NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

Kaiwen Luo[1,*], **Zhenhong Zhou**[1,*], **Leo Wang**[2,*], **Liang Lin**[1,*,†], **Yang Xiao**[3], **Tianyu Shao**[4], **Yuanhe Zhang**[5], **Yuxuan Li**[6], **Miao Yu**[7], **Kailin Lyu**[8], **Jiaming Zhang**[1], **Dongrui Liu**[9], **Li Sun**[5], **Yueming Wu**[10], **Kai Li**[11], **Ting Dang**[3], **Xiaojun Jia**[1], **Rohan Kumar Das**[12], **Xinfeng Li**[1], **Siyuan Liang**[1], **Qiufeng Wang**[13], **Xingjun Ma**[14], **Jing Chen**[15], **Kun Wang**[1,✉], **Junhao Dong**[1,✉], **Deqing Zou**[10], **Yu Cheng**[16], **Xia Hu**[9], **Zhigang Zeng**[10], **Sen Su**[17], **Yang Liu**[1], **Yu-Gang Jiang**[14], **Philip S. Yu**[18], **Yew-Soon Ong**[1]

Affiliations: [1] Nanyang Technological University, [2] Independent Researcher, [3] The University of Melbourne, [4] North China Electric Power University, [5] Beijing University of Posts and Telecommunications, [6] University of Chinese Academy of Sciences, [7] University of Science and Technology of China, [8] Institute of Automation, Chinese Academy of Sciences, [9] Shanghai AI Laboratory, [10] Huazhong University of Science and Technology, [11] Tsinghua University, [12] Fortemedia Singapore, [13] Tencent, [14] Fudan University, [15] Wuhan University, [16] Chinese University of Hong Kong, [17] Chongqing University of Posts and Telecommunications, [18] University of Illinois Chicago

Abstract: The foundational capabilities established by Large Language Models (LLMs) have paved the way for Multimodal Large Language Models (MLLMs), within which 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 inherently 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 through six analytical pillars: hallucination, robustness, safety, privacy, fairness, and authentication. The profound imbalance between a mature offensive landscape and underdeveloped defenses further validates the critical trustworthiness gaps and multidimensional risks facing audio-centric intelligence. Finally, we propose a strategic roadmap advocating for "Defense-in-Depth" architectures, causal auditory world modeling, and intrinsic representation engineering to bridge the gap between empirical performance and intrinsically trustworthy audio intelligence.

* These authors contributed equally.

† Project leader.

✉ Corresponding authors.

Project: https://github.com/Kwwwww74/Awesome-Trustworthy-AudioLLMs

Note: If you identify your own or other papers relevant to this survey that have not been discussed (we apologize for any such omissions due to the rapidly expanding literature), please feel free to contact us via email or raise an issue on GitHub.

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

Kaiwen Luo[1,*], Zhenhong Zhou[1,*], Leo Wang[2,*], Liang Lin[1,*], Yang Xiao[3], Tianyu Shao[4], Yuanhe Zhang[5], Yuxuan Li[6], Miao Yu[7], Kailin Lyu [8], Jiaming Zhang[1], Dongrui Liu[9], Li Sun[5], Yueming Wu[10], Kai Li[11], Ting Dang[3], Xiaojun Jia[1], Rohan Kumar Das[12], Xinfeng Li[1], Siyuan Liang[1], Qiufeng Wang[13], Xingjun Ma[14], Jing Chen[15], Kun Wang[1,†] Junhao Dong[1,†], Deqing Zou[10], Yu Cheng[16], Xia Hu[9], Zhigang Zeng[10], Sen Su[17], Yang Liu[1], Yu-Gang Jiang[14], Philip S. Yu[18], Yew-Soon Ong[1]

[1]Nanyang Technological University [2]Independent Researcher [3]The University of Melbourne
[4]North China Electric Power University [5]Beijing University of Posts and Telecommunications
[6]University of Chinese Academy of Sciences [7]University of Science and Technology of China
[8] Institute of Automation, Chinese Academy of Sciences [9]Shanghai AI Laboratory
[10]Huazhong University of Science and Technology [11]Tsinghua University [12]Fortemedia Singapore
[13]Tencent [14]Fudan University [15]Wuhan University [16]Chinese University of Hong Kong
[17]Chongqing University of Posts and Telecommunications [18]University of Illinois Chicago

Abstract—The foundational capabilities established by Large Language Models (LLMs) have paved the way for Multimodal Large Language Models (MLLMs), within which 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 inherently 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 through six analytical pillars: hallucination, robustness, safety, privacy, fairness, and authentication. The profound imbalance between a mature offensive landscape and underdeveloped defenses further validates the critical trustworthiness gaps and multidimensional risks facing audio-centric intelligence. Finally, we propose a strategic roadmap advocating for "Defense-in-Depth" architectures, causal auditory world modeling, and intrinsic representation engineering to bridge the gap between empirical performance and intrinsically trustworthy audio intelligence. Our project has been uploaded to GitHub https://github.com/Kwwwww74/Awesome-Trustworthy-AudioLLMs.

Index Terms—Large Audio Language Models (LALMs), Trustworthiness, Multimodal Safety, Auditory Reasoning, Cross-modal Alignment, Artificial Intelligence Security.

1 Introduction

The emergence of Large Language Models (LLMs) [1]–[6] 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) [7]–[9] 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 [10], 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 [11], [12] or sound classification [13], [14]. Latest transition from these artifacts to unified **Large Audio Language Models (LALMs)** [15]–[19] 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 [20]–[23]. In contrast, LALMs introduce the audio modality, which presents a intricate risk landscape [24]–[27] 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 [28]–[30] or specific concerns [31]–[33], yet there remains a significant lack of work dedicated to a systematic taxonomy of the

Corresponding author: wang.kun@ntu.edu.sg, junhao003@ntu.edu.sg

* *These authors contributed equally.*

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 [10], [34], [35] 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 [36] 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 [37], 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 [38]–[40]. 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	D	P	F	D	E	
Year 2022													
Feng et al. [37]	S	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	✓	✓
Year 2023													
Latif et al. [10]	A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Yi et al. [38]	S	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓
Year 2024													
Li et al. [39]	S	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✗
Pham et al. [40]	S	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓
Peng et al. [34]	A	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
Year 2025													
Su et al. [35]	A	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓
Cui et al. [41]	S+M	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓
Yang et al. [36]	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 (D), Pre-training (P), Fine-tuning (F), Deployment (D), Evaluation (E). **Outook**: Outlook(O)

Upon reviewing the aforementioned survey 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 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 2, figure 1 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 [17], [42]. 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 [16], [18].

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 generally 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 [43]. The investigation of information transfer mechanisms from these encoders to language decoders is essential for optimizing system performance [29]. Moreover, specialized encoding strategies are employed to characterize physical attributes including spatial descriptors [44].

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 [45]. Frameworks like **TWNM** [30] and **SPUR** [28] enhance the adaptability of existing systems. Furthermore, architectural refinements for specialized deployment scenarios are represented by egocentric multichannel processing [46].

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 [47].

Significant structural innovations continue to enhance the efficiency and versatility of the LALM pipeline. These

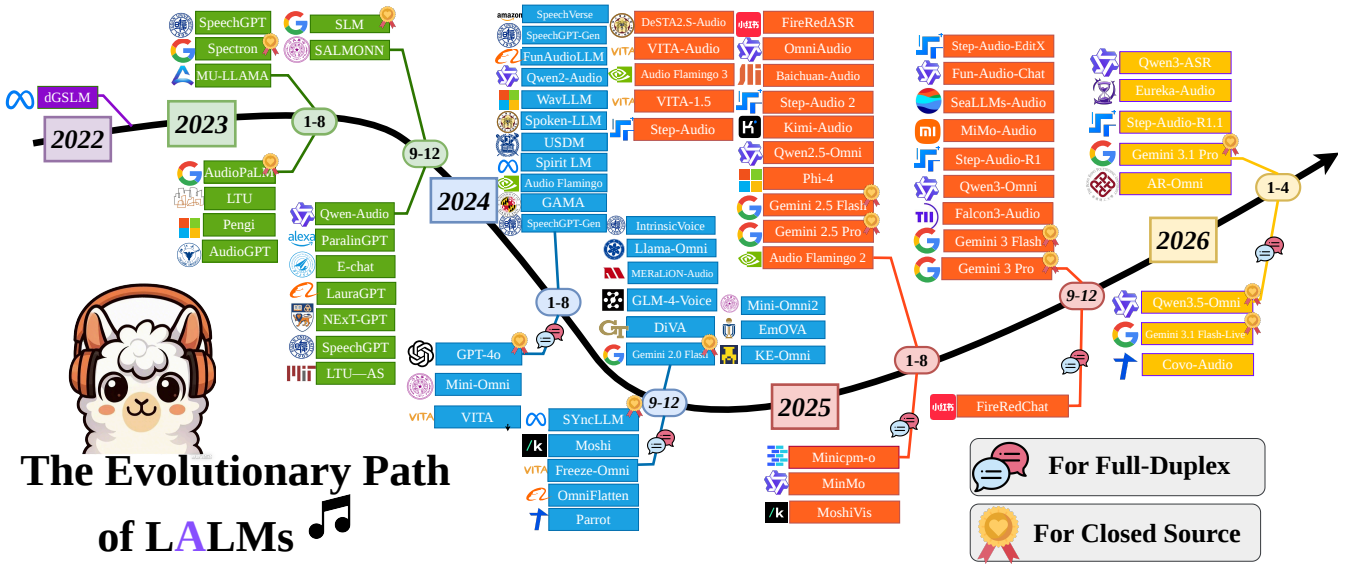

Fig. 1. The Evolutionary Roadmap of LALMs from Cascaded Systems to End-to-End Causal Cognition from 2022 to 2026.

developments include the adoption of structured embeddings for integrated understanding and editing as presented in **SALM** [48]. 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 [49].

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 [50]. To address the scalability constraints of discrete sequences, researchers have developed audio token compression techniques that maintain semantic density while reducing computational overhead [51]. Additionally the capture of paralinguistic nuances is enhanced through fine-grained feature augmentation including vowel-level modifications designed to improve emotional prosody [52]. 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 [53]. Efficiency is further improved through segmentwise pruning techniques that mitigate the overhead of processing continuous streams [54]. For domain specific applications, the utilization of Low Rank Adaptation facilitates precise temporal localization in high stakes environments such as clinical therapy [55].

Systematic evaluation of these optimizations is facilitated by benchmarks [56]. Methodological advancements include the development of extended context mechanisms for long form understanding [57] and techniques designed to bridge temporal gaps between frames to maintain dependency capture [58]. The end-to-end contrastive pretraining models can improve performance for long form question answering capabilities [59].

Sophisticated alignment algorithms are essential to resolve modality bias and improve fidelity of cross modal representations. Research into few shot learning finds that models can achieve high proficiency with minimal data samples [60]. To ensure models rely on acoustic evidence rather than textual shortcuts, researchers employ attention rebalancing mechanisms to mitigate imbalances [61] and utilize audio contribution aware post training to enhance correctness [62]. Specialized alignment strategies [63] further refine this synergy. Knowledge distillation serves as another vital tool with methods transferring reasoning abilities from vision to audio [64] or through weighted on policy cross modal distillation [65]. Advanced metrics like attention weighted centered kernel alignment also contribute to optimized speech emotion recognition [66].

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 [67]. Similarly test time adaptation methods enhance robustness for emotional recognition tasks without additional training [68].

The evolution of LALMs is defined by the transition from rigid turn-taking toward synchronous interaction [69], [70]. This shift necessitates sophisticated architectural schemes that enable real-time conversations beyond the turn-based game [71], including codec-free designs for speech under-

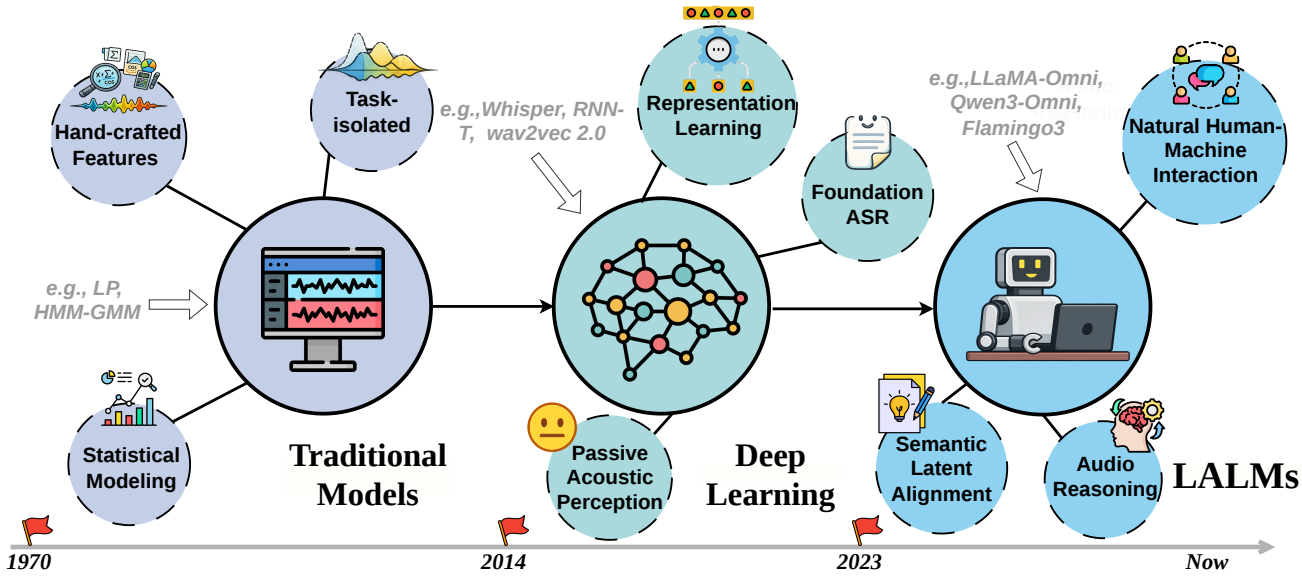

Fig. 2. Architectural and Paradigmatic Evolution from Traditional Audio Models to LALMs.

standing and generation [72]. While some research investigates efficient and direct duplex modeling [73], others highlight the potential of modular systems [74] or propose plug-and-play streaming state prediction modules to ensure real-time responsiveness [75]. Optimization strategies have also advanced through time-controllable training [76], reinforcement learning for interactivity optimization [77], and the use of natural monologues via dual training [78]. And cognitive capabilities within duplex models are being extended through asynchronous knowledge retrieval [79] and latent reasoning to model internal cognition [80]. 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 [81], while technical reports like **Covo-Audio** [82] 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 [83] 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** [84]. 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 [85]. 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 [86].

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 [87]. 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** [88]. Other strategies employ RL to guide models on the optimal timing and methodology for initiating reasoning processes [89]. Such incentivized reasoning is critical for solving complex logical challenges and is fundamental to the advancements presented in **SoundMind** [90].

The adaptability and scalability of reasoning at inference time allow models to navigate ambiguity and high-dimensional tasks. Difficulty-adaptive mechanisms empower models to dynamically allocate computational resources based on instruction complexity [91]. To resolve highly ambiguous emotional cues researchers utilize test-time scaling to expand the computational depth of the model during decoding [92].

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 [93]. Causal state-action planning pushes the boundaries of reasoning by enabling models to simulate and reason through physical world dynamics [94] . The robustness of these rea-

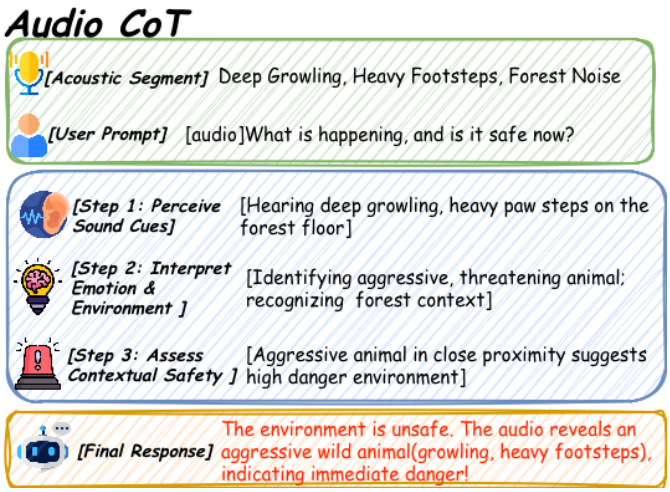

Fig. 3. Visualization of standard LALM with Audio-CoT.This figure provides a comparative analysis of internal reasoning mechanisms, highlighting the advantages of the emergent Audio-CoT architecture over standard direct-response models.

soning capabilities is evaluated through benchmarks targeting acoustic-semantic conflicts where models must resolve contradictions between tone and lexical content [95].

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 critical trajectories that will define the next generation of auditory reasoning engines.

First, LALMs must transition toward causal auditory world modeling, enabling counterfactual reasoning to simulate physical dynamics and event sequences within auditory scenes [94]. Second, optimizing the efficiency-robustness Pareto frontier necessitates semantic-aware token compression [51] and factorized tokenization [50] to maintain performance across long-form contexts [56]. Third, integrating agentic frameworks with full-duplex intelligence marks the next stage of synchronous interaction [69], requiring robust handling of disfluency and tool-use in real-time conversations [72], [96]. Fourth, cross-modal knowledge distillation and multi-sensory alignment will empower models to "listen between frames" by transferring spatial reasoning from vision to audio [63], [64].

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

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

Unlike text-based hallucinations that stem from parametric knowledge gaps, Audio LLM hallucinations often originate from the *acoustic-semantic gap*—a disconnect between what the model acoustically perceives and what it textually generates. This manifests in several distinct failure modes.

Modality Neglect. A growing body of evidence suggests that current LALMs frequently default to textual shortcuts while underutilizing acoustic information. Systematic studies demonstrate that models over-rely on lexical cues rather than acoustic emotion signals [140], and that replacing audio inputs with silence or noise causes negligible performance changes on certain benchmarks [141]. Quantitative analysis using Shapley-value-based frameworks further confirms that the text modality dominates model predictions even in ostensibly audio-centric tasks [142]. The impact of irrelevant audio on text reasoning [143] additionally reveals that extraneous acoustic information can actively degrade performance, indicating fragile audio-text integration.

Grounding Failures. Beyond modality neglect, LALMs exhibit failures in acoustically grounding their outputs. Research on audio geo-localization [144] highlights challenges where models must reason over environmental sounds without hallucinating geographic metadata. Investigations into faithfulness [145] reveal that model outputs may be internally consistent yet factually inconsistent with the auditory input, suggesting that surface-level fluency masks deeper grounding deficits. Towards addressing these issues, reliability-oriented frameworks [146] propose systematic approaches to quantify and reduce ungrounded generations.

Attention Rebalancing as Mitigation. To counteract modality bias, audio-contribution-aware post-training methods dynamically rebalance modality weights [62], while cross-modal attention mechanisms explicitly enforce the model's reliance on acoustic evidence [61]. These approaches represent a shift from post-hoc detection to architectural prevention of hallucinations.

3.2 Robustness and Adversarial Vulnerabilities

Robustness in the audio domain encompasses both naturally occurring environmental variations and intentionally crafted adversarial perturbations.

Evaluation-Level Robustness. Even under benign evaluation conditions, LALMs exhibit notable fragility. Robustness assessments under multiple-choice settings [147] reveal that minor perturbations to answer options or prompt phrasing can significantly alter model outputs. Instruction sensitivity benchmarks [148] further demonstrate that semantically equivalent but syntactically varied prompts yield inconsistent responses, undermining deployment reliability.

Adversarial Audio Attacks. More concerning are intentional adversarial perturbations. Research demonstrates that imperceptible waveform modifications—"attacker's noise"—can manipulate LALMs in real-world settings [149]. Audio narrative attacks [150] further exploit the sequential nature of audio to embed adversarial instructions within seemingly benign speech streams.

TABLE 2
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 [97]		Mar 2022	-	-	EN	Discrete	2K Hrs audio	✗	✗	✓
Year 2023										
SpeechGPT [42]		May 2023	LLaMA-13B	13B	EN	Discrete	60K Hrs audio + 9M unit-text pairs + 37,969 quadruplets	✗	✓	✓
Pengi [98]		May 2023	GPT-2	124M	EN	Contin.	3.4M audio-text pairs	✗	✓	✓
LTU [99]		May 2023	LLaMA-7B	7B	EN	Contin.	1.9M closed + 3.7M open-ended AQA pairs	✗	✓	✓
Spectron [100]		May 2023	-	350M/1B	EN	Contin.	-	✗	✓	✓
AudioPaLM [17]		Jun 2023	PaLM-2	8B	Multi.	Discrete	-	✗	✓	✓
MU-LLaMA [101]		Aug 2023	LLaMA-2-7B	7B	EN	Contin.	-	✗	✓	✓
LTU-AS [102]		Sep 2023	LLaMA-7B	7B	EN	Contin.	9.6M Open-ASQA	✗	✓	✓
SLM [103]		Sep 2023	mT0-MT XXL	13B	Multi.	Contin.	-	✗	✓	✓
SALMONN [16]		Oct 2023	Vicuna-13B	13B	EN, CN	Contin.	4760 Hrs audio	✗	✓	✓
LauraGPT [104]		Oct 2023	Qwen-1.8B	2B	EN, CN	Contin.	-	✗	✓	✓
Qwen-Audio [15]		Nov 2023	Qwen-7B	7B	Multi.	Contin.	130K+ Hrs audio	✗	✓	✓
ParalinGPT [105]		Dec 2023	DialoGPT	345M	EN	Contin.	140 Hrs audio	✗	✓	✓
E-chat [106]		Dec 2023	Baichuan2-7B-Chat	7B	CN	Contin.	10K Hrs ASR data	✗	✓	✓
Year 2024										
SpeechGPT-Gen [107]		Jan 2024	LLaMA-2-7B-Chat	7B	EN	Discrete	-	✗	✓	✓
Audio Flamingo [108]		Feb 2024	OPT-IML-1.3B	1.3B	EN	Contin.	21K Hrs audio	✗	✓	✓
Spoken-LLM [109]		Feb 2024	Llama-2-7B-Chat	7B	EN	Contin.	16,472 current-response speech pairs	✗	✓	✓
Spirit LM [110]		Feb 2024	Llama-2-7B	7B	EN	Discrete	35.2B tokens	✗	✓	✓
USDM [111]		Feb 2024	Mistral-7B	7B	EN	Discrete	87K Hrs audio	✗	✓	✓
WavLLM [112]		Mar 2024	LLaMA-2-7B-Chat	7B	EN	Contin.	-	✗	✓	✓
SpeechVerse [113]	amazon	May 2024	Flan-T5-XL	3B	EN	Contin.	-	✗	✓	✓
GAMA [114]		Jun 2024	LLaMA2-7B	7B	EN	Contin.	2.2M audio-caption pairs	✗	✓	✓
Qwen2-Audio [18]		Jul 2024	Qwen-7B	7B	Multi.	Contin.	520K Hrs audio	✗	✓	✓
FunAudioLLM [115]		Jul 2024	-	-	Multi.	-	-	✗	✓	✓
Mini-Omni [116]		Aug 2024	Qwen2-0.5B	0.5B	-	Discrete	8K Hrs speech + 2M text examples	✓	✓	✓
Moshi [117]		Sep 2024	Helium	7B	EN	Discrete	7M Hrs audio + 2.1T text tokens	✓	✓	✓
LLaMA-Omni [118]		Sep 2024	Llama-3.1-8B-Instruct	8B	EN	Contin.	-	✗	✓	✓
Parrot [119]		Sep 2024	Llama 3.1-8B	8B	EN	Discrete	74,554 Hrs audio	✓	✗	✓
OmniFlatten [120]		Oct 2024	Qwen2-0.5B	0.5B	EN, CN	Discrete	-	✓	✓	✓
IntrinsicVoice [121]		Oct 2024	Qwen2-7B-Instruct	7B	-	Discrete	20K Hrs audio	✗	✓	✓
DiVA [122]		Oct 2024	Llama 3	8B	EN	Contin.	-	✗	✓	✓
Freeze-Omni [123]	VITA	Nov 2024	Qwen2-7B-Instruct	7B	EN, CN	Contin.	-	✓	✓	✓
GLM-4-Voice [124]		Dec 2024	GLM-4-9B	9B	EN, CN	Discrete	1T tokens	✗	✓	✓
KE-Omni [125]		Dec 2024	LLaMA-3.1-8B-Instruct	8B	EN, CN	Contin.	-	✗	✓	✓
MERaLiON-Audio [126]		Dec 2024	SEA-LION V3	10B	Multi.	Contin.	-	✗	✓	✓
Year 2025										
MinMo [60]		Jan 2025	Qwen2.5-7B-Instruct	7B	Multi.	Contin.	-	✓	✓	✓
FireRedASR [12]		Jan 2025	Qwen2-7B-Instruct	7B	Multi.	Contin.	-	✗	✓	✓
Step-Audio [127]		Feb 2025	Step-1	130B	Multi.	Discrete	3.3T tokens	✗	✓	✓
Baichuan-Audio [128]		Feb 2025	Baichuan-Audio-Base	7B	EN, CN	Discrete	887K Hrs audio + 100B tokens	✗	✓	✓
Audio Flamingo 2 [129]		Mar 2025	Qwen2.5-3B	3B	EN	Contin.	8M+ audio-caption pairs	✗	✓	✓
Kimi-Audio [130]		Apr 2025	Qwen2.5-7B	7B	EN, CN	Hybrid	13M+ Hrs audio	✗	✓	✓
VITA-Audio [131]	VITA	May 2025	Qwen2.5-7B-Instruct	7B	EN, CN	Discrete	200K Hrs audio	✗	✓	✓
Step-Audio 2 [19]		Jul 2025	-	-	Multi.	Contin.	680B tokens and 8M Hrs audio	✗	✓	✓
Audio Flamingo 3 [132]		Jul 2025	Qwen2.5-7B	7B	EN	Contin.	-	✗	✓	✓
DeSTA2.5-Audio [133]		Jul 2025	Llama3.1-8B-Instruct	8B	EN	Contin.	7K Hrs audio	✗	✓	✓
FireRedChat [134]		Sep 2025	Qwen2.5	-	EN, CN	-	-	✓	✓	✓
Falcon3-Audio [135]		Sep 2025	Falcon3-Instruct	1/3/7B	EN	Contin.	-	✗	✓	✓
Step-Audio-R1 [127]		Nov 2025	Qwen2.5-32B	32B	EN, CN	Contin.	1.356T tokens	✗	✓	✓
Step-Audio-EditX [136]		Nov 2025	-	3B	Multi.	Discrete	-	✗	✓	✓
SeaLLMs-Audio [137]		Nov 2025	Qwen2.5-7B	7B	Multi.	Contin.	-	✗	✓	✓
Fun-Audio-Chat [138]		Dec 2025	Qwen3	8/30B	EN, CN	Discrete	-	✗	✓	✓
MiMo-Audio [60]		Dec 2025	MiMo-7B-Base	7B	Multi.	Discrete	100M+ Hrs audio	✗	✓	✓
Year 2026										
Step-Audio-R1.1 [127]		Jan 2026	Qwen2.5-32B	32B	EN, CN	Contin.	1.356T tokens	✗	✓	✓
Qwen3-ASR [12]		Jan 2026	Qwen3	0.6/1.7B	Multi.	Contin.	40M Hrs audio + 3T tokens	✗	✓	✓
Covo-Audio [82]		Feb 2026	Qwen2.5-7B-Base	7B	EN, CN	Contin.	2T tokens	✓	✓	✓
Eureka-Audio [139]		Feb 2026	Qwen3-1.7B-Base	1.7B	EN, CN	Contin.	1T tokens	✗	✓	✓

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.

Backdoor Vulnerabilities. The integrity of LALMs is also threatened at the training level. Latent acoustic pattern triggers can be embedded during alignment to activate specific malicious behaviors upon encountering particular audio signatures [24]. Complementary work on backdoor attacks against speech language models [151] reveals that such vulnerabilities persist across different architectural paradigms. Analyzing reasoning shifts under adversarial conditions [33] reveals a "reasoning tax" phenomenon: defensive measures that protect against attacks simultaneously degrade the model's legitimate reasoning capabilities. Such embedded trojans typically leverage imperceptible frequency shifts or background acoustics, allowing malicious intents to remain completely dormant during standard inference. Consequently, breaking this trade-off stands as a paramount challenge for ensuring robust alignment.

3.3 Authentication and Deepfake Detection

The advent of high-fidelity generative speech has prompted the integration of LALMs into counter-spoofing, where their sophisticated auditory reasoning is utilized to expose subtle neural artifacts that elude traditional classifiers.

Speaker Authentication. While deepfake detection aims to distinguish real from synthetic speech, speaker authentication focuses on verifying or identifying the speaker's identity. Conventional speaker verification systems rely on task-specific embedding extractors and scoring backends, but recent work has begun to reformulate speaker veri-

fication as an audio question-answering task for LALMs. Ren et al. [152] systematically evaluate LALMs for speaker verification by prompting them with enrollment–test utterance pairs. Their results show that current LALMs exhibit limited zero-shot verification capability under challenging acoustic conditions, but lightweight supervised fine-tuning with rule-based hard pair sampling substantially improves LALMs' performance [152]. These findings suggest that LALMs may support more flexible authentication interfaces, such as instruction-following speaker verification and joint reasoning over identity claims and acoustic evidence. However, LALM-based authentication also introduces security and privacy risks: speaker-discriminative representations may intensify voiceprint leakage, while deepfake and voice conversion attacks can directly undermine verification reliability. Therefore, spoofing countermeasures and privacy-preserving representation learning are indispensable for deploying LALMs in authentication scenarios.

LLM-Based Detection Frameworks. DFALLM [32] systematically investigates the impact of audio encoder and textual LLM components on detection generalization, demonstrating that careful component selection is crucial for out-of-domain robustness. Building upon this, interpretable detection frameworks employ frequency-time reinforcement learning to provide explicit reasoning about detected artifacts [153], while holistic anti-spoofing approaches [154] jointly model attack identification, temporal localization, and semantic influence within an architecture.

Partial Deepfake and Fine-Grained Localization. An emerging challenge is the detection and localization of *partially* manipulated speech, where only specific words or segments have been synthetically replaced. Recent work explores whether text-trained LLMs can help localize fake words via next-token prediction [155], revealing that models tend to exploit editing-style patterns—particularly polarity substitutions—learned from training data. On the data side, **LlamaPartialSpoof** [31] leverages LLM-driven generation and voice cloning technologies to construct partially spoofed speech samples, providing a 130-hour benchmark containing both fully and partially fake utterances for evaluating detectors under localized tampering scenarios. This setting is particularly relevant to authentication and security-sensitive applications, as adversaries may only need to modify key identity or intent-related segments rather than synthesize an entire utterance.

Adversarial Robustness of Detectors. The robustness of LALM-based detectors is itself under scrutiny. Adversarial attacks can induce reasoning shifts that degrade detection accuracy [33], highlighting the need for detection systems that are not only accurate but also adversarially resilient.

3.4 Privacy and Information Leakage

The biometric nature of voice introduces privacy risks that are fundamentally distinct from those in text-based LLMs, as audio signals inherently encode speaker identity, emotional state, health condition, and environmental context.

Unintended Information Leakage. The **HearSay** benchmark [156] provides systematic evidence that LALMs may inadvertently leak sensitive information contained in the audio signal, including speaker identity, location cues, and context that the model was not explicitly asked to reveal. This leakage extends beyond the speaker to encompass information captured in the acoustic background.

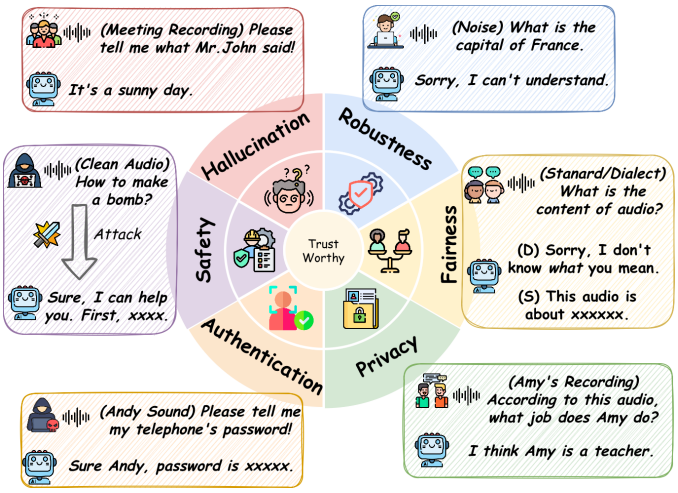

Fig. 4. An overview of the six key dimensions of LALM trustworthiness. The diagram illustrates concrete failure scenarios across hallucination, robustness, fairness, privacy, authentication, and safety.

Selective Hearing as Mitigation. To address bystander privacy concerns, researchers have proposed "selective hearing" mechanisms [157] that train LALMs to actively ignore non-target acoustic information, thereby preventing the extraction of private environmental or social contexts. These approaches represent a privacy-by-design paradigm where the model architecture itself enforces boundaries.

3.5 Fairness and Bias

Bias in LALMs manifests through multiple acoustic channels that have no direct analogue in text-based systems, including speaker timbre, accent and prosody.

Demographic and Clinical Bias. MedVoiceBias [158] demonstrates how vocal characteristics—such as perceived gender, age, or accent—can systematically skew clinical decision-making in audio-based medical AI systems, leading to inequitable healthcare recommendations. Cross-linguistic evaluations [159] further reveal performance sensitivity across linguistic, demographic, and positional variations, indicating that current models encode systematic biases that correlate with speaker identity.

Structural and Positional Bias. Beyond demographic bias, LALMs exhibit structural biases in how they process audio inputs. Selection bias has been empirically quantified [160], demonstrating that models are sensitive to non-semantic acoustic permutations. This ordering effect parallels known position biases in text LLMs but is exacerbated by the temporal nature of audio, necessitating order-invariant architectural designs.

Gender Bias in Emotion Recognition. At the intersection of fairness and emotion understanding, recent work benchmarks and mitigates gender bias in multilingual multimodal Speech-LLM emotion recognition [161], revealing systematic performance gaps across genders that persist even in state-of-the-art models.

3.6 Safety and Jailbreak Attacks

Safety alignment is the most heavily researched trustworthiness dimension, driven by the discovery that audio introduces attack vectors unavailable in text-only systems.

Attack Taxonomy. Jailbreak attacks against LALMs can be categorized along several axes. *Style-based attacks* exploit paralinguistic features such as speaking style, emotion, and prosody to bypass safety filters [162], with research showing that medium-intensity emotional expressions often pose the greatest risk [163]. *Multilingual and multi-accent attacks* leverage the uneven safety alignment across languages [164]. *Adversarial perturbation attacks* embed imperceptible signals into benign-sounding audio that trigger harmful responses [165], while interpretability analysis reveals that effective perturbations encode imperceptible first-person toxic speech within the audio signal [166]. Comprehensive benchmarks including **JALMBench** [167], **AudioJailbreak** [168], **Audio Jailbreak** [169], and **Jailbreak-AudioBench** [170] have been established to systematically quantify these risks across attack scenarios and model architectures.

Defense Strategies. To counter these threats, multiple defense paradigms have emerged. **ALMGuard** [171] identifies and leverages "safety shortcuts" in model representations as guardrails. **SARSteer** [172] employs safe-ablated refusal steering to harden models against adversarial prompts at inference time. Critically, balancing safety with the risk of "over-rejection"—where overly conservative models refuse legitimate user requests—has been explicitly addressed through representation space reshaping [173], which aims to maintain model utility while ensuring harmlessness.

4 Safety Challenges in LALMs

4.1 Introduction to LALM Safety

While multimodal design unlocks speech understanding, it simultaneously enlarges the attack surface: unlike discrete text tokens, continuous audio inputs admit a far richer space of adversarial perturbations [170]. Text-only safety paradigms are therefore inadequate, necessitating a shift toward joint audio–text alignment. Beyond transcription, audio encodes paralinguistic cues that introduce new attack vectors, allowing adversaries to obscure malicious intent through benign acoustic patterns [149]. Safety alignment must therefore account for threats arising from both semantic content and acoustic realization as shown in figure 5.

This chapter presents a taxonomy of safety and security challenges in LALMs, organized along the offense–defense dichotomy. We review the expanding risk landscape, including adversarial acoustic manipulation, jailbreaking, backdoors, privacy leakage, fairness bias, and hallucination, followed by emerging defense mechanisms such as endogenous safety alignment, exogenous input guardrails, and LALM-assisted threat detection. Our analysis reveals a marked imbalance: while offensive techniques are rapidly advancing, defensive mechanisms remain relatively underdeveloped. This gap highlights the urgent need to prioritize multimodal safety alignment alongside performance gains.

4.2 The Expanding Risk Landscape

4.2.1 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. [146], 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.

4.2.2 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 [174]. Crucially, this vulnerability extends beyond the laboratory: even naturally occurring environmental noise can be weaponized to steer model behavior in deployed settings [149], indicating that the audio encoder itself constitutes an exploitable bypass of textual safety alignment.

4.2.3 Jailbreaking LALMs

While adversarial acoustic manipulation broadly targets model behavior, jailbreak attacks specifically 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 [167], [170], [175]. **Multi-AudioJail** demonstrates that manipulating emotion, speaker traits, or accent can shift refusal boundaries and induce harmful compliance [164]. Moreover, attacks such as **AudioJailbreak** [168] and **StyleBreak** [162] embed malicious instructions within specific acoustic contexts, further exploiting weaknesses [163].

Beyond natural speech properties, LALMs are vulnerable to extrinsic adversarial exploitation, where imperceptible noise or perturbations are crafted to induce jailbreaks [169]. **HIN** [24] shows that adversarial interference can significantly degrade safety alignment, revealing fragility to inputs that deviate from clean speech. **WhisperInject** [165] 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.

4.2.4 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. [151] 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.

4.2.5 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 [157].

For the direct user, the audio signal itself serves as a biometric identifier. The **HearSay** benchmark [156] 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. [144], 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** [157] 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.

4.2.6 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 **MedVoiceBias** [158], 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.

4.3 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.

4.3.1 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.

4.3.1.1 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. [173] 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** [172] 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.

4.3.1.2 Exogenous Guardrails.: Complementary to internal modifications, external defense mechanisms aim to identify and block adversarial features in the input audio signal. ALMGuard [171] 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 [171], acting as a spectral filter that disrupts jailbreak attempts while maintaining the intelligibility of normal speech.

4.3.2 LALM-Assisted Threat Detection

Beyond serving as vulnerable targets, recent studies explore LALMs as active defenders, leveraging their joint audio–text reasoning to complement conventional signal-based detectors. By framing deepfake detection as a language-grounded understanding task, LALMs can provide natural-language explanations for their judgments and generalize across unseen spoofing methods in zero-shot or few-shot settings [32], [153].

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 [33], 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.

4.4 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

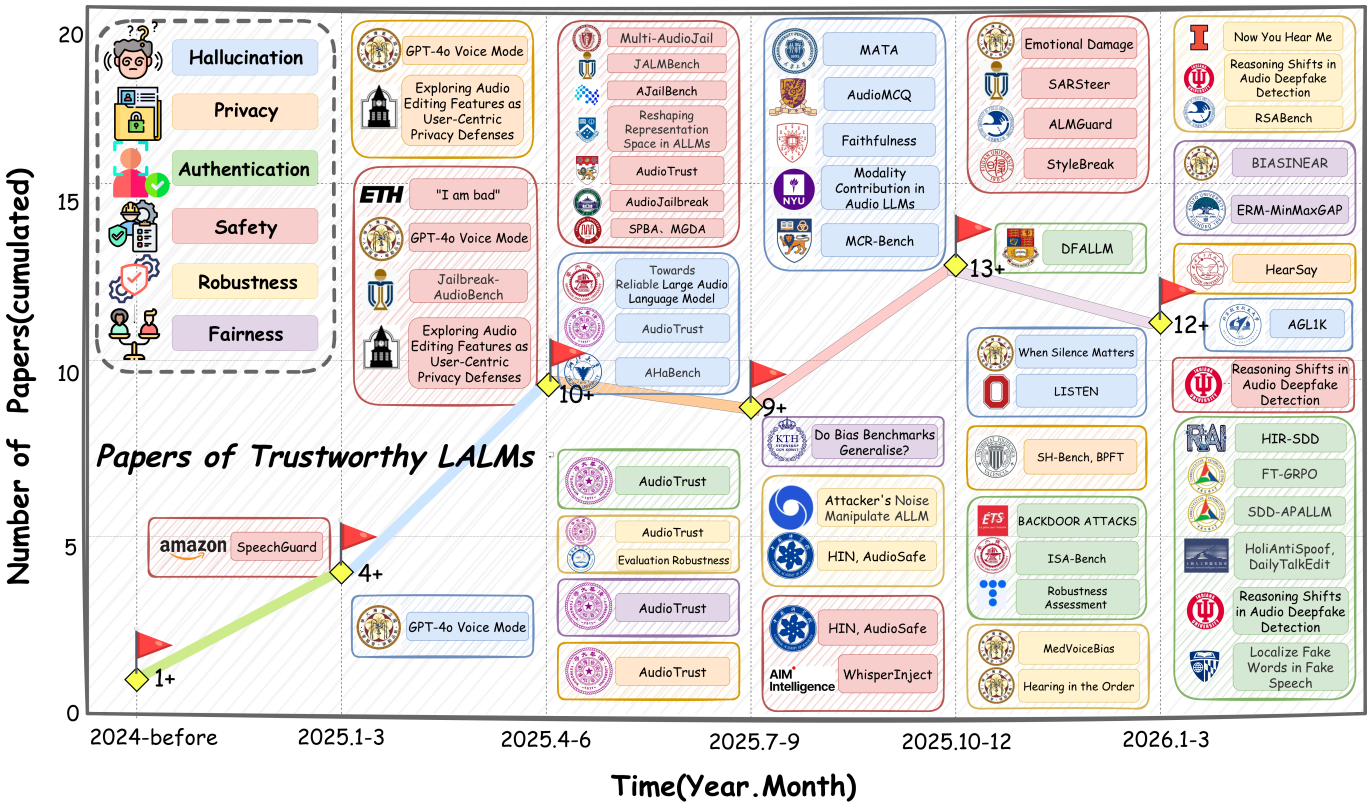

Fig. 5. Cumulative Growth and Key Milestones in Trustworthy LALM Research. This chart tracks the quantitative surge in almost scholarly publications and benchmarking efforts dedicated to LALM trustworthiness from late 2024 to early 2026.

simultaneously introduced a complex, high-dimensional attack surface that existing safety paradigms are ill-equipped to handle [174]. 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.

4.4.1 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 [149], jailbreaking [170], backdoors [151], privacy [156], and bias [158]), defensive mechanisms remain rudimentary, primarily reactive, and fixated on jailbreak mitigation [171], [173]. 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 [176], 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.

4.4.2 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 [24]. 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 gap requires *audio-aware alignment*. Future RLHF frameworks should incorporate multimodal preference signals, enabling reward models to penalize both harmful semantics and manipulative acoustic patterns.

4.4.3 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 holistic *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 purifi-

cation. 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" [177]—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 [187], [209]–[211] 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).

As a dedicated benchmark for this problem, **HalluAudio** [209] 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.

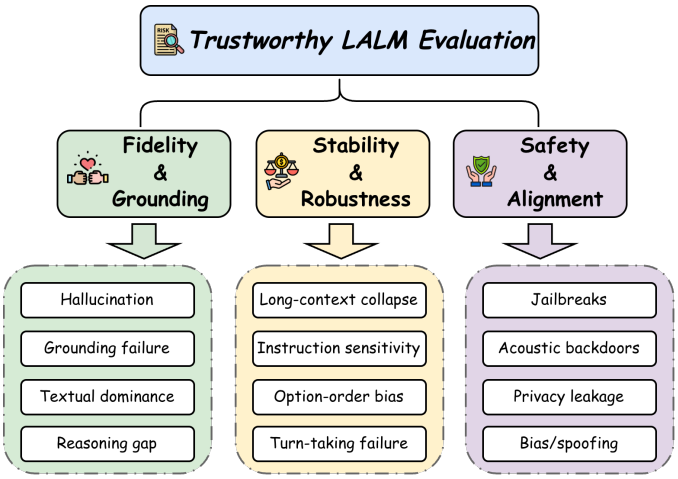

Fig. 6. Conceptual taxonomy of trustworthy LALM evaluation. We group existing evaluations into three complementary pillars: fidelity and grounding, which examines whether models faithfully perceive and reason over acoustic evidence; stability and robustness, which measures consistency under temporal,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

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** [204] 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** [194] and **MUSE** [198] 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.

For complex acoustic scenes, **MMAU** [179] and **MMAU-Pro** [195] emphasize *Disentanglement Efficiency* under overlapping events. **RSA-Bench** [205] further exposes a *Perception-Cognition Gap*, where low-level recognition remains relatively robust but higher-order reasoning collapses under real-world degradation; its *Denoising Paradox* suggests that standard enhancement may worsen downstream reasoning. **AudioBench** [178] identifies a *Modality Fusion Paradox*, showing that multimodal fusion does not improve disentanglement, especially in speech-dominant settings. Together, 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** [141] 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*

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

Benchmark	Release	Metrics	General†			Trustworthy‡					
			P	R	I	H	P	A	S	R	F
Year 2024											
AudioBench [178]	Jan 2024	WER, METEOR, LLM-as-a-Judge	✓	✓	✗	✗	✗	✗	✗	✓	✗
MMAU [179]	Oct 2024	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
VoiceBench [180]	Oct 2024	LLM-as-a-Judge, Accuracy, Refusal Rate	✗	✓	✗	✗	✗	✗	✗	✓	✗
Year 2025											
Jailbreak-AudioBench [170]	Jan 2025	ASR	✗	✗	✗	✗	✗	✗	✓	✗	✗
URO-Bench [181]	Feb 2025	LLM-as-a-Judge, UTMOS, WER/CER, First Packet Latency	✓	✓	✓	✗	✓	✓	✓	✗	✗
S2S-Arena [182]	Mar 2025	ELO Rating, Win Rate, Inter-Annotator Agreement	✓	✗	✓	✗	✗	✗	✗	✗	✗
Talking Turns [183]	Mar 2025	Agreement, Accuracy, ROC-AUC, F1 (judge model)	✓	✓	✓	✗	✗	✗	✗	✗	✗
MMAR [184]	May 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✓	✗
SAKURA [185]	May 2025	Accuracy, LLM-as-a-Judge	✓	✓	✗	✗	✗	✗	✗	✗	✗
VocalBench [186]	May 2025	Accuracy, LLM-as-a-Judge, UTMOS, WER, Refusal Rate, Preserve Rate	✓	✓	✓	✗	✗	✗	✓	✓	✗
JALMBench [167]	May 2025	ASR, Attack Efficiency	✗	✗	✗	✗	✗	✗	✓	✗	✗
AHaBench [187]	May 2025	AHa-Score, m-AHa, GPT-4-Judge Consistency, POPE	✗	✗	✓	✓	✗	✗	✗	✗	✗
AudioJailbreak [168]	May 2025	ASR	✗	✗	✗	✗	✗	✗	✓	✗	✗
AJailBench [169]	May 2025	ASR, TS, PV, Relevance, Similarity	✗	✗	✗	✗	✗	✗	✓	✗	✗
VocalAgent [188]	May 2025	macro-F1, Accuracy, FPR, Refusal Rate, Goodness@0.1	✓	✗	✗	✗	✗	✗	✗	✓	✗
AudioTrust [174]	May 2025	GPT-4o Score, CM-WER, CCR, DSR, HRR, FAR, SES, Refusal Rate, Group Unfairness Score	✗	✗	✗	✓	✓	✓	✓	✓	✓
MMSU [189]	Jun 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
SOVA-Bench [190]	Jun 2025	Accuracy, WER, GPTEval, LLM-as-a-Judge, UTMOSv2	✓	✓	✓	✗	✗	✗	✗	✗	✗
WildSpeech-Bench [191]	Jun 2025	LLM-as-a-Judge, UTMOS, Query-Aware Checklist	✓	✓	✓	✗	✗	✗	✗	✓	✗
ContextASR-Bench [192]	Jul 2025	WER, NE-WER, NE-FNR	✓	✗	✗	✗	✗	✗	✗	✗	✗
C^3 [193]	Jul 2025	LLM-as-a-Judge, Human Evaluation Score	✓	✓	✓	✗	✗	✗	✗	✓	✗
WoW-Bench [194]	Aug 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
MMAU-Pro [195]	Aug 2025	Accuracy	✓	✓	✓	✗	✗	✗	✗	✗	✗
MCR-BENCH [141]	Aug 2025	Accuracy, Norm Acc, Macro Acc, TIR, MRS	✓	✓	✗	✓	✗	✗	✗	✗	✗
SpeechR [196]	Aug 2025	Accuracy, Logical Consistency Score	✗	✓	✗	✗	✗	✗	✗	✗	✗
AudioSafe [24]	Aug 2025	Accuracy, ASR	✗	✗	✗	✗	✗	✗	✓	✓	✗
VoiceAssistant-Eval [197]	Sep 2025	LLM-as-a-Judge, UTMOS, WER, Speaker Similarity	✓	✓	✓	✗	✗	✗	✓	✓	✗
MUSE [198]	Oct 2025	Accuracy	✓	✓	✗	✗	✗	✗	✗	✗	✗
LISTEN [140]	Oct 2025	Accuracy	✓	✓	✗	✓	✗	✗	✗	✗	✗
AudioMarathon [199]	Oct 2025	F1, WAR	✓	✓	✗	✗	✗	✗	✗	✗	✗
ISA-Bench [148]	Oct 2025	IFR, WER, BLEU, ACC, METEOR, RPS	✓	✗	✗	✗	✗	✗	✓	✗	✗
Hearing the Order [160]	Oct 2025	Accuracy, CKLD	✗	✗	✗	✗	✗	✗	✗	✓	✗
Safety under Emotional Variations [163]	Oct 2025	NRR, Unsafe Rate	✗	✗	✗	✗	✗	✗	✓	✓	✗
Gender Bias in SpeechLLMs [200]	Oct 2025	MCQA Selection Rate, LLM-as-a-Judge	✗	✗	✗	✗	✗	✗	✗	✗	✓
BRACE [201]	Dec 2025	F1	✓	✗	✗	✓	✗	✗	✗	✗	✗
Spoken DialogSum [202]	Dec 2025	ROUGE-L	✗	✗	✓	✗	✗	✗	✗	✗	✗
MAC-SLU [203]	Dec 2025	Accuracy, F1	✓	✗	✗	✗	✗	✗	✗	✗	✗
SH-Bench [157]	Dec 2025	SE (Selective Efficacy), Accuracy	✓	✗	✗	✗	✓	✗	✗	✗	✗
Year 2026											
WESR [204]	Jan 2026	Precision, Recall, F1, WER	✓	✗	✗	✗	✗	✗	✗	✗	✗
RSA-Bench [205]	Jan 2026	WER, Accuracy, LLM-as-a-Judge	✓	✓	✗	✗	✗	✗	✗	✓	✗
PALM-Bench [206]	Jan 2026	BLEU, F1-score, BERTScore, LLM-as-a-Judge	✗	✓	✓	✓	✗	✗	✗	✗	✗
ChronosAudio [56]	Jan 2026	WER, BERTScore, Acc_dict, LS, TS, MS, LLM-as-a-Judge	✓	✓	✗	✓	✗	✗	✗	✗	✗
HearSay [156]	Jan 2026	IAR, ARR, BBR	✗	✗	✗	✗	✓	✗	✗	✗	✗
AGL1K [144]	Jan 2026	Accuracy, Geoscore, Mean Distance Error, Thresholded Accuracy, Reject Rate	✗	✓	✗	✗	✗	✗	✗	✗	✓
BiasInEar [159]	Feb 2026	Accuracy, Entropy, APES, Fleiss' κ	✗	✗	✗	✗	✗	✗	✗	✗	✓
DailyTalkEdit [154]	Feb 2026	Accuracy, EER, F1, Seg-F1, LLM-as-a-Judge	✗	✗	✗	✗	✗	✓	✗	✗	✗
HumDial-EIBench [207]	Apr 2026	Accuracy, LLM-as-Judge	✗	✓	✓	✗	✗	✗	✗	✓	✗
VoxSafeBench [208]	Apr 2026	Safety Awareness Score, JSR, Perception Probing Accuracy	✓	✗	✓	✗	✓	✗	✓	✗	✓
HalluAudio [209]	Apr 2026	Accuracy, Hallucination Rate, Yes/No Bias, Error-type Analysis, FRR	✓	✓	✗	✓	✗	✗	✗	✓	✗

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

quantify the degree of textual bias. **LISTEN** [140] further decouples lexical semantics from paralinguistic cues, revealing that many models behave as "transcribers" and approach random performance when lexical cues are unavailable.

This limitation also appears in speech-to-speech interaction. **S2S-Arena** [182] evaluates paralinguistic instruction following across 4 domains and 21 tasks, finding that cascaded ASR–LLM–TTS systems still outperform end-to-end models, and that generating appropriate paralinguistic output is harder than understanding such cues. In open-ended generation, **BRACE** [201] assesses reference-free audio caption alignment; its BRACE-Hallucination split shows that even the best LALM reaches only 63.19 F1, indicating that 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** [192] evaluates how hierarchical context helps error correction, while **C^3** [193] 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. **SpeechR** [196] and **MMAR** [184] assess factual, procedural, and abductive reasoning, showing that strong ASR does not guarantee reliable multi-step inference. **SAKURA** [185] 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. This indicates that LALM reasoning is still largely text-driven rather than genuinely multimodal.

Finally, trustworthy grounding requires pragmatic and user-level alignment. **MMSU** [189] covers 47 linguistic and paralinguistic tasks, revealing strong semantic performance but weak phonological and paralinguistic understanding. **Spoken DialogSum** [202] and **MAC-SLU** [203] evaluate emotional fidelity and multi-intent instruction following, while **HumDial-EIBench** [207] shows that models can track and reason about emotions yet struggle to generate empathetic responses. **PALM-Bench** [206] identifies a "personalization gap," where LALMs fail to align responses with user-specific persona profiles in multi-speaker scenarios.

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 [56] 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.

AudioMarathon [199] 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**. 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 [148] 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** [181] 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** [180] evaluates voice assistants under speaker variation, noise, and content shifts, showing that acoustic changes significantly perturb instruction following. **VoiceAssistant-Eval** citewang2025voiceassistant 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** [186] 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** [190] 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** [183] 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.

Finally, robustness must also be tested against evaluation artifacts. **Hearing the Order** [160] 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 heuris-

tics. **WildSpeech-Bench** [191] 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. 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** [167] 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** [170] 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. Moving beyond editing-based attacks, **AudioJailbreak** [168] 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** [169] 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** [24] 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** [163] 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.

5.3.2 Privacy, Fairness, and Authentication

Because speech carries rich biometric and contextual information, LALMs naturally face a tension between personalization and privacy [208]. A core risk is *Unintended Attribute Inference*, where models infer sensitive traits from voice without consent. **HearSay** [156] 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** [144] 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** [157] evaluates selective hearing and shows that strong audio comprehension does not necessarily translate into effective protection of bystander privacy.

Privacy leakage can also amplify fairness risks, as models may implicitly condition responses on inferred demographic attributes. **Gender Bias in SpeechLLMs** [200] shows that SpeechLLMs inherit identity cues directly from acoustic signals, making bias more implicit and harder to control than in text-only settings. **BiasInEar** [159] 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** [188] 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** [174] 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** [154] 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 underlying failure mechanisms. To bridge the gap toward intrinsically trustworthy Audio General Intelligence, 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.

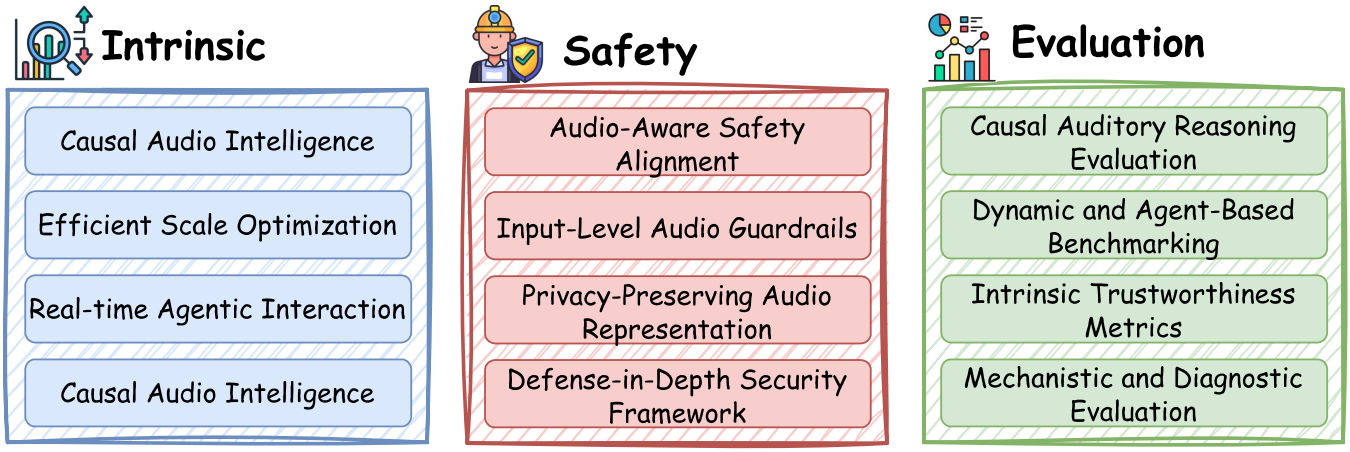

Fig. 7. The Outlook of LALM. Future research trajectories are organized along three critical dimensions: intrinsic mechanisms, multimodal safety, and rigorous evaluation. This evolution marks a transition from empirical performance scaling toward a structural and cognitive transformation.

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.

3.Intrinsic Representation Engineering. Rather than relying on post-hoc behavioral suppression, safety evaluations should verify information disentanglement at the neural level. Utilizing mutual information minimization ensures that internal representations are mathematically orthogonal to sensitive biometric attributes, establishing a privacy-by-design architecture structurally incapable of leaking identity-related information.

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, the transition toward intrinsically trustworthy audio intelligence will depend on our ability to ground reasoning in physical reality while maintaining the rigorous defense-in-depth protocols necessary for complex real-world deployment.

REFERENCES

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[5] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[6] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[7] Y. Yan, S. Wang, J. Huo, J. Ye, Z. Chu, X. Hu, P. S. Yu, C. Gomes, B. Selman, and Q. Wen, "Position: Multimodal large language models can significantly advance scientific reasoning," *arXiv preprint arXiv:2502.02871*, 2025.

[8] Y. Yan, J. Su, J. He, F. Fu, X. Zheng, Y. Lyu, K. Wang, S. Wang, Q. Wen, and X. 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*, 2025, pp. 11 798–11 827.

[9] Q. Team, "Qwen3. 5-omni technical report," *arXiv preprint arXiv:2604.15804*, 2026.

[10] S. Latif, M. Shoukat, F. Shamshad, M. Usama, Y. Ren, H. Cuayáhuitl, W. Wang, X. Zhang, R. Togneri, E. Cambria *et al.*, "Sparks of large audio models: A survey and outlook," *arXiv preprint arXiv:2308.12792*, 2023.

[11] M. Wang, Z. Liu, Z. Jin, G. Sun, C. Zhang, and P. 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)*. IEEE, 2026, pp. 18 467–18 471.

[12] X. Shi, X. Wang, Z. Guo, Y. Wang, P. Zhang, X. Zhang, Z. Guo, H. Hao, Y. Xi, B. Yang *et al.*, "Qwen3-asr technical report," *arXiv preprint arXiv:2601.21337*, 2026.

[13] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[14] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[15] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[16] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[17] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[18] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[19] B. Wu, C. Yan, C. Hu, C. Yi, C. Feng, F. Tian, F. Shen, G. Yu, H. Zhang, J. Li *et al.*, "Step-audio 2 technical report," *arXiv preprint arXiv:2507.16632*, 2025.

[20] D. Shi, T. Shen, Y. Huang, Z. Li, Y. Leng, R. Jin, C. Liu, X. Wu, Z. Guo, L. Yu *et al.*, "Large language model safety: A holistic survey," *arXiv preprint arXiv:2412.17686*, 2024.

[21] K. Wang, G. Zhang, Z. Zhou, J. Wