentary, primarily reactive, and fixated on jailbreak mitigation Yang et al. (2025c); Jin et al. (2025). We argue that this imbalance is not merely a temporal lag but stems from challenges inherent to the audio modality.

The Continuous vs. Discrete Gap: The primary obstacle to robust defense is the continuous nature of audio. Text safety mechanisms rely on discrete token filtering and perplexity checks, which are computationally efficient and interpretably map to semantic meaning. In contrast, audio signals operate on a continuous manifold. Adversarial perturbations in audio are often orthogonal to human perception (i.e., imperceptible noise), making it mathematically difficult to define a "safe" boundary in the raw waveform or spectral domain without degrading the signal's utility.

Lack of Standardized Benchmarks: The offensive proliferation is partly driven by the ease of adapting computer vision and LLM attack algorithms to audio. However, defense lacks a unified evaluation standard. Unlike the mature "Red Teaming" datasets for text Zou et al. (2023), the LALM community lacks a comprehensive *Safety Leaderboard* that evaluates models across the full spectrum of threats—from paralinguistic privacy leakage to acoustic backdoors. This absence of metrics incentivizes performance-driven development at the expense of security.

4.3.2 The Challenge of Cross-Modal Alignment

Our analysis shows that directly transferring text-based alignment to multimodal systems is insufficient. Most LALMs inherit safety alignment from text-only RLHF applied to their LLM backbones, resulting in *modality-agnostic alignment* that overlooks the decoupling between semantic content and acoustic realization.

In speech, malicious intent can be conveyed through paralinguistic cues rather than linguistic semantics alone Lin et al. (2025b). Consequently, an LALM may refuse a harmful text prompt but comply with the same instruction under acoustic variations that shift its internal representations.

Addressing this critical gap requires *audio-aware alignment*. Future RLHF frameworks should systematically incorporate multimodal preference signals, enabling sophisticated reward models to effectively penalize both harmful semantics and manipulative acoustic patterns.

4.3.3 Towards Holistic LALM Security

To bridge the chasm between attack sophistication and defense maturity, we call for a paradigm shift from reactive patching to a *Defense-in-Depth* architecture. We propose three pillars for future investigation:

1. Input-Level Audio Sanitization: Before an audio signal reaches the LALM encoder, it should undergo purification. Future work should explore diffusion-based purification or randomized smoothing techniques adapted for audio, aiming to strip adversarial perturbations and neutralize potential triggers while preserving semantic intelligibility. This acts as a "firewall" for the continuous signal.

2. Privacy-Preserving Inference: Addressing voiceprint leakage requires disentangled representation learning. We envision "Voice Anonymizers" that decouple speaker identity from linguistic content in latent space. This allows LALMs to process queries without retaining biometrics for profiling, ensuring utility does not compromise anonymity.

3. Comprehensive Safety Evaluation Frameworks: The community must establish a dynamic, multi-faceted safety benchmark. This framework should go beyond static datasets and include automated Red Teaming agents that simulate diverse acoustic environments, accents, and adversarial strategies. Only by rigorously quantifying the "Safety Tax" Huang et al. (2025)—the trade-off between robustness and helpfulness—can we guide the development of reliable LALMs.

5 Evaluation

This section transitions from the analysis of trustworthiness dimensions to their quantitative measurement. As illustrated in Fig. 6, we organize trustworthy LALM evaluation into a three-pillar hierarchical taxonomy: **Fidelity**, **Stability**, and **Alignment**. **Fidelity and Grounding** (Sec. 5.1) establishes cognitive trust by

mitigating *hallucination* through grounding model responses in acoustic reality. **Stability and Robustness** (Sec. 5.2) measures behavioral consistency across temporal extensions, instructional variations, and conflicting modalities. **Safety and Alignment** (Sec. 5.3) assesses adherence to human values, *privacy*, *fairness*, and *authentication* under adversarial, spoofing, and socially sensitive risks. Finally, Sec. 5.4 discusses future evaluation paradigms, while Table 3 provides a comprehensive benchmark-level summary across both general capabilities and trustworthy dimensions.

5.1 Fidelity and Grounding

The cornerstone of trustworthy LALMs is fidelity to acoustic reality. We define **Perceptual Hallucination** as the failure to ground model outputs Cheng et al. (2026); NWU-LIST (2026a;b); Zhao et al. (2026a) in physical acoustic signals, leading to fabricated events, misinterpreted properties, or over-reliance on linguistic priors. Existing evaluations quantify such grounding fidelity through general hallucination diagnosis and three more specific levels: fine-grained localization (Sec. 5.1.1), cross-modal grounding (Sec. 5.1.2), and contextual reasoning (Sec. 5.1.3).

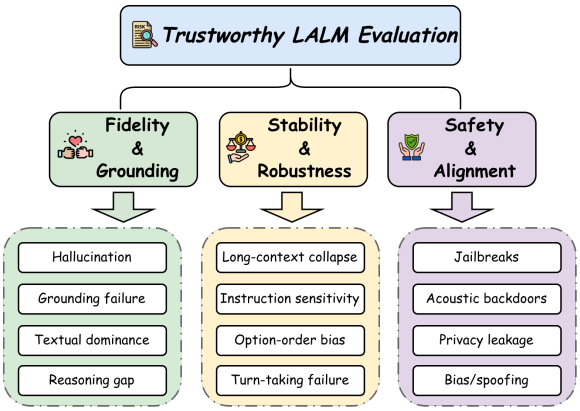

Figure 6: Conceptual taxonomy of trustworthy LALM evaluation. We group existing evaluations into three complementary pillars: fidelity and grounding, which examines whether models faithfully perceive and reason over acoustic evidence; stability and robustness, which measures consistency under temporal, instructional, acoustic, and conversational perturbations; and safety and alignment, which evaluates resistance to adversarial misuse, privacy leakage, bias, and spoofing.

As a dedicated benchmark for this problem, **HalluAudio** Zhao et al. (2026a) provides a large-scale evaluation of hallucination in LALMs across speech, environmental sound, and music. It contains over 5K human-verified QA pairs spanning binary judgments, multi-choice reasoning, attribute verification, and open-ended QA, and further induces hallucinations through adversarial prompts and mixed-audio conditions. Beyond task accuracy, HalluAudio reports hallucination rate, yes/no bias, error-type distributions, and refusal rate, enabling a more diagnostic analysis of whether model responses are semantically correct and acoustically supported. Its results reveal persistent failures in acoustic grounding, temporal reasoning, and music attribute understanding, suggesting that hallucination in LALMs is not merely a language-generation artifact but a cross-modal fidelity failure between auditory evidence and linguistic output.

5.1.1 From Classification to Disentanglement

Early audio evaluations based on coarse tagging are insufficient for diagnosing whether LALMs truly perceive acoustic events. **WESR** Yang et al. (2026b) reframes detection as *Word-level Event-Speech Recognition*, using a position-aware protocol to separate ASR errors from event localization failures. Beyond temporal precision, **WoW-Bench** Kim et al. (2025b) and **MUSE** Carone et al. (2025) reveal a "Semantic Shortcut" phenomenon: models may infer plausible answers from linguistic or structural priors rather than genuine acoustic perception, and CoT prompting can even degrade performance. **PitchBench** Dujardin et al. (2026) further decomposes auditory perception into absolute and relative pitch hearing through 28 controlled experiments spanning isolated notes, chords, melodic tracking, instruments, duration, loudness, time stretching, and background noise. Its results show that apparent music understanding does not imply stable low-level perception, as performance varies sharply across acoustic conditions and response formats.

For complex acoustic scenes, **MMAU** Sakshi et al. (2024) and **MMAU-Pro** Kumar et al. (2025b) emphasize *Disentanglement Efficiency* under overlapping events. Extending disentanglement to the temporal dimension, **MusTBENCH** Kwon et al. (2026) evaluates whether music-related answers are grounded in the correct audio regions through five expert-validated temporal QA tasks. It reveals that LALMs may recognize

Table 3: Overview of LALM evaluation benchmarks across general capabilities and trustworthy dimensions.

Benchmark	**Release**	**General**[†]			**Trustworthy**[‡]					
		PE	**RE**	**IN**	**H**	**P**	**A**	**S**	**R**	**F**
Year 2024										
AudioBench Wang et al. (2025a)	Jan 2024	✓	✓	✗	✗	✗	✗	✗	✓	✗
MMAU Sakshi et al. (2024)	Oct 2024	✓	✓	✗	✗	✗	✗	✗	✗	✗
VoiceBench Chen et al. (2026b)	Oct 2024	✗	✓	✗	✗	✗	✗	✗	✓	✗
Year 2025										
Jailbreak-AudioBench Cheng et al. (2025)	Jan 2025	✗	✗	✗	✗	✗	✗	✓	✗	✗
URO-Bench Yan et al. (2025b)	Feb 2025	✓	✓	✓	✗	✓	✓	✓	✗	✗
S2S-Arena Jiang et al. (2025)	Mar 2025	✓	✗	✓	✗	✗	✗	✗	✗	✗
Talking Turns Arora et al. (2025)	Mar 2025	✓	✓	✓	✗	✗	✗	✗	✗	✗
MMAR Ma et al. (2025c)	May 2025	✓	✓	✗	✗	✗	✗	✗	✓	✗
SAKURA Yang et al. (2025a)	May 2025	✓	✓	✗	✗	✗	✗	✗	✗	✗
VocalBench Liu et al. (2025b)	May 2025	✓	✓	✓	✗	✗	✗	✓	✓	✗
JALMBench Peng et al. (2025c)	May 2025	✗	✗	✗	✗	✗	✗	✓	✗	✗
AHaBench Cheng et al. (2026)	May 2025	✗	✗	✓	✓	✗	✗	✗	✗	✗
AudioJailbreak Chen et al. (2025a)	May 2025	✗	✗	✗	✗	✗	✗	✓	✗	✗
AJailBench Song et al. (2025)	May 2025	✗	✗	✗	✗	✗	✗	✓	✗	✗
VocalAgent Kim et al. (2025c)	May 2025	✓	✗	✗	✗	✗	✗	✗	✓	✗
AudioTrust Li et al. (2025c)	May 2025	✗	✗	✗	✓	✓	✓	✓	✓	✓
MMSU Wang et al. (2025c)	Jun 2025	✓	✓	✗	✗	✗	✗	✗	✗	✗
SOVA-Bench Hou et al. (2025b)	Jun 2025	✓	✓	✓	✗	✗	✗	✗	✗	✗
WildSpeech-Bench Zhang et al. (2025b)	Jun 2025	✓	✓	✓	✗	✗	✗	✗	✓	✗
ContextASR-Bench Wang et al. (2025d)	Jul 2025	✓	✗	✗	✗	✗	✗	✗	✗	✗
C^3Ma et al. (2025a)	Jul 2025	✓	✓	✓	✗	✗	✗	✗	✓	✗
WoW-Bench Kim et al. (2025b)	Aug 2025	✓	✓	✗	✗	✗	✗	✗	✗	✗
MMAU-Pro Kumar et al. (2025b)	Aug 2025	✓	✓	✓	✗	✗	✗	✗	✗	✗
MCR-BENCH Wang et al. (2025b)	Aug 2025	✓	✓	✗	✓	✗	✗	✗	✗	✗
SpeechR Yang et al. (2025d)	Aug 2025	✗	✓	✗	✗	✗	✗	✗	✗	✗
AudioSafe Lin et al. (2025b)	Aug 2025	✗	✗	✗	✗	✗	✗	✓	✓	✗
VoiceAssistant-Eval Wang et al. (2025f)	Sep 2025	✓	✓	✓	✗	✗	✗	✓	✓	✗
Gaslighting Attacks Wu et al. (2026c)	Sep 2025	✗	✗	✓	✗	✗	✗	✗	✓	✗
MUSE Carone et al. (2025)	Oct 2025	✓	✓	✗	✗	✗	✗	✗	✗	✗
LISTEN Chen et al. (2025c)	Oct 2025	✓	✓	✗	✓	✗	✗	✗	✗	✗
AudioMarathon He et al. (2025b)	Oct 2025	✓	✓	✗	✗	✗	✗	✗	✗	✗
ISA-Bench Li et al. (2025a)	Oct 2025	✓	✗	✗	✗	✗	✗	✓	✗	✗
Hearing the Order Lin et al. (2025d)	Oct 2025	✗	✗	✗	✗	✗	✗	✗	✓	✗
Safety under Emotional Variations Feng et al. (2025)	Oct 2025	✗	✗	✗	✗	✗	✗	✓	✓	✗
Gender Bias in SpeechLLMs Satish et al. (2025)	Oct 2025	✗	✗	✗	✗	✗	✗	✗	✗	✓
BRACE Guo et al. (2025b)	Dec 2025	✓	✗	✗	✓	✗	✗	✗	✗	✗
Spoken DialogSum Lu et al. (2025a)	Dec 2025	✗	✗	✓	✗	✗	✗	✗	✗	✗
MAC-SLU Peng et al. (2025b)	Dec 2025	✓	✗	✗	✗	✗	✗	✗	✗	✗
SH-Bench Zhan et al. (2025)	Dec 2025	✓	✗	✗	✗	✓	✗	✗	✗	✗

[†] **General**: Perception (PE), Reasoning (RE), Interaction (IN).
[‡] **Trustworthy**: Hallucination (H), Privacy (P), Authentication (A), Safety (S), Robustness (R), Fairness (F).

Table 3: Overview of LALM evaluation benchmarks across general capabilities and trustworthy dimensions (continued).

Benchmark	Release	General†			Trustworthy‡					
		PE	RE	IN	H	P	A	S	R	F
Year 2026										
WESR Yang et al. (2026b)	Jan 2026	✓	✗	✗	✗	✗	✗	✗	✗	✗
RSA-Bench Zhang et al. (2026e)	Jan 2026	✓	✓	✗	✗	✗	✗	✗	✓	✗
PALM-Bench Wang et al. (2026g)	Jan 2026	✗	✓	✓	✓	✗	✗	✗	✗	✗
ChronosAudio Luo et al. (2026)	Jan 2026	✓	✓	✗	✓	✗	✗	✗	✗	✗
HearSay Wang et al. (2026b)	Jan 2026	✗	✗	✗	✗	✓	✗	✗	✗	✗
AGL1K Zhang et al. (2026d)	Jan 2026	✗	✓	✗	✗	✗	✗	✗	✗	✓
VoxPrivacy Wang et al. (2026h)	Jan 2026	✗	✗	✓	✗	✓	✓	✗	✗	✗
BiasInEar Wei et al. (2026)	Feb 2026	✗	✗	✗	✗	✗	✗	✗	✗	✓
DailyTalkEdit Xu et al. (2026)	Feb 2026	✗	✗	✗	✗	✗	✓	✗	✗	✗
HumDial-EIBench Wang et al. (2026d)	Apr 2026	✗	✓	✓	✗	✗	✗	✗	✓	✗
VoxSafeBench Wang et al. (2026i)	Apr 2026	✓	✗	✓	✗	✓	✗	✓	✗	✓
HalluAudio Zhao et al. (2026a)	Apr 2026	✓	✓	✗	✓	✗	✗	✗	✓	✗
MUSA Koo (2026)	May 2026	✓	✓	✗	✗	✗	✗	✗	✓	✗
PitchBench Dujardin et al. (2026)	May 2026	✓	✗	✗	✗	✗	✗	✗	✓	✗
EnvMem Xiao et al. (2026)	May 2026	✓	✗	✓	✗	✗	✗	✗	✓	✗
VoiceGiraffe Ye et al. (2026)	May 2026	✓	✓	✗	✗	✗	✗	✗	✓	✗
VoxParadox Pang et al. (2026a)	May 2026	✓	✗	✗	✓	✗	✗	✗	✓	✗
MusTBENCH Kwon et al. (2026)	May 2026	✓	✓	✗	✓	✗	✗	✗	✗	✗
SpeechJBB Ceccatelli et al. (2026)	Jun 2026	✓	✗	✗	✗	✗	✗	✓	✓	✗
GlobeAudio Tan & Zhang (2026)	Jun 2026	✓	✓	✗	✗	✗	✗	✗	✓	✓
AudioProcessBench Zhao et al. (2026b)	Jun 2026	✓	✓	✗	✓	✗	✗	✗	✗	✗
RAIL Jin et al. (2026)	Jun 2026	✓	✓	✗	✗	✗	✗	✗	✗	✗
IHBench Salimi et al. (2026)	Jun 2026	✗	✓	✓	✗	✗	✗	✗	✓	✗
AOR-Bench Yang et al. (2026d)	Jun 2026	✓	✓	✗	✗	✗	✗	✓	✓	✗
MSU-Bench Sun et al. (2026)	Jun 2026	✓	✓	✓	✓	✗	✗	✗	✗	✗
ParaPairAudioBench Jeon et al. (2026)	Jun 2026	✓	✗	✗	✗	✗	✗	✗	✓	✗
CASU Zhang et al. (2026c)	Jun 2026	✓	✓	✗	✓	✗	✗	✗	✗	✗
SpeechEQ Wu et al. (2026d)	Jun 2026	✓	✓	✓	✗	✗	✗	✗	✓	✗

† **General**: Perception (PE), Reasoning (RE), Interaction (IN).
‡ **Trustworthy**: Hallucination (H), Privacy (P), Authentication (A), Safety (S), Robustness (R), Fairness (F).

musical content while failing to localize the evidence that supports their answers. **RSA-Bench** Zhang et al. (2026e) exposes a *Perception-Cognition Gap*, where low-level recognition remains robust but higher-order reasoning collapses under real-world degradation; its *Denoising Paradox* suggests standard enhancement may worsen downstream reasoning. **AudioBench** Wang et al. (2025a) identifies a *Modality Fusion Paradox*, showing multimodal fusion does not improve disentanglement, especially in speech-dominant settings. These benchmarks motivate a shift from coarse classification toward fine-grained perceptual modeling.

5.1.2 From Textual Bias to Genuine Listening

A central grounding failure is *Textual Dominance*, where models rely more on linguistic priors than acoustic evidence. **MCR-BENCH** Wang et al. (2025b) evaluates this problem through *Modal Conflict Resolution*, showing that under adversarial text-audio conflicts, accuracy drops sharply while confidence remains high;

metrics such as *Text Influence Rate* quantify the degree of textual bias. **LISTEN** Chen et al. (2025c) further decouples lexical semantics from paralinguistic cues, revealing that many models behave as "transcribers" and approach random performance when lexical cues are unavailable. **VoxParadox** Pang et al. (2026a) extends this diagnosis to 2,000 adversarial examples across 10 paralinguistic tasks, showing that models often follow transcript-implied labels over contradictory acoustic cues. **CASU** Zhang et al. (2026c) extends this diagnosis from speech–text conflicts to complete auditory scenes composed of speech, acoustic events, and background environments. Through contextual QA, entity extraction, speaker-role inference, and counterfactual reasoning, it shows that current models often process individual sound layers correctly but fail to integrate their logical relationships.

This limitation also appears in speech-to-speech interaction. **S2S-Arena** Jiang et al. (2025) evaluates paralinguistic instruction following across 4 domains and 21 tasks, finding cascaded ASR–LLM–TTS systems outperform end-to-end models, and generating paralinguistic output is harder than understanding such cues. In open-ended generation, **BRACE** Guo et al. (2025b) assesses reference-free audio caption alignment; its BRACE-Hallucination split shows even the best LALM reaches only 63.19 F1, indicating hallucinated, acoustically ungrounded content remains common beyond classification tasks.

5.1.3 From Fact Retrieval to Personal Alignment

Fidelity also concerns whether models reason correctly over spoken contexts. **ContextASR-Bench** Wang et al. (2025d) evaluates how hierarchical context helps error correction, while **C**^{3} Ma et al. (2025a) probes phonological, semantic, and discourse-level ambiguities in bilingual speech. These studies show that current speech dialogue models even still struggle when lexical cues are insufficient and acoustic-prosodic information is strongly needed for disambiguation.

Higher-level reasoning benchmarks further reveal a gap between transcription and cognition. **RAIL** Jin et al. (2026) organizes auditory intelligence into CHC-grounded perception, reasoning, memory, efficiency, and knowledge dimensions, revealing highly uneven cognitive profiles across LALMs. **SpeechR** Yang et al. (2025d) and **MMAR** Ma et al. (2025c) assess factual, procedural, and abductive reasoning, showing that strong ASR does not guarantee reliable multi-step inference. **SAKURA** Yang et al. (2025a) extends this to multi-hop reasoning over speaker gender, language, emotion, and animal sounds; performance drops substantially from single-hop perception to multi-hop integration, and remains much stronger in text-only settings than in speech/audio settings. **GlobeAudio** Tan & Zhang (2026) evaluates 5,637 natural-audio questions across six languages, revealing pronounced weaknesses in low-resource languages and culturally grounded reasoning. This indicates that LALM reasoning is still largely text-driven rather than genuinely multimodal. **AudioProcessBench** Zhao et al. (2026b) further identifies step-level errors in audio-grounded reasoning, showing that correct final answers can conceal hallucinated events, faulty grounding, or invalid inference.

Finally, trustworthy grounding requires pragmatic and user-level alignment. **MMSU** Wang et al. (2025c) covers 47 linguistic and paralinguistic tasks, revealing strong semantic performance but weak phonological and paralinguistic understanding. **Spoken DialogSum** Lu et al. (2025a) and **MAC-SLU** Peng et al. (2025b) evaluate emotional fidelity and multi-intent instruction following, while **HumDial-EIBench** Wang et al. (2026d) shows that models can track and reason about emotions yet struggle to generate empathetic responses. **SpeechEQ** Wu et al. (2026d) extends emotional evaluation to multi-turn social reasoning, exposing text reliance and contextual amnesia when models track evolving affective trajectories. **PALM-Bench** Wang et al. (2026g) identifies a "personalization gap," where LALMs fail to align responses with user-specific persona profiles in multi-speaker scenarios. **MSU-Bench** Sun et al. (2026) evaluates 16 multi-speaker tasks, revealing persistent errors in binding speaker identities, utterances, and conversational evidence.

5.2 Stability and Robustness

While perceptual fidelity concerns accurate grounding in static acoustic inputs, trustworthy LALMs must also exhibit **Behavioral Stability**: the ability to maintain reliable performance under dynamic real-world variations. In this context, stability refers to invariance against non-semantic changes, including extended audio duration, prompt phrasing, output format, option ordering, speaker variability, and conversational

dynamics. Existing evaluations mainly diagnose two failure modes: **Long-Context Collapse** in extended audio understanding (Sec. 5.2.1) and **Instructional Sensitivity** in interactive use (Sec. 5.2.2).

5.2.1 Temporal Robustness in Long-Form Context

LALMs often perform well on short clips but degrade sharply as audio duration increases. We define **Temporal Robustness** as the capacity to preserve attention, memory, and reasoning quality under long-form acoustic contexts. Recent benchmarks reveal a widespread **Long-Context Collapse**, where document-level audio understanding suffers severe information loss and reasoning degradation.

ChronosAudio Luo et al. (2026) systematically quantifies this failure across 36,000 test instances and multiple task types, using metrics for verbatim transcription, temporal localization, and high-level comprehension. By stratifying inputs into short, medium, and long durations, it reveals a non-linear degradation pattern, with some tasks dropping by over 90% in long-context settings. The benchmark attributes this to **Structural Attention Dilution**, where attention mechanisms fail to preserve temporal locality, motivating evaluation beyond aggregate accuracy toward degradation-rate analysis. **EnvMem** Xiao et al. (2026) isolates multi-turn acoustic-memory failures and identifies representational trajectory drift, rather than diffuse attention, as the main source of non-speech information loss.

AudioMarathon He et al. (2025b) further examines the trade-off between long-context understanding and computational efficiency. Its dual evaluation framework combines task accuracy with latency and memory measurements, exposing the quadratic cost of attention-based processing. Although acceleration methods such as token pruning or sparse attention can partially recover retrieval performance, they often fail to restore high-fidelity reasoning, revealing a **Restorative Ceiling**. **VoiceGiraffe** Ye et al. (2026) extends evaluation to real hour-long audio, revealing particular difficulty in retaining and integrating sparse evidence over long temporal spans. These findings suggest that temporal robustness cannot be achieved by enlarging context windows; instead, models must optimize the accuracy–efficiency Pareto frontier while preserving long-range reasoning.

5.2.2 Interaction and Instruction Robustness

Beyond temporal stability, trustworthy LALMs should display **Cognitive Conviction**: semantically equivalent inputs should yield consistent outputs despite changes in phrasing, format, ordering, speaker, or dialogue context. However, existing evaluations reveal severe **Interactional Fragility**, indicating that current systems remain sensitive to superficial interaction artifacts.

ISA-Bench Li et al. (2025a) evaluates instruction sensitivity by varying prompt description, output format, and task composition. It shows that even strong models degrade when instructions deviate from standard templates, with structured-output compliance such as JSON often falling below 50%. Its **Instruction-Following Rate** and **Relative Performance Score** quantify robustness relative to each model's best-case behavior. Importantly, ISA-Bench reveals a plasticity–stability dilemma: instruction tuning improves compliance but can cause catastrophic forgetting of acoustic capabilities. **URO-Bench** Yan et al. (2025b) corroborates this finding for end-to-end speech dialogue models, showing that conversational instruction-following gains often come at the cost of paralinguistic and audio comprehension.

Interactional fragility also emerges from acoustic variability and multimodal conversation. **VoiceBench** Chen et al. (2026b) evaluates voice assistants under speaker variation, noise, and content shifts, showing that acoustic changes significantly perturb instruction following. **VoiceAssistant-Eval** Wang et al. (2025f) broadens the evaluation to listening, speaking, and viewing tasks, revealing that models often speak fluently but lag in audio understanding and multimodal integration. Similarly, **VocalBench** Liu et al. (2025b) decomposes robustness into semantic quality, acoustic performance, conversational ability, and robustness, showing that models may preserve semantic coherence while failing in acoustic naturalness or multi-turn consistency. **SOVA-Bench** Hou et al. (2025b) reaches a similar conclusion: current speech LLMs can generate semantically plausible responses, but their acoustic quality and speech-level conversational robustness remain limited.

More fine-grained interaction studies expose failures in natural dialogue flow. **Talking Turns** Arora et al. (2025) benchmarks turn-taking prediction, including yielding, backchanneling, interruption, and floor-holding. Compared with human-human interaction, current systems often miss turn yields, interrupt too aggressively, rarely produce backchannels, and perform near randomly on backchannel understanding. These results show that robust spoken interaction requires modeling real-time conversational dynamics, not only following explicit instructions. **IHBench** Salimi et al. (2026) evaluates post-interruption recovery in structured workflows, showing that detecting an interruption does not guarantee correct task resumption or non-redundant responses.

Finally, robustness must also be tested against evaluation artifacts. **Hearing the Order** Lin et al. (2025d) identifies selection bias in multiple-choice audio evaluations: simply permuting answer options can change accuracy by up to 24% and alter model rankings. It advocates permutation-based protocols to separate true understanding from positional heuristics. **ParaPairAudioBench** Jeon et al. (2026) evaluates LALM judges across five paralinguistic dimensions, revealing large human–model gaps, position bias, and poor calibration on tied pairs. **WildSpeech-Bench** Zhang et al. (2025b) evaluates end-to-end speech LLMs on realistic speech phenomena such as prosody, homophones, stuttering, and speaker diversity, showing large performance variation across non-ideal conditions. **MUSA** Koo (2026) introduces multilingual distractors under controlled SNRs, showing that source separation does not fully prevent confident reasoning from the wrong speech stream. To probe vulnerability under deliberate conversational distortion, **Gaslighting Attacks** Wu et al. (2026c) evaluates cognitive robustness against five psychological negation strategies, revealing significant performance degradation accompanied by behavioral anomalies like unsolicited apologies. Together, these benchmarks indicate that robust evaluation must actively perturb interaction forms, acoustic conditions, and dialogue structures to determine whether LALMs genuinely reason or merely exploit superficial patterns.

5.3 Safety and Alignment

Safety and alignment evaluate whether LALMs adhere to human values and resist malicious exploitation. Compared with text-only systems, the auditory modality introduces additional risks: acoustic signals can serve as adversarial carriers, backdoor triggers, biometric identifiers, or sources of demographic bias. Existing work mainly examines two aspects: defensive security against jailbreaks and acoustic backdoors (Sec. 5.3.1), and broader risks in privacy, fairness, and authentication (Sec. 5.3.2).

5.3.1 Jailbreaks and Acoustic Backdoors

Audio input substantially expands the attack surface of LALMs, enabling both text-transferred jailbreaks and audio-native attacks. **JALMBench** Peng et al. (2025c) provides a large-scale comparison between textual and auditory jailbreaks, showing that audio attacks achieve higher success rates than text attacks and that prompt-level defenses can reduce but not eliminate vulnerabilities, often at the cost of utility. **Jailbreak-AudioBench** Cheng et al. (2025) further shows that simple audio edits, such as changes in speed, pitch, emotion, or accent, can significantly increase attack success, indicating that non-semantic acoustic features can bypass safety guardrails. **SpeechJBB** Ceccatelli et al. (2026) evaluates harmful code-switched speech and pseudo-word obfuscation, showing that multilingual and phonological variations substantially weaken refusal alignment. Moving beyond editing-based attacks, **AudioJailbreak** Chen et al. (2025a) targets end-to-end LALMs with stealthy, universal, and over-the-air robust adversarial audio, achieving high success rates even under weak-adversary assumptions. These studies show that jailbreaks are not spoken versions of textual attacks but exploit modality-specific vulnerabilities.

Signal-level perturbations create another safety risk. **AJailBench** Song et al. (2025) uses optimized audio perturbations such as time stretching and fading to lower refusal rates while preserving semantic transcription similarity. This suggests that current safety mechanisms often rely too heavily on clean ASR-like representations and fail to detect adversarial information embedded in non-standard acoustic patterns.

More insidious threats arise from hidden or socially natural triggers. **AudioSafe** Lin et al. (2025b) studies acoustic backdoors where background noise, prosody, emotion, or speaking rate can trigger unsafe behavior. It shows that highly effective backdoors can be implanted with only small amounts of poisoned data, implying

that seemingly harmless acoustic conditions may become latent switches for malicious outputs. **Safety under Emotional Variations** Feng et al. (2025) further reveals *Emotional Hijacking*: malicious requests expressed with certain emotional intensities are more likely to bypass refusal mechanisms. This indicates that affective cues may interfere with safety alignment, motivating affect-aware safety evaluation and defense. Conversely, **AOR-Bench** Yang et al. (2026d) exposes widespread over-refusal on pseudo-harmful queries, showing that models often ignore benign intent supplied by the surrounding acoustic context.

5.3.2 Privacy, Fairness, and Authentication

Because speech carries rich biometric and contextual information, LALMs naturally face a tension between personalization and privacy Wang et al. (2026i). A core risk is *Unintended Attribute Inference*, where models infer sensitive traits from voice without consent. **HearSay** Wang et al. (2026b) evaluates inference of attributes such as gender, socioeconomic status, and health conditions from short speech clips, showing that models can recover sensitive information with high accuracy and that chain-of-thought reasoning may further amplify privacy leakage. **AGL1K** Zhang et al. (2026d) extends this concern to audio geo-localization, demonstrating that environmental sounds and linguistic cues can be combined to infer user location, raising surveillance risks. In multi-speaker settings, **SH-Bench** Zhan et al. (2025) evaluates selective hearing and shows that strong audio comprehension does not necessarily translate into effective protection of bystander privacy. **VoxPrivacy** Wang et al. (2026h) extends this to shared devices by formalizing *interactional privacy* across multi-turn dialogs. By evaluating a model's ability to bind speaker identities with contextual secrets, it reveals that most open-source LALMs achieve near-chance accuracy due to cross-attention failures.

Privacy leakage can also amplify fairness risks, as models may implicitly condition responses on inferred demographic attributes. **Gender Bias in SpeechLLMs** Satish et al. (2025) shows that SpeechLLMs inherit identity cues directly from acoustic signals, making bias more implicit and harder to control than in text-only settings. **BiasInEar** Wei et al. (2026) broadens the analysis to multilingual speech and finds that models are particularly sensitive to language and option ordering, suggesting that speech interfaces can amplify structural biases in evaluation. In high-stakes domains, **VocalAgent** Kim et al. (2025c) evaluates vocal health diagnostics and warns that demographic and class imbalance may lead to unfair or unsafe medical recommendations, especially for general-purpose commercial audio models.

Finally, LALMs must authenticate speaker identity rather than become tools for impersonation. **AudioTrust** Li et al. (2025c) evaluates identity verification bypass and voice-cloning spoofing, showing that open-source models remain highly vulnerable and that advanced cloning systems can successfully impersonate speakers with substantial success rates. **DailyTalkEdit** Xu et al. (2026) complements this by reformulating anti-spoofing as a holistic generation task, jointly reasoning over spoofing methods, manipulated attributes, and semantic impacts. Together, these benchmarks indicate that safe deployment of LALMs requires not only refusal policies but also robust defenses against acoustic attacks, privacy leakage, demographic bias, and synthetic-voice impersonation.

5.4 Future Horizons of LALMs' Evaluation

Current evaluation methodologies primarily offer phenomenological snapshots that measure performance errors without elucidating their underlying failure mechanisms. To advance the systematic evaluation of LALM trustworthiness, the field must transition from static behavioral testing to rigorous structural verification across four paradigmatic shifts.

1.Causal Auditory World Modeling. Evaluation must move beyond statistical correlation to assess if models comprehend the physical dynamics governing auditory scenes. Future benchmarks should prioritize counterfactual reasoning to ensure reasoning is grounded in a coherent internal physics engine rather than superficial pattern recognition.

2.Agent-Based Dynamic Red-Teaming. To mitigate the contamination and overfitting risks of static datasets, stability assessments should evolve into real-time ecosystems where adaptive adversarial agents probe decision boundaries via noise injection or language switching. Metrics must shift from static accuracy to attack-defense curves that quantify the interaction turns to break consistency.

Figure 7: The Outlook of LALM. Future research trajectories are organized along three critical dimensions: intrinsic mechanisms, multimodal safety, and rigorous evaluation. This evolution marks a transition from empirical performance scaling toward a structural and cognitive transformation.

3.Intrinsic Representation Engineering. Rather than relying on post-hoc behavioral suppression, safety evaluations should verify information disentanglement at the neural level. Evaluation should examine whether representation-learning objectives, including mutual-information minimization, reduce the encoding and recoverability of sensitive biometric attributes, while explicitly measuring residual leakage under realistic adversarial probes.

4.Mechanistic Interpretability. To move beyond "opaque behaviorism," audio mechanistic interpretability tools must map specific neural circuits to auditory functions. Future benchmarks should incorporate internal consistency checks that monitor states for uncertainty or conflict before generation, transforming evaluation from probabilistic guessing into predictive failure detection.

6 Outlook and Conclusion

6.1 Future Outlook

The evolution of Large Audio-Language Models is transitioning from empirical performance scaling toward a structural and cognitive transformation. We identify several critical research trajectories organized along the dimensions of intrinsic mechanisms, multimodal safety, and rigorous evaluation. A comprehensive roadmap of these emerging directions is synthesized in Fig. 7.

6.2 Conclusion

This survey has presented a systematic investigation into the landscape of LALMs, analyzing their architectural evolution from task-specific cascaded systems to unified multimodal generative frameworks. Our exploration of endogenous mechanisms reveals that while sophisticated cross-modal alignment and reinforcement learning strategies have unlocked emergent reasoning abilities, they have simultaneously introduced a complex, high-dimensional attack surface. Our critical analysis through the six analytical pillars—hallucination, robustness, safety, privacy, fairness, and authentication—highlights a significant developmental asymmetry. While offensive research has matured into a diverse ecosystem of adversarial manipulations and stealthy jailbreaks, defensive mechanisms remain rudimentary and largely reactive. Bridging this chasm requires the community to prioritize multimodal safety alignment as a core architectural property rather than a post-hoc constraint. Ultimately, progress toward more trustworthy audio intelligence is likely to require both stronger grounding in physical reality and rigorous defense-in-depth protocols for complex real-world deployment.

7 Broader Impact Statement

LALMs have the potential to improve accessibility, human–computer interaction, education, healthcare, and multilingual communication. However, their deployment also introduces risks that extend beyond those of text-only systems because audio signals may encode sensitive information such as speaker identity, demographic attributes, emotional state, health conditions, and environmental context. As surveyed in this work, failures involving hallucination, adversarial manipulation, privacy leakage, demographic bias, unsafe responses, and synthetic-voice impersonation may cause disproportionate harm in high-stakes or socially sensitive applications.

This survey aims to support the responsible development and evaluation of LALMs by organizing existing evidence on these risks and highlighting corresponding defensive and evaluation strategies. And systematizing attack methods may have dual-use implications by making the threat landscape more accessible to malicious actors. We present attacks primarily at the level needed for scientific comparison and risk assessment, without providing implementation-level instructions intended to facilitate misuse. We encourage researchers and practitioners to evaluate LALMs only with appropriately licensed or consented audio data, protect biometric and contextual information, report performance across relevant demographic and linguistic groups, and follow responsible disclosure practices when identifying vulnerabilities. Ultimately, the taxonomies and research directions presented in this survey should be viewed as tools for improving accountability and defense, rather than as evidence that current LALMs are ready for unrestricted deployment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Tony Alex, Wish Suharitdamrong, Sara Atito, Armin Mustafa, Philip JB Jackson, Imran Razzak, and Muhammad Awais. Pal: Probing audio encoders via llms-audio information transfer into llms. *arXiv preprint arXiv:2506.10423*, 2025.

Ranya Aloufi, Srishti Gupta, Soumya Shaw, Battista Biggio, and Lea Schönherr. Evaluation of audio language models for fairness, safety, and security. *arXiv preprint arXiv:2603.13262*, 2026.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*, 2024.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. *arXiv preprint arXiv:2503.01174*, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Saurabhchand Bhati, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Towards audio token compression in large audio language models. *arXiv preprint arXiv:2511.20973*, 2025.

Suhas BN, Andrew M Sherrill, Jyoti Alaparthi, Dominik Mattioli, Rosa I Arriaga, Chris W Wiese, and Saeed Abdullah. Fine-tuning large audio-language models with lora for precise temporal localization of prolonged exposure therapy elements. *arXiv preprint arXiv:2506.09707*, 2025.

Brandon James Carone, Iran R Roman, and Pablo Ripollés. The muse benchmark: Probing music perception and auditory relational reasoning in audio llms. *arXiv preprint arXiv:2510.19055*, 2025.

Virginia Ceccatelli, Yejin Jeon, and David Ifeoluwa Adelani. Speechjbb: Probing safety alignment and comprehension in large audio language models under code-switched speech, 2026. URL `https://arxiv.org/abs/2606.06037`.

Yuatyong Chaichana, Pittawat Taveekitworachai, Warit Sirichotedumrong, Potsawee Manakul, and Kunat Pipatanakul. Extending audio context for long-form understanding in large audio-language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6046–6066, 2026.

Kai-Wei Chang, Wei-Chih Chen, En-Pei Hu, Hung-yi Lee, and James Glass. Tico: Time-controllable training for spoken dialogue models. *arXiv preprint arXiv:2603.22267*, 2026.

Guangke Chen, Fu Song, Zhe Zhao, Xiaojun Jia, Yang Liu, Yanchen Qiao, and Weizhe Zhang. Audiojailbreak: Jailbreak attacks against end-to-end large audio-language models. *arXiv preprint arXiv:2505.14103*, 2025a.

Guangke Chen, Yuhui Wang, Shouling Ji, Xiapu Luo, and Ting Wang. Synthetic voices, real threats: Evaluating large text-to-speech models in generating harmful audio. *arXiv preprint arXiv:2511.10913*, 2025b.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio llms really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. *arXiv preprint arXiv:2510.10444*, 2025c.

Junjie Chen, Yao Hu, Junjie Li, Kangyue Li, Kun Liu, Wenpeng Li, Xu Li, Ziyuan Li, Feiyu Shen, Xu Tang, et al. Fireredchat: A pluggable, full-duplex voice interaction system with cascaded and semi-cascaded implementations. *arXiv preprint arXiv:2509.06502*, 2025d.

Meng Chen, Kun Wang, Li Lu, Jiaheng Zhang, and Tianwei Zhang. Hijacking large audio-language models via context-agnostic and imperceptible auditory prompt injection. *arXiv preprint arXiv:2604.14604*, 2026a.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14: 378–398, 2026b.

Yuxuan Chen and Haoyuan Yu. From turn-taking to synchronous dialogue: A survey of full-duplex spoken language models. *arXiv preprint arXiv:2509.14515*, 2025.

Hao Cheng, Erjia Xiao, Jing Shao, Yichi Wang, Le Yang, Chao Shen, Philip Torr, Jindong Gu, and Renjing Xu. Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models. *arXiv preprint arXiv:2501.13772*, 2025.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. Aha-bench: Benchmarking audio hallucinations in large audio-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Chung-Ming Chien, Manu Orsini, Eugene Kharitonov, Neil Zeghidour, Karen Livescu, and Alexandre Défossez. Moshirag: Asynchronous knowledge retrieval for full-duplex speech language models. *arXiv preprint arXiv:2604.12928*, 2026.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

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

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13943–13970, 2025.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*, 2024.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108, 2023.

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. Soundmind: Rl-incentivized logic reasoning for audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 528–540, 2025.

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

Zhihao Du, Jiaming Wang, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, et al. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673*, 2023.

Milan Liessens Dujardin, Song-Ze Yu, Craver Corbyn Thomas-Smith, David M. Chan, and Karina Nguyen. Pitchbench: Measuring pitch hearing in audio-language models, 2026. URL `https://arxiv.org/abs/2605.26176`.

Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. Incentivizing consistent, effective and scalable reasoning capability in audio llms via reasoning process rewards. *arXiv preprint arXiv:2510.20867*, 2025.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

Qingkai Fang, Shoutao Guo, and Yang Feng. Bayling-duplex: Native full-duplex speech dialogue with a single autoregressive llm. *arXiv preprint arXiv:2606.14528*, 2026.

Bo-Han Feng, Chien-Feng Liu, Yu-Hsuan Li Liang, Chih-Kai Yang, Szu-Wei Fu, Zhehuai Chen, Ke-Han Lu, Sung-Feng Huang, Chao-Han Huck Yang, Yu-Chiang Frank Wang, et al. Investigating safety vulnerabilities of large audio-language models under speaker emotional variations. *arXiv preprint arXiv:2510.16893*, 2025.

Tiantian Feng, Rajat Hebbar, Nicholas Mehlman, Xuan Shi, Aditya Kommineni, et al. A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness. *arXiv preprint arXiv:2212.09006*, 2022.

Alexandrine Fortier, Thomas Thebaud, Jesús Villalba, Najim Dehak, and Patrick Cardinal. Backdoor attacks against speech language models. *arXiv preprint arXiv:2510.01157*, 2025.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

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

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

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

Marcel Gibier, Raphaël Duroselle, Pierre Serrano, Olivier Boeffard, and Jean-François Bonastre. Segmentwise pruning in audio-language models. *arXiv preprint arXiv:2511.14293*, 2025.

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

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Petr Grinberg and Hassan Shahmohammadi. Alarm: Audio-language alignment for reasoning models. *arXiv preprint arXiv:2603.09556*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Tianyu Guo, Hongyu Chen, Hao Liang, Meiyi Qiang, Bohan Zeng, Linzhuang Sun, Bin Cui, and Wentao Zhang. Brace: A benchmark for robust audio caption quality evaluation. *arXiv preprint arXiv:2512.10403*, 2025b.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, et al. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. *arXiv preprint arXiv:2509.21060*, 2025a.

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, et al. Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. *arXiv preprint arXiv:2510.07293*, 2025b.

Yingxu He, Zhuohan Liu, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F Chen, and Ai Ti Aw. Meralion-audiollm: Technical report. *arXiv e-prints*, pp. arXiv–2412, 2024.

William Held, Yanzhe Zhang, Minzhi Li, Weiyan Shi, Michael J Ryan, and Diyi Yang. Distilling an end-to-end voice assistant without instruction training data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7876–7891, 2025.

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

Yixuan Hou, Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. Sova-bench: Benchmarking the speech conversation ability for llm-based voice assistant. *arXiv preprint arXiv:2506.02457*, 2025b.

Chi-Yuan Hsiao, Ke-Han Lu, Yu-Kuan Fu, Guan-Ting Lin, Hsiao-Tsung Hung, and Hung-yi Lee. Aspirin: Action space projection for interactivity-optimized reinforcement learning in full-duplex speech language models. *arXiv preprint arXiv:2604.10065*, 2026.

Jiliang Hu, Zuchao Li, Baoyuan Qi, Guoming Liu, and Ping Wang. End-to-end contrastive language-speech pretraining model for long-form spoken question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31041–31049, 2026a.

Jinbo Hu, Yin Cao, Ming Wu, Zhenbo Luo, and Jun Yang. Salm: Spatial audio language model with structured embeddings for understanding and editing. *arXiv preprint arXiv:2507.16724*, 2025a.

Jing Hu, Danxiang Zhu, Xianlong Luo, Dan Zhang, Shuwei He, Yishu Lei, Haitao Zheng, Shikun Feng, Jingzhou He, Yu Sun, et al. Cord: Bridging the audio-text reasoning gap via weighted on-policy cross-modal distillation. *arXiv preprint arXiv:2601.16547*, 2026b.

Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. Efficient and direct duplex modeling for speech-to-speech language model. *arXiv preprint arXiv:2505.15670*, 2025b.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4552–4572, 2024.

Dawei Huang, Yongjie Lv, Ruijie Xiong, Chunxiang Jin, and Xiaojiang Peng. When tone and words disagree: Towards robust speech emotion recognition under acoustic-semantic conflict. *arXiv preprint arXiv:2601.04564*, 2026.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

Lok-Lam Ieong, Chia-Chien Chen, Chih-Kai Yang, Yu-Han Huang, An-Yu Cheng, and Hung-yi Lee. Nudging hidden states: Training-free model steering for chain-of-thought reasoning in large audio-language models. *arXiv preprint arXiv:2603.14636*, 2026.

Lovenya Jain, Pooneh Mousavi, Mirco Ravanelli, and Cem Subakan. Investigating faithfulness in large audio language models. *arXiv preprint arXiv:2509.22363*, 2025.

Jisu Jeon, Seungyeon Jwa, Joosung Lee, Jinhyeon Kim, Woojin Chung, Hwiyeol Jo, Jeonghoon Kim, Jonghyun Choi, and Soyoon Kim. Parapairaudiobench: Paralinguistic pairwise audio benchmark for lalm-as-a-judge, 2026. URL `https://arxiv.org/abs/2606.24648`.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*, 2024.

Hong Jia, Weibin Li, Jingyao Wu, Xiaofeng Yu, Yan Gao, Jintao Cheng, Xiaoyu Tang, Feng Xia, and Ting Dang. Decoding ambiguous emotions with test-time scaling in audio-language models. *arXiv preprint arXiv:2602.03873*, 2026.

Feng Jiang, Zhiyu Lin, F. Bu, Yuhao Du, Benyou Wang, and Haizhou Li. S2s-arena, evaluating speech2speech protocols on instruction following with paralinguistic information. *arXiv preprint arXiv:2503.05085*, 2025.

Xilin Jiang, Hannes Gamper, and Sebastian Braun. Sci-phi: A large language model spatial audio descriptor. *IEEE Open Journal of Signal Processing*, 2026.

Hongyu Jin, Siyi Wang, Yang Xiao, Jiaheng Dong, Shihong Tan, Kaiyuan peng, Georgiana Juravle, Shanquan Chen, Gongping Huang, Hong Jia, Eun-Jung Holden, James Bailey, and Ting Dang. Rail: Rethinking auditory intelligence in large audio-language models with a chc-grounded benchmark, 2026. URL `https://arxiv.org/abs/2606.11260`.

Weifei Jin, Yuxin Cao, Junjie Su, Minhui Xue, Jie Hao, Ke Xu, Jin Song Dong, and Derui Wang. Almguard: Safety shortcuts and where to find them as guardrails for audio-language models. *arXiv preprint arXiv:2510.26096*, 2025.

Bodam Kim, Hiskias Dingeto, Taeyoun Kwon, Dasol Choi, DongGeon Lee, Haon Park, JaeHoon Lee, and Jongho Shin. When good sounds go adversarial: Jailbreaking audio-language models with benign inputs. *arXiv preprint arXiv:2508.03365*, 2025a.

Jaeyeon Kim, Heeseung Yun, Sang Hoon Woo, Chao-Han Huck Yang, and Gunhee Kim. Wow-bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations. *arXiv preprint arXiv:2508.20976*, 2025b.

Yubin Kim, Taehan Kim, Wonjune Kang, Eugene Park, Joonsik Yoon, Dongjae Lee, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Cynthia Breazeal, et al. Vocalagent: Large language models for vocal health diagnostics with safety-aware evaluation. *arXiv preprint arXiv:2505.13577*, 2025c.

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

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024.

Heejoon Koo. Can large audio language models ignore multilingual distractors? an evaluation of their selective auditory attention capabilities, 2026. URL `https://arxiv.org/abs/2605.17225`.

Gokul Karthik Kumar, Rishabh Saraf, Ludovick Lepauloux, Abdul Muneer, Billel Mokeddem, and Hakim Hacid. Competitive audio-language models with data-efficient single-stage training on public data. *arXiv preprint arXiv:2509.07526*, 2025a.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025b.

Nikita Kuzmin, Tao Zhong, Jiajun Deng, Yingke Zhu, Tristan Tsoi, Tianxiang Cao, Simon Lui, Kong Aik Lee, and Eng Siong Chng. Privacy-preserving end-to-end full-duplex speech dialogue models. *arXiv preprint arXiv:2603.08179*, 2026.

Daeyong Kwon, Qiyu Wu, Shinobu Kuriya, Junghyun Koo, Shuyang Cui, Zhi Zhong, Wei-Hsiang Liao, Hiromi Wakaki, and Yuki Mitsufuji. Mustbench: Benchmarking and advancing temporal grounding in music llms, 2026. URL `https://arxiv.org/abs/2605.29300`.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, et al. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*, 2023.

Yishu Lei, Shuwei He, Jing Hu, Dan Zhang, Xianlong Luo, Danxiang Zhu, Shikun Feng, Rui Liu, Jingzhou He, Yu Sun, et al. Moe adapter for large audio language models: Sparsity, disentanglement, and gradient-conflict-free. *arXiv preprint arXiv:2601.02967*, 2026.

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

Hongyi Li, Chengxuan Zhou, Chu Wang, Sicheng Liang, Yanting Chen, Qinlin Xie, Jiawei Ye, and Jie Wu. Stylebreak: Revealing alignment vulnerabilities in large audio-language models via style-aware audio jailbreak. *arXiv preprint arXiv:2511.10692*, 2025b.

Jiaqi Li, Chaoren Wang, Xiaohai Tian, Mingjie Chen, Xinyu Liang, Xu Li, Yufan Lin, Junwen Qiu, Jun Zhang, Lu Lu, et al. Flexislm: A dynamic and controllable frame rate spoken language model. *arXiv preprint arXiv:2606.31247*, 2026.

Kai Li, Can Shen, Yile Liu, Jirui Han, Kelong Zheng, Xuechao Zou, Zhe Wang, Shun Zhang, Xingjian Du, Hanjun Luo, et al. Audiotrust: Benchmarking the multifaceted trustworthiness of audio large language models. *arXiv preprint arXiv:2505.16211*, 2025c.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38, 2025d.

Pengcheng Li, Botao Zhao, Zuheng Kang, Junqing Peng, Xiaoyang Qu, Yayun He, and Jianzong Wang. Emo-rl: Emotion-rule-based reinforcement learning enhanced audio-language model for generalized speech emotion recognition. *arXiv preprint arXiv:2509.15654*, 2025e.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, et al. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*, 2025f.

Yupei Li, Li Wang, Yuxiang Wang, Lei Wang, Rizhao Cai, Jie Shi, Björn W Schuller, and Zhizheng Wu. Dfallm: Achieving generalizable multitask deepfake detection by optimizing audio llm components. *arXiv preprint arXiv:2512.08403*, 2025g.

Zhiying Li, Jason E Smerdon, Richard Seager, Noel Siegert, and Justin S Mankin. Emergent trends complicate the interpretation of the united states drought monitor (usdm). *AGU Advances*, 5(2):e2023AV001070, 2024.

Bin Lin, Bo Zhao, Boyong Wu, Chao Yan, Chen Wu, Cheng Yi, Chengyuan Yao, Daijiao Liu, Fei Tian, Feng Tian, et al. Stepaudio 2.5 technical report. *arXiv preprint arXiv:2605.23463*, 2026.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6626–6642, 2024a.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-yi Lee, and Ivan Bulyko. Paralinguistics-enhanced large language modeling of spoken dialogue. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10316–10320. IEEE, 2024b.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, G. Anumanchipalli, Alexander H. Liu, and Hung-yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*, 2025a.

Liang Lin, Miao Yu, Kaiwen Luo, Yibo Zhang, Lilan Peng, Dexian Wang, Xuehai Tang, Yuanhe Zhang, Xikang Yang, Zhenhong Zhou, et al. Hidden in the noise: Unveiling backdoors in audio llms alignment through latent acoustic pattern triggers. *arXiv preprint arXiv:2508.02175*, 2025b.

Weilin Lin, Jianze Li, Hui Xiong, and Li Liu. Sarsteer: Safeguarding large audio language models via safe-ablated refusal steering. *arXiv preprint arXiv:2510.17633*, 2025c.

Yu-Xiang Lin, Chen-An Li, Sheng-Lun Wei, Po-Chun Chen, Hsin-Hsi Chen, and Hung-yi Lee. Hearing the order: Investigating selection bias in large audio-language models. *arXiv preprint arXiv:2510.00628*, 2025d.

Zhaojiang Lin, Yong Xu, Kai Sun, Jing Zheng, Yin Huang, Surya Teja Appini, Krish Narang, Renjie Tao, Ishan Kapil Jain, Siddhant Arora, et al. Wearvox: An egocentric multichannel voice assistant benchmark for wearables. *arXiv preprint arXiv:2601.02391*, 2025e.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and Wenxuan Zhang. Seallms-audio: Large audio-language models for southeast asia. *arXiv preprint arXiv:2511.01670*, 2025a.

Heyang Liu, Yuhao Wang, Ziyang Cheng, Hongcheng Liu, Yiqi Li, Yixuan Hou, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. Vocalbench: Benchmarking the vocal conversational abilities for speech interaction models. *arXiv preprint arXiv:2505.15727*, 2025b.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 286–290. IEEE, 2024b.

Zhanxun Liu, Yifan Duan, Mengmeng Wang, Pengchao Feng, Haotian Zhang, Xiaoyu Xing, Y. Shan, Haina Zhu, Yuhang Dai, Chaochao Lu, Xipeng Qiu, Lei Xie, Lan Wang, Nan Yan, Zilong Zheng, Ziyang Ma, Kai Yu, and Xie Chen. X-talk: On the underestimated potential of modular speech-to-speech dialogue system. *arXiv preprint arXiv:2512.18706*, 2025c.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, et al. Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *arXiv preprint arXiv:2505.03739*, 2025.

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

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, et al. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *IEEE Transactions on Audio, Speech and Language Processing*, 2026a.

Ke-Han Lu, Szu-Wei Fu, Chao-Han Huck Yang, Zhehuai Chen, Sung-Feng Huang, Chih-Kai Yang, Yi-Cheng Lin, Chi-Yuan Hsiao, Wenze Ren, En-Pei Hu, Yu-Han Huang, An-Yu Cheng, Cheng-Han Chiang, Yu Tsao, Yu-Chiang Frank Wang, and Hung-yi Lee. How auditory knowledge in llm backbones shapes audio language models: A holistic evaluation. *arXiv preprint arXiv:2603.19195*, 2026b.

Yen-Ju Lu, Kunxiao Gao, Mingrui Liang, Helin Wang, Thomas Thebaud, Laureano Moro-Velazquez, Najim Dehak, and Jesus Villalba. Spoken dialogsum: An emotion-rich conversational dataset for spoken dialogue summarization. *arXiv preprint arXiv:2512.14687*, 2025a.

Yen-Ju Lu, Yashesh Gaur, Wei Zhou, Benjamin Muller, Jesus Villalba, Najim Dehak, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Srinivasan Iyer, et al. Latent speech-text transformer. *arXiv preprint arXiv:2510.06195*, 2025b.

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

Hieu-Thi Luong, Haoyang Li, Lin Zhang, Kong Aik Lee, and Eng Siong Chng. Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Chengqian Ma, Wei Tao, and Steven Y Guo. C3: A bilingual benchmark for spoken dialogue models exploring challenges in complex conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22789–22807, 2025a.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, et al. Safety at scale: A comprehensive survey of large model and agent safety. *Foundations and Trends in Privacy and Security*, 8(3-4):1–240, 2026a.

Ziyang Ma, Xiquan Li, Yakun Song, Wenxi Chen, Chenpeng Du, Jian Wu, Yuanzhe Chen, Zhuo Chen, Yuping Wang, Yuxuan Wang, et al. Towards reliable large audio language model. *arXiv preprint arXiv:2505.19294*, 2025b.

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

Ziyang Ma, Ruiyang Xu, Yinghao Ma, Chao-Han Huck Yang, Bohan Li, Jaeyeon Kim, Jin Xu, Jinyu Li, Carlos Busso, Kai Yu, et al. The interspeech 2026 audio reasoning challenge: Evaluating reasoning process quality for audio reasoning models and agents. *arXiv preprint arXiv:2602.14224*, 2026b.

Ziqiao Meng, Qichao Wang, Wenqian Cui, Yifei Zhang, Bingzhe Wu, Irwin King, Liang Chen, and Peilin Zhao. Parrot: Autoregressive spoken dialogue language modeling with decoder-only transformers. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michele Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. In *International Conference on Learning Representations*, volume 2024, pp. 51883–51898, 2024.

Binh Nguyen and Thai Le. Analyzing reasoning shifts in audio deepfake detection under adversarial attacks: The reasoning tax versus shield bifurcation. *arXiv preprint arXiv:2601.03615*, 2026.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, et al. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025.

NWU-LIST. Aabench: A comprehensive benchmark for evaluating audio adversarial robustness in llms. `https://huggingface.co/datasets/NWULIST/AABench`, 2026a. Accessed: 2026-05-08.

NWU-LIST. Speechjailbreaker: A framework for evaluating and breaking audio-language model safety. `https://github.com/NWULIST/SpeechJailbreaker`, 2026b. GitHub repository, Accessed: 2026-05-08.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio llms listen or read? analyzing and mitigating paralinguistic failures with voxparadox. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026a.

Zi Haur Pang, Xiaoxue Gao, Tatsuya Kawahara, and Nancy F Chen. Erm-minmaxgap: Benchmarking and mitigating gender bias in multilingual multimodal speech-llm emotion recognition. *arXiv preprint arXiv:2603.21050*, 2026b.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, et al. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*, 2025a.

Yuezhang Peng, Chonghao Cai, Ziang Liu, Shuai Fan, Sheng Jiang, Hua Xu, Yuxin Liu, Qiguang Chen, Kele Xu, Yao Li, et al. Mac-slu: Multi-intent automotive cabin spoken language understanding benchmark. *arXiv preprint arXiv:2512.01603*, 2025b.

Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Zeren Luo, Jingyi Zheng, Wenhan Dong, Xinlei He, Xuechao Wang, Yingjie Xue, et al. Jalmbench: Benchmarking jailbreak vulnerabilities in audio language models. *arXiv preprint arXiv:2505.17568*, 2025c.

Lam Pham, Phat Lam, Dat Tran, Hieu Tang, Tin Nguyen, Alexander Schindler, Florian Skopik, Alexander Polonsky, and Hai Canh Vu. A comprehensive survey with critical analysis for deepfake speech detection. *Computer Science Review*, 57:100757, 2025.

Amith Ramaswamy. *Enhancing listen, think, and understand (LTU) for temporal marine audio segmentation: from ship classification to multi-source maritime soundscape analysis: a thesis in Data Science*. PhD thesis, University of Massachusetts Dartmouth, 2025.

Yiming Ren, Xuenan Xu, Baoxiang Li, Shuai Wang, and Chao Zhang. Can audio large language models verify speaker identity?, 2025. URL `https://arxiv.org/abs/2509.19755`.

Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr. Multilingual and multi-accent jailbreaking of audio llms. *arXiv preprint arXiv:2504.01094*, 2025.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Vinu Sankar Sadasivan, Soheil Feizi, Rajiv Mathews, and Lun Wang. Attacker's noise can manipulate your audio-based llm in the real world. *arXiv preprint arXiv:2507.06256*, 2025.

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

S Sakshi, Vaibhavi Lokegaonkar, Neil Zhang, Ramani Duraiswami, Sreyan Ghosh, Dinesh Manocha, and Lie Lu. Spur: A plug-and-play framework for integrating spatial audio understanding and reasoning into large audio-language models. *arXiv preprint arXiv:2511.06606*, 2025.

Ahmad Salimi, Wentao Ma, Yuzhi Tang, Dongming Shen, Mu Li, and Alex Smola. Ihbench: Evaluating post-interruption recovery in voice agents with structured workflows, 2026. URL `https://arxiv.org/abs/2606.19595`.

Shree Harsha Bokkahalli Satish, Gustav Eje Henter, and Éva Székely. Do bias benchmarks generalise? evidence from voice-based evaluation of gender bias in speechllms. *arXiv preprint arXiv:2510.01254*, 2025.

Zhichao Sheng, Shilin Zhou, Chen Gong, and Zhenghua Li. Think smart, not hard: Difficulty adaptive reasoning for large audio language models. *arXiv preprint arXiv:2509.21960*, 2025.

Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, et al. Large language model safety: A holistic survey. *arXiv preprint arXiv:2412.17686*, 2024.

Jiacheng Shi, Hongfei Du, Y Alicia Hong, and Ye Gao. Emo-tta: Improving test-time adaptation of audio-language models for speech emotion recognition. *arXiv preprint arXiv:2509.25495*, 2025.

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

Yi-Jen Shih, Desh Raj, Chunyang Wu, Wei Zhou, SK Bong, Yashesh Gaur, Jay Mahadeokar, Ozlem Kalinli, and Mike Seltzer. Can speech llms think while listening? *arXiv preprint arXiv:2510.07497*, 2025.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, et al. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *arXiv preprint arXiv:2505.15406*, 2025.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*, 2025.

Zhaokai Sun, Shuai Wang, Zhennan Lin, Chengyou Wang, Dehui Gao, Yuang Cao, Chunjiang He, Pan Zhou, and Lei Xie. Msu-bench: Towards speaker-centric understanding in conversational multi-speaker scenarios, 2026. URL `https://arxiv.org/abs/2606.22868`.

Zhi Rui Tam and Yun-Nung Chen. Medvoicebias: A controlled study of audio llm behavior in clinical decision-making. *arXiv preprint arXiv:2511.06592*, 2025.

Ryner Tan and Wenxuan Zhang. Globeaudio: A multilingual multicultural benchmark for naturalistic evaluation of large audio-language models, 2026. URL `https://arxiv.org/abs/2606.08194`.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, et al. Fun-audio-chat technical report. *arXiv preprint arXiv:2512.20156*, 2025.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. When audio and text disagree: Revealing text bias in large audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4878–4888, 2025b.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025c.

He Wang, Linhan Ma, Dake Guo, Xiong Wang, Lei Xie, Jin Xu, and Junyang Lin. Contextasr-bench: A massive contextual speech recognition benchmark. *arXiv preprint arXiv:2507.05727*, 2025d.

Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. Listening between the frames: Bridging temporal gaps in large audio-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26233–26241, 2026a.

Jin Wang, Liang Lin, Kaiwen Luo, Weiliu Wang, Yitian Chen, Moayad Aloqaily, Xuehai Tang, Zhenhong Zhou, Kun Wang, Li Sun, et al. Hearsay benchmark: Do audio llms leak what they hear? *arXiv preprint arXiv:2601.03783*, 2026b.

Junyu Wang, Ziyang Ma, Zhengding Luo, Tianrui Wang, Meng Ge, Xiaobao Wang, and Longbiao Wang. Pay more attention to audio: Mitigating imbalance of cross-modal attention in large audio language models. *arXiv preprint arXiv:2509.18816*, 2025e.

Ke Wang, Houxing Ren, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Voiceassistant-eval: Benchmarking ai assistants across listening, speaking, and viewing. *arXiv preprint arXiv:2509.22651*, 2025f.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*, 2025g.

Mengqi Wang, Zhan Liu, Zengrui Jin, Guangzhi Sun, Chao Zhang, and Philip C Woodland. Audio-conditioned diffusion llms for asr and deliberation processing. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 18467–18471. IEEE, 2026c.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K Rubenstein, et al. Slm: Bridge the thin gap between speech and text foundation models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Qiaolin Wang, Xilin Jiang, Linyang He, Junkai Wu, and Nima Mesgarani. Sightsound-r1: Cross-modal reasoning distillation from vision to audio language models. *arXiv preprint arXiv:2509.15661*, 2025h.

Shuiyuan Wang, Zhixian Zhao, Hongfei Yue, Chengyou Wang, Shuai Wang, Hui Bu, Xin Xu, and Lei Xie. Humdial-eibench: A human-recorded multi-turn emotional intelligence benchmark for audio language models, 2026d. URL `https://arxiv.org/abs/2604.11594`.

Wenfu Wang, Chenxing Li, Liqiang Zhang, Yiyang Zhao, Yuxiang Zou, Hanzhao Li, Mingyu Cui, Hao Zhang, Kun Wei, Le Xu, et al. Covo-audio technical report. *arXiv preprint arXiv:2602.09823*, 2026e.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*, 2024.

Yancheng Wang, Osama Hanna, Ruiming Xie, Xianfeng Rui, Maohao Shen, Xuedong Zhang, Christian Fuegen, Jilong Wu, Debjyoti Paul, Arthur Guo, et al. Vowelprompt: Hearing speech emotions from text via vowel-level prosodic augmentation. *arXiv preprint arXiv:2602.06270*, 2026f.

Yuwen Wang, Xinyuan Qian, Tian-Hao Zhang, Jiaran Gao, Yuchen Pan, Xin Wang, Zhou Pan, Chen Wei, and Yiming Wang. Palm-bench: A comprehensive benchmark for personalized audio-language models. *arXiv preprint arXiv:2601.03531*, 2026g.

Yuxiang Wang, HongYu Liu, Dekun Chen, Xueyao Zhang, and Zhizheng Wu. Voxprivacy: A benchmark for evaluating interactional privacy of speech language models. In *The Fourteenth International Conference on Learning Representations*, 2026h. URL `https://openreview.net/forum?id=GNo1qMqgPD`.

Yuxiang Wang, Hongyu Liu, Yijiang Xu, Qinke Ni, Li Wang, Wan Lin, Kunyu Feng, Dekun Chen, Xu Tan, Lei Wang, et al. Voxsafebench: Not just what is said, but who, how, and where. *arXiv preprint arXiv:2604.14548*, 2026i.

Sheng-Lun Wei, Yu-Ling Liao, Yen-Hua Chang, Hen-Hsen Huang, and Hsin-Hsi Chen. Bias in the ear of the listener: Assessing sensitivity in audio language models across linguistic, demographic, and positional variations. *arXiv preprint arXiv:2602.01030*, 2026.

Gijs Wijngaard, Elia Formisano, Michel Dumontier, and Jenia Jitsev. Audiotoolagent: An agentic framework for audio-language models. *arXiv preprint arXiv:2510.02995*, 2025.

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

Daiqing Wu, Xuan Zhang, Dongbao Yang, Jiashu Yao, Longfei Chen, Qingsong Liu, Sicheng Zhao, Can Ma, Yangyang Kang, and Yu Zhou. Echo: Towards advanced audio comprehension via audio-interleaved reasoning. *arXiv preprint arXiv:2602.11909*, 2026a.

Donghang Wu, Tianyu Zhang, Yuxin Li, Hexin Liu, Chen Chen, Eng Siong Chng, and Yoshua Bengio. The silent thought: Modeling internal cognition in full-duplex spoken dialogue models via latent reasoning. *arXiv preprint arXiv:2603.17837*, 2026b.

Jinyang Wu, Bin Zhu, Xiandong Zou, Qiquan Zhang, Xu Fang, and Pan Zhou. Benchmarking gaslighting attacks against speech large language models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19867–19871. IEEE, 2026c.

Liang-Yuan Wu, Zih-Ching Chen, Tongshuang Wu, Chao-Han Huck Yang, and Hua Shen. Speecheq: Benchmarking emotional intelligence quotient in socially aware voice conversational models, 2026d. URL https://arxiv.org/abs/2606.25990.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025b.

Yang Xiao, Siyi Wang, Han Yin, Hong Jia, Vidhyasaharan Sethu, Eun-Jung Holden, and Ting Dang. Why can't they remember? uncovering representation and retrieval bottlenecks in multi-turn acoustic memory, 2026. URL https://arxiv.org/abs/2605.27039.

Yuankun Xie, Xiaoxuan Guo, Jiayi Zhou, Tao Wang, Jian Liu, Ruibo Fu, Xiaopeng Wang, Haonan Cheng, and Long Ye. Interpretable all-type audio deepfake detection with audio llms via frequency-time reinforcement learning. *arXiv preprint arXiv:2601.02983*, 2026.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024.

Zhen Xiong, Yujun Cai, Zhecheng Li, Junsong Yuan, and Yiwei Wang. Thinking with sound: Audio chain-of-thought enables multimodal reasoning in large audio-language models. *arXiv preprint arXiv:2509.21749*, 2025.

Xuenan Xu, Yiming Ren, Liwei Liu, Wen Wu, Baoxiang Li, Chaochao Lu, Shuai Wang, and Chao Zhang. Holiantispoof: Audio llm for holistic speech anti-spoofing. *arXiv preprint arXiv:2602.04535*, 2026.

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. E-chat: Emotion-sensitive spoken dialogue system with large language models. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 586–590. IEEE, 2024.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Li Xie, Yuxin Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, et al. Step-audio-editx technical report. *arXiv preprint arXiv:2511.03601*, 2025a.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*, 2025b.

Ruiqi Yan, Wenxi Chen, Zhanxun Liu, Ziyang Ma, Haopeng Lin, Hanlin Wen, Hanke Xie, Jun Wu, Yuzhe Liang, Yuxiang Zhao, et al. Soulx-duplug: Plug-and-play streaming state prediction module for realtime full-duplex speech conversation. *arXiv preprint arXiv:2603.14877*, 2026.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11798–11827, 2025c.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*, 2025d.

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

Chenchen Yang, Kexin Huang, Liwei Fan, Qian Tu, Botian Jiang, Dong Zhang, Linqi Yin, Shimin Li, Zhaoye Fei, Qinyuan Cheng, et al. Wesr: Scaling and evaluating word-level event-speech recognition. *arXiv preprint arXiv:2601.04508*, 2026b.

Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung-yi Lee. Sakura: On the multi-hop reasoning of large audio-language models based on speech and audio information. *arXiv preprint arXiv:2505.13237*, 2025a.

Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10155–10181, 2025b.

Dongchao Yang, Yuanyuan Wang, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. Uniaudio 2.0: A unified audio language model with text-aligned factorized audio tokenization. *arXiv preprint arXiv:2602.04683*, 2026c.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. Reshaping representation space to balance the safety and over-rejection in large audio language models. *arXiv preprint arXiv:2505.19670*, 2025c.

Jiaxi Yang, Chaewan Chun, Jason Lucas, Yuchen Yang, and Dongwon Lee. Aor-bench: Do large audio language models over-refuse pseudo-harmful queries?, 2026d. URL `https://arxiv.org/abs/2606.21147`.

Qingran Yang, Botao Zhao, Zuheng Kang, Xue Li, Yayun He, Chuhang Liu, Xulong Zhang, Xiaoyang Qu, Junqing Peng, and Jianzong Wang. Attention-weighted centered kernel alignment for knowledge distillation in large audio-language models applied to speech emotion recognition. *arXiv preprint arXiv:2602.01547*, 2026e.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. Speechr: A benchmark for speech reasoning in large audio-language models. *arXiv preprint arXiv:2508.02018*, 2025d.

Yiqun Yao, Xiang Li, Xin Jiang, Xuezhi Fang, Naitong Yu, Wenjia Ma, Aixin Sun, and Yequan Wang. Flm-audio: Natural monologues improves native full-duplex chatbots via dual training. *arXiv preprint arXiv:2509.02521*, 2025.

Jashin Ye, Dongxiao Wang, Yixuan Ye, Sashuai Zhou, Weihuang Lin, Mingyang Han, Kunpeng Wang, Zeyu Yuan, Boyu Li, Haoxiang Shi, Jingchen Shu, Jun Song, and Bo Zheng. Voicegiraffe: A benchmark for extreme long-context audio-language understanding, 2026. URL `https://arxiv.org/abs/2605.27976`.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.

Yuhuan You, Lai Wei, Xihong Wu, and Tianshu Qu. The world is not mono: Enabling spatial understanding in large audio-language models. *arXiv preprint arXiv:2601.02954*, 2026.

Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6216–6226, 2025.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*, 2024.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Xiao Zhan, Guangzhi Sun, Jose Such, and Phil Woodland. Protecting bystander privacy via selective hearing in lalms. *arXiv preprint arXiv:2512.06380*, 2025.

Dan Zhang, Yishu Lei, Jing Hu, Shuwei He, Songhe Deng, Xianlong Luo, Danxiang Zhu, Shikun Feng, Rui Liu, Jingzhou He, et al. Eureka-audio: Triggering audio intelligence in compact language models. *arXiv preprint arXiv:2602.13954*, 2026a.

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

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechgpt-gen: Scaling chain-of-information speech generation. *arXiv preprint arXiv:2401.13527*, 2024a.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025a.

Haoyang Zhang, Jun Chen, Donghang Wu, Yuxin Li, Yuxin Zhang, Xiangyu Tony Zhang, Che Liu, Qingjian Lin, Yizhou Peng, Hexin Liu, et al. Duplexsla: A full-duplex spoken language model with synchronized speech, language, and action. *arXiv preprint arXiv:2605.20755*, 2026b.

Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. Wildspeech-bench: Benchmarking end-to-end speechllms in the wild. *arXiv preprint arXiv:2506.21875*, 2025b.

Pengfei Zhang, Hoang H Nguyen, Kazi Shaharair Sharif, Yutong Song, Wenjun Huang, Henry Peng Zou, Pinxin Liu, Honghui Xu, and Amir M. Rahmani. From sounds to scenes: A benchmark for evaluating context-aware auditory scene understanding in large audio language models, 2026c. URL `https://arxiv.org/abs/2606.25391`.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, et al. Omniflatten: An end-to-end gpt model for seamless voice conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14570–14580, 2025c.

Ruixing Zhang, Zihan Liu, Leilei Sun, Tongyu Zhu, and Weifeng Lv. The sonar moment: Benchmarking audio-language models in audio geo-localization. *arXiv preprint arXiv:2601.03227*, 2026d.

Xin Zhang, Xiang Lyu, Zhihao Du, Qian Chen, Dong Zhang, Hangrui Hu, Chaohong Tan, Tianyu Zhao, Yuxuan Wang, Bin Zhang, et al. Intrinsicvoice: Empowering llms with intrinsic real-time voice interaction abilities. *arXiv preprint arXiv:2410.08035*, 2024b.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. Beyond the turn-based game: Enabling real-time conversations with duplex models. *Conference on Empirical Methods in Natural Language Processing*, pp. 11543–11557, 2024c.

Yibo Zhang, Liang Lin, Kaiwen Luo, Shilinlu Yan, Jin Wang, Yaoqi Guo, Yitian Chen, Yalan Qin, Zhenhong Zhou, Kun Wang, et al. Rsa-bench: Benchmarking audio large models in real-world acoustic scenarios. *arXiv preprint arXiv:2601.10384*, 2026e.

Yuxin Zhang, Xiangyu Tony Zhang, Daijiao Liu, Fei Tian, Yayue Deng, Jun Chen, Qingjian Lin, Haoyang Zhang, Yuxin Li, Jinglan Gong, et al. Step-audio-r1. 5 technical report. *arXiv preprint arXiv:2604.25719*, 2026f.

Feiyu Zhao, Yiming Chen, Wenhuan Lu, Daipeng Zhang, Xianghu Yue, and Jianguo Wei. Halluaudio: A comprehensive benchmark for hallucination detection in large audio-language models, 2026a. URL `https://arxiv.org/abs/2604.19300`.

Junqi Zhao, Chenxing Li, Jinzheng Zhao, Rilin Chen, Dong Yu, Mark D Plumbley, and Wenwu Wang. Feedback-driven retrieval-augmented audio generation with large audio language models. *arXiv preprint arXiv:2511.01091*, 2025a.

Shuaijiang Zhao, Tingwei Guo, Cheng Wen, Bajian Xiang, and Wei Zou. Ke-omni-r: Achieving advanced audio reasoning with a concise 50-words think process. `https://github.com/shuaijiang/Ke-Omni-R`, 2025b. GitHub repository.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. It hears, it sees too: Multi-modal llm for depression detection by integrating visual understanding into audio language models. *arXiv preprint arXiv:2511.19877*, 2025c.

Xiangyu Zhao, Junyu Yan, Yaling Shen, Zimu Wang, Yiwen Jiang, Stephanie Fong, Qingyang Xu, Jiahe Liu, Dominic Dwyer, and Zongyuan Ge. Audioprocessbench: Benchmark for identifying process errors in audio-grounded reasoning, 2026b. URL https://arxiv.org/abs/2606.09925.

Xuanru Zhou, Jiachen Lian, Henry Hong, Xinyi Yang, and Gopala Anumanchipalli. Speech world model: Causal state-action planning with explicit reasoning for speech. *arXiv preprint arXiv:2512.05933*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Appendix

A.1 Release-Date Convention

To ensure consistent chronological comparisons, all works in this survey are organized according to their earliest publicly verifiable release date rather than the publication year of the version cited in the bibliography. This convention is applied consistently to Tables 1–5 and the timeline-based analyses in Figures 1 and 5.

For research papers and surveys, we use the date of the initial arXiv submission whenever available. If no arXiv version exists, we use the earliest verifiable date of an official technical report, project page, or public repository. For models and systems, the release date refers to their first public availability through an official technical report, model repository, model card, or announcement. For benchmarks, it refers to the earliest public release of the associated paper, dataset, evaluation code, or official repository.

When multiple public versions are available, later revisions, repository updates, or peer-reviewed publications do not change the chronological placement of the work. The bibliography cites the version consulted in this survey and may therefore display a later year. Consequently, a work grouped under one release year may have a different bibliographic publication year.

The literature search and release records were last updated and verified on June 30, 2026.

A.2 Detailed Comparison of Large Audio Language Models

Table 4 extends the compact comparison in Table 2. In addition to release date, parameter scale, full-duplex capability, and supported modalities, the detailed table reports the base LLM, supported languages, audio input representation, and publicly disclosed pre-training data scale. A dash indicates that the corresponding information was not publicly reported or could not be reliably verified from the available technical documentation.

Table 4: Summary of Large Audio Language Models from 2022 to 2026

Model	Institute	Release	Base LLM	Base LLM Params	Lang.	Input Repr.	Pre-train Data Scale	Full-Duplex	Multimodality	
									Text	Audio
Year 2022										
dGSLM Nguyen et al. (2023)	Meta	Mar 2022	-	-	EN	Discrete	2K Hrs audio	✗	✗	✓
Year 2023										
SpeechGPT Zhang et al. (2023)	OpenLMLab	May 2023	LLaMA-13B	13B	EN	Discrete	60K Hrs audio + 9M unit-text pairs + 37,969 quadruplets	✗	✓	✓
Pengi Deshmukh et al. (2023)	Microsoft	May 2023	GPT-2	124M	EN	Contin.	3.4M audio-text pairs	✗	✓	✓
LTU Ramaswamy (2025)	MIT	May 2023	LLaMA-7B	7B	EN	Contin.	1.9M closed + 3.7M open-ended AQA pairs	✗	✓	✓
Spectron Nachmani et al. (2024)	Google	May 2023	-	350M/1B	EN	Contin.	-	✗	✓	✓
AudioPaLM Rubenstein et al. (2023)	Google	Jun 2023	PaLM-2	8B	Multi.	Discrete	-	✗	✓	✓
MU-LLaMA Liu et al. (2024b)	Adobe	Aug 2023	LLaMA-2-7B	7B	EN	Contin.	-	✗	✓	✓
LTU-AS Gong et al. (2023)	MIT	Sep 2023	LLaMA-7B	7B	EN	Contin.	9.6M Open-ASQA	✗	✓	✓
SLM Wang et al. (2023)	Google	Sep 2023	mT0-MT XXL	13B	Multi.	Contin.	-	✗	✓	✓
SALMONN Tang et al. (2023)	Tencent	Oct 2023	Vicuna-13B	13B	EN, CN	Contin.	4760 Hrs audio	✗	✓	✓
LauraGPT Du et al. (2023)	Alibaba	Oct 2023	Qwen-1.8B	2B	EN, CN	Contin.	-	✗	✓	✓
Qwen-AudioChu et al. (2023)	Alibaba	Nov 2023	Qwen-7B	7B	Multi.	Contin.	130K+ Hrs audio	✗	✓	✓
ParalinGPT Lin et al. (2024b)	Meta	Dec 2023	DialoGPT	345M	EN	Contin.	140 Hrs audio	✗	✓	✓
E-chat Xue et al. (2024)	Tencent	Dec 2023	Baichuan2-7B-Chat	7B	CN	Contin.	10K Hrs ASR data	✗	✓	✓
Year 2024										
SpeechGPT-Gen Zhang et al. (2024a)	OpenLMLab	Jan 2024	LLaMA-2-7B-Chat	7B	EN	Discrete	-	✗	✓	✓
Audio Flamingo Kong et al. (2024)	NVIDIA	Feb 2024	OPT-IML-1.3B	1.3B	EN	Contin.	21K Hrs audio	✗	✓	✓
Spoken-LLM Lin et al. (2024a)	Meta	Feb 2024	Llama-2-7B-Chat	7B	EN	Contin.	16,472 current-response speech pairs	✗	✓	✓
Spirit LM Nguyen et al. (2025)	Meta	Feb 2024	Llama-2-7B	7B	EN	Discrete	35.2B tokens	✗	✓	✓
USDM Li et al. (2024)	Meta	Feb 2024	Mistral-7B	7B	EN	Discrete	87K Hrs audio	✗	✓	✓
WavLLM Hu et al. (2024)	Microsoft	Mar 2024	LLaMA-2-7B-Chat	7B	EN	Contin.	-	✗	✓	✓
SpeechVerse Das et al. (2024)	Amazon	May 2024	Flan-T5-XL	3B	EN	Contin.	-	✗	✓	✓
GAMA Ghosh et al. (2024)	Maryland	Jun 2024	LLaMA2-7B	7B	EN	Contin.	2.2M audio-caption pairs	✗	✓	✓
Qwen2-Audio Chu et al. (2024)	Alibaba	Jul 2024	Qwen-7B	7B	Multi.	Contin.	520K Hrs audio	✗	✓	✓
FunAudioLLM An et al. (2024)	Alibaba	Jul 2024	-	-	Multi.	-	-	✗	✓	✓
Mini-Omni Xie & Wu (2024)	Tencent	Aug 2024	Qwen2-0.5B	0.5B	-	Discrete	8K Hrs speech + 2M text examples	✓	✓	✓
Moshi Défossez et al. (2024)	Kyutai	Sep 2024	Helium	7B	EN	Discrete	7M Hrs audio + 2.1T text tokens	✓	✓	✓
LLaMA-Omni Fang et al. (2024)	ICT	Sep 2024	Llama-3.1-8B-Instruct	8B	EN	Contin.	-	✗	✓	✓
Parrot Meng et al. (2024)	Apple	Sep 2024	Llama 3.1-8B	8B	EN	Discrete	74,554 Hrs audio	✓	✗	✓
OmniFlatten Zhang et al. (2025c)	Alibaba	Oct 2024	Qwen2-0.5B	0.5B	EN, CN	Discrete	-	✓	✓	✓
IntrinsicVoice Zhang et al. (2024b)	OpenLMLab	Oct 2024	Qwen2-7B-Instruct	7B	-	Discrete	20K Hrs audio	✗	✓	✓
DiVA Held et al. (2025)	Stanford	Oct 2024	Llama 3	8B	EN	Contin.	-	✗	✓	✓
Freeze-Omni Wang et al. (2024)	VITA	Nov 2024	Qwen2-7B-Instruct	7B	EN, CN	Contin.	-	✓	✓	✓
GLM-4-Voice Zeng et al. (2024)	Zhipu	Dec 2024	GLM-4-9B	9B	EN, CN	Discrete	1T tokens	✗	✓	✓
KE-Omni Zhao et al. (2025b)	Kuaishou	Dec 2024	LLaMA-3.1-8B-Instruct	8B	EN, CN	Contin.	-	✗	✓	✓
MERaLiON-Audio He et al. (2024)	A*STAR	Dec 2024	SEA-LION V3	10B	Multi.	Contin.	-	✗	✓	✓

Table 4: Summary of Large Audio Language Models from 2022 to 2026 (continued)

Model	Institute	Release	Base LLM	Base LLM Params	Lang.	Input Repr.	Pre-train Data Scale	Full-Duplex	Multimodality	
									Text	Audio
Year 2025										
MinMo Zhang et al. (2025a)		Jan 2025	Qwen2.5-7B-Instruct	7B	Multi.	Contin.	-	✓	✓	✓
FireRedASR Shi et al. (2026)		Jan 2025	Qwen2-7B-Instruct	7B	Multi.	Contin.	-	✗	✓	✓
Step-Audio Tian et al. (2025)		Feb 2025	Step-1	130B	Multi.	Discrete	3.3T tokens	✗	✓	✓
Baichuan-Audio Li et al. (2025f)		Feb 2025	Baichuan-Audio-Base	7B	EN, CN	Discrete	887K Hrs audio + 100B tokens	✗	✓	✓
Audio Flamingo 2 Ghosh et al. (2025)		Mar 2025	Qwen2.5-3B	3B	EN	Contin.	8M+ audio-caption pairs	✗	✓	✓
Kimi-Audio Ding et al. (2025)		Apr 2025	Qwen2.5-7B	7B	EN, CN	Hybrid	13M+ Hrs audio	✗	✓	✓
VITA-Audio Long et al. (2025)		May 2025	Qwen2.5-7B-Instruct	7B	EN, CN	Discrete	200K Hrs audio	✗	✓	✓
Step-Audio 2 Wu et al. (2025a)		Jul 2025	-	-	Multi.	Contin.	680B tokens and 8M Hrs audio	✗	✓	✓
Audio Flamingo 3 Goel et al. (2025)		Jul 2025	Qwen2.5-7B	7B	EN	Contin.	-	✗	✓	✓
DeSTA2.5-Audio Lu et al. (2026a)		Jul 2025	Llama3.1-8B-Instruct	8B	EN	Contin.	7K Hrs audio	✗	✓	✓
FireRedChat Chen et al. (2025d)		Sep 2025	Qwen2.5	-	EN, CN	-	-	✓	✓	✓
Falcon3-Audio Kumar et al. (2025a)		Sep 2025	Falcon3-Instruct	1/3/7B	EN	Contin.	-	✗	✓	✓
Step-Audio-R1 Tian et al. (2025)		Nov 2025	Qwen2.5-32B	32B	EN, CN	Contin.	1.356T tokens	✗	✓	✓
Step-Audio-EditX Yan et al. (2025a)		Nov 2025	-	3B	Multi.	Discrete	-	✗	✓	✓
SeaLLMs-Audio Liu et al. (2025a)		Nov 2025	Qwen2.5-7B	7B	Multi.	Contin.	-	✗	✓	✓
Fun-Audio-Chat Team et al. (2025)		Dec 2025	Qwen3	8/30B	EN, CN	Discrete	-	✗	✓	✓
MiMo-Audio Zhang et al. (2025a)		Dec 2025	MiMo-7B-Base	7B	Multi.	Discrete	100M+ Hrs audio	✗	✓	✓
Year 2026										
Step-Audio-R1.1 Tian et al. (2025)		Jan 2026	Qwen2.5-32B	32B	EN, CN	Contin.	1.356T tokens	✗	✓	✓
Qwen3-ASR Shi et al. (2026)		Jan 2026	Qwen3	0.6/1.7B	Multi.	Contin.	40M Hrs audio + 3T tokens	✗	✓	✓
Covo-Audio Wang et al. (2026e)		Feb 2026	Qwen2.5-7B-Base	7B	EN, CN	Contin.	2T tokens	✓	✓	✓
Eureka-Audio Zhang et al. (2026a)		Feb 2026	Qwen3-1.7B-Base	1.7B	EN, CN	Contin.	1T tokens	✗	✓	✓
Audio Flamingo Next Ghosh et al. (2026)		Apr 2026	Qwen2.5-7B	7B	Multi.	Contin.	363.5K Hrs audio	✗	✓	✓
Step-Audio-R1.5 Zhang et al. (2026f)		Apr 2026	Qwen2.5-32B	32B	EN, CN	Contin.	-	✗	✓	✓
DuplexSLA Zhang et al. (2026b)		May 2026	Qwen2.5-7B	7B	EN, CN	Contin.	500K Hrs audio + 1.92M text samples	✓	✓	✓
StepAudio 2.5 Lin et al. (2026)		May 2026	-	-	EN, CN	Contin.	2.2T text and audio tokens	✗	✓	✓
MOSS-Audio Yang et al. (2026a)		Jun 2026	Qwen3-4B/8B	4B/8B	EN, CN	Contin.	1.2T tokens	✗	✓	✓
BayLing-Duplex Fang et al. (2026)		Jun 2026	GLM-4-9B	9B	EN, CN	Discrete	-	✓	✓	✓
FlexiSLM Li et al. (2026)		Jun 2026	Qwen2.5-7B-Instruct	7B	EN	Contin.	-	✗	✓	✓

Note: This table only summarizes large language models with audio modality, excluding models with image or video modality support. "Lang." is short for language, where "EN" denotes English, "CN" denotes Chinese, and "Multi." is short for multiple languages, indicating support for more than two languages. "Input Repr." is short for input representation, and "Contin." is short for continuous representation.

A.3 Detailed Comparison of LALM Evaluation Benchmarks

Table 5 extends the benchmark coverage summary in Table 3 by reporting the original evaluation metrics used by each benchmark. The general-capability dimensions comprise perception, reasoning, and interaction, while the trustworthiness dimensions comprise hallucination, privacy, authentication, safety, robustness, and fairness.

A check mark is assigned only when a benchmark explicitly includes evaluation examples, annotations, or metrics for the corresponding dimension. A dimension that is mentioned only as motivation or future work is not marked as covered. This criterion is applied consistently to both the compact and detailed tables.

Table 5: Overview of LALM evaluation benchmarks across general capabilities and trustworthy dimensions.

Benchmark	Release	Metrics	General[†]			Trustworthy[‡]					
			PE	RE	IN	H	P	A	S	R	F
Year 2024											
AudioBench Wang et al. (2025a)	Jan 2024	WER, METEOR, LLM-as-a-Judge	✓	✓	✗	✗	✗	✗	✗	✓	✗
MMAU Sakshi et al. (2024)	Oct 2024	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
VoiceBench Chen et al. (2026b)	Oct 2024	LLM-as-a-Judge, Accuracy, Refusal Rate	✗	✓	✗	✗	✗	✗	✗	✓	✗
Year 2025											
Jailbreak-AudioBench Cheng et al. (2025)	Jan 2025	ASR	✗	✗	✗	✗	✗	✗	✓	✗	✗
URO-Bench Yan et al. (2025b)	Feb 2025	LLM-as-a-Judge, UTMOS, WER/CER, First Packet Latency	✓	✓	✓	✗	✓	✓	✓	✗	✗
S2S-Arena Jiang et al. (2025)	Mar 2025	ELO Rating, Win Rate, Inter-Annotator Agreement	✓	✗	✓	✗	✗	✗	✗	✗	✗
Talking Turns Arora et al. (2025)	Mar 2025	Agreement, Accuracy, ROC-AUC, F1 (judge model)	✓	✓	✓	✗	✗	✗	✗	✗	✗
MMAR Ma et al. (2025c)	May 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✓	✗
SAKURA Yang et al. (2025a)	May 2025	Accuracy, LLM-as-a-Judge	✓	✓	✗	✗	✗	✗	✗	✗	✗
VocalBench Liu et al. (2025b)	May 2025	Accuracy, LLM-as-a-Judge, UTMOS, WER, Refusal Rate, Preserve Rate	✓	✓	✓	✗	✗	✗	✓	✓	✗
JALMBench Peng et al. (2025c)	May 2025	ASR, Attack Efficiency	✗	✗	✗	✗	✗	✗	✓	✗	✗
AHaBench Cheng et al. (2026)	May 2025	AHa-Score, m-AHa, GPT-4-Judge Consistency, POPE	✗	✗	✓	✓	✗	✗	✗	✗	✗
AudioJailbreak Chen et al. (2025a)	May 2025	ASR	✗	✗	✗	✗	✗	✗	✓	✗	✗
AJailBench Song et al. (2025)	May 2025	ASR, TS, PV, Relevance, Similarity	✗	✗	✗	✗	✗	✗	✓	✗	✗
VocalAgent Kim et al. (2025c)	May 2025	macro-F1, Accuracy, FPR, Refusal Rate, Goodness@0.1	✓	✗	✗	✗	✗	✗	✗	✓	✗
AudioTrust Li et al. (2025c)	May 2025	GPT-4o Score, CM-WER, CCR, DSR, HRR, FAR, SES, Refusal Rate, Group Unfairness Score	✗	✗	✗	✓	✓	✓	✓	✓	✓
MMSU Wang et al. (2025c)	Jun 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
SOVA-Bench Hou et al. (2025b)	Jun 2025	Accuracy, WER, GPTEval, LLM-as-a-Judge, UTMOSv2	✓	✓	✓	✗	✗	✗	✗	✗	✗
WildSpeech-Bench Zhang et al. (2025b)	Jun 2025	LLM-as-a-Judge, UTMOS, Query-Aware Checklist	✓	✓	✓	✗	✗	✗	✗	✓	✗
ContextASR-Bench Wang et al. (2025d)	Jul 2025	WER, NE-WER, NE-FNR	✓	✗	✗	✗	✗	✗	✗	✗	✗
C^3Ma et al. (2025a)	Jul 2025	LLM-as-a-Judge, Human Evaluation Score	✓	✓	✓	✗	✗	✗	✗	✓	✗
WoW-Bench Kim et al. (2025b)	Aug 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
MMAU-Pro Kumar et al. (2025b)	Aug 2025	Accuracy	✓	✓	✓	✗	✗	✗	✗	✗	✗
MCR-BENCH Wang et al. (2025b)	Aug 2025	Accuracy, Norm Acc, Macro Acc, TIR, MRS	✓	✓	✗	✓	✗	✗	✗	✗	✗
SpeechR Yang et al. (2025d)	Aug 2025	Accuracy, Logical Consistency Score	✗	✓	✗	✗	✗	✗	✗	✗	✗
AudioSafe Lin et al. (2025b)	Aug 2025	Accuracy, ASR	✗	✗	✗	✗	✗	✗	✓	✓	✗
VoiceAssistant-Eval Wang et al. (2025f)	Sep 2025	LLM-as-a-Judge, UTMOS, WER, Speaker Similarity	✓	✓	✓	✗	✗	✗	✓	✓	✗
Gaslighting Attacks Wu et al. (2026c)	Sep 2025	Accuracy, Behavioral Consistency	✗	✗	✓	✗	✗	✗	✗	✓	✗
MUSE Carone et al. (2025)	Oct 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
LISTEN Chen et al. (2025c)	Oct 2025	Accuracy	✓	✓	✗	✓	✗	✗	✗	✗	✗
AudioMarathon He et al. (2025b)	Oct 2025	F1, WAR	✓	✓	✗	✗	✗	✗	✗	✗	✗
ISA-Bench Li et al. (2025a)	Oct 2025	IFR, WER, BLEU, ACC, METEOR, RPS	✓	✗	✗	✗	✗	✗	✓	✗	✗
Hearing the Order Lin et al. (2025d)	Oct 2025	Accuracy, CKLD	✗	✗	✗	✗	✗	✗	✗	✓	✗
Safety under Emotional Variations Feng et al. (2025)	Oct 2025	NRR, Unsafe Rate	✗	✗	✗	✗	✗	✗	✓	✓	✗
Gender Bias in SpeechLLMs Satish et al. (2025)	Oct 2025	MCQA Selection Rate, LLM-as-a-Judge	✗	✗	✗	✗	✗	✗	✗	✗	✓
BRACE Guo et al. (2025b)	Dec 2025	F1	✓	✗	✗	✓	✗	✗	✗	✗	✗
Spoken DialogSum Lu et al. (2025a)	Dec 2025	ROUGE-L	✗	✗	✓	✗	✗	✗	✗	✗	✗
MAC-SLU Peng et al. (2025b)	Dec 2025	Accuracy, F1	✓	✗	✗	✗	✗	✗	✗	✗	✗
SH-Bench Zhan et al. (2025)	Dec 2025	SE (Selective Efficacy), Accuracy	✓	✗	✗	✗	✓	✗	✗	✗	✗

Table 5: Overview of LALM evaluation benchmarks across general capabilities and trustworthy dimensions (continued).

Benchmark	Release	Metrics	General†			Trustworthy‡					
			PE	RE	IN	H	P	A	S	R	F
Year 2026											
WESR Yang et al. (2026b)	Jan 2026	Precision, Recall, F1, WER	✓	✗	✗	✗	✗	✗	✗	✗	✗
RSA-Bench Zhang et al. (2026e)	Jan 2026	WER, Accuracy, LLM-as-a-Judge	✓	✓	✗	✗	✗	✗	✗	✓	✗
PALM-Bench Wang et al. (2026g)	Jan 2026	BLEU, F1-score, BERTScore, LLM-as-a-Judge	✗	✓	✓	✓	✗	✗	✗	✗	✗
ChronosAudio Luo et al. (2026)	Jan 2026	WER, BERTScore, Acc_dict, LS, TS, MS, LLM-as-a-Judge	✓	✓	✗	✓	✗	✗	✗	✗	✗
HearSay Wang et al. (2026b)	Jan 2026	IAR, ARR, BBR	✗	✗	✗	✗	✓	✗	✗	✗	✗
AGL1K Zhang et al. (2026d)	Jan 2026	Accuracy, Geoscore, Mean Distance Error, Thresholded Accuracy, Reject Rate	✗	✓	✗	✗	✗	✗	✗	✗	✓
VoxPrivacy Wang et al. (2026h)	Jan 2026	Accuracy, Precision, Recall, F1-Score, IRR, EER	✗	✗	✓	✗	✓	✓	✗	✗	✗
BiasInEar Wei et al. (2026)	Feb 2026	Accuracy, Entropy, APES, Fleiss' κ	✗	✗	✗	✗	✗	✗	✗	✗	✓
DailyTalkEdit Xu et al. (2026)	Feb 2026	Accuracy, EER, F1, Seg-F1, LLM-as-a-Judge	✗	✗	✗	✗	✗	✓	✗	✗	✗
HumDial-EIBench Wang et al. (2026d)	Apr 2026	Accuracy, LLM-as-Judge	✗	✓	✓	✗	✗	✗	✗	✓	✗
VoxSafeBench Wang et al. (2026i)	Apr 2026	Safety Awareness Score, JSR, Perception Probing Accuracy	✓	✗	✓	✗	✓	✗	✓	✗	✓
HalluAudio Zhao et al. (2026a)	Apr 2026	Accuracy, Hallucination Rate, Yes/No Bias, Error-type Analysis, FRR	✓	✓	✗	✓	✗	✗	✗	✓	✗
MUSA Koo (2026)	May 2026	Accuracy, ECE	✓	✓	✗	✗	✗	✗	✗	✓	✗
PitchBench Dujardin et al. (2026)	May 2026	Accuracy	✓	✗	✗	✗	✗	✗	✗	✓	✗
EnvMem Xiao et al. (2026)	May 2026	Accuracy, $\Delta(N)$, CKA, cov_{90}	✓	✗	✓	✗	✗	✗	✗	✓	✗
VoiceGiraffe Ye et al. (2026)	May 2026	Macro-Average Accuracy	✓	✓	✗	✗	✗	✗	✗	✓	✗
VoxParadox Pang et al. (2026a)	May 2026	GT Accuracy, ALA	✓	✗	✗	✓	✗	✗	✗	✓	✗
MusTBENCH Kwon et al. (2026)	May 2026	Accuracy, Hit@T, METEOR, CLAP Score, Temporal IoU, F1	✓	✓	✗	✓	✗	✗	✗	✗	✗
SpeechJBB Ceccatelli et al. (2026)	Jun 2026	RR, DR, JSR	✓	✗	✗	✗	✗	✗	✓	✓	✗
GlobeAudio Tan & Zhang (2026)	Jun 2026	Accuracy	✓	✓	✗	✗	✗	✗	✗	✓	✓
AudioProcessBench Zhao et al. (2026b)	Jun 2026	PRMScore, Balanced Accuracy, AUROC, BoN/MV Accuracy	✓	✓	✗	✓	✗	✗	✗	✗	✗
RAIL Jin et al. (2026)	Jun 2026	Accuracy, LLM-as-a-Judge, B-AUC	✓	✓	✗	✗	✗	✗	✗	✗	✗
IHBench Salimi et al. (2026)	Jun 2026	TF Win Rate, RQ Pass Rate	✗	✓	✓	✗	✗	✗	✗	✓	✗
AOR-Bench Yang et al. (2026d)	Jun 2026	ORR, TRR, MB-Score	✓	✓	✗	✗	✗	✗	✓	✓	✗
MSU-Bench Sun et al. (2026)	Jun 2026	Exact-Match Accuracy, Error-type Analysis	✓	✓	✓	✓	✗	✗	✗	✗	✗
ParaPairAudioBench Jeon et al. (2026)	Jun 2026	Accuracy, Acc@A, Acc@B, Position-Bias Gap	✓	✗	✗	✗	✗	✗	✗	✓	✗
CASU Zhang et al. (2026c)	Jun 2026	BLEU-4, BERTScore, LLM-as-a-Judge, 1-WER, Accuracy	✓	✓	✗	✓	✗	✗	✗	✗	✗
SpeechEQ Wu et al. (2026d)	Jun 2026	Acc_1, Acc_2, Acc_{traj}, SEQ	✓	✓	✓	✗	✗	✗	✗	✓	✗

† **General**: Perception (PE), Reasoning (RE), Interaction (IN).
‡ **Trustworthy**: Hallucination (H), Privacy (P), Authentication (A), Safety (S), Robustness (R), Fairness (F).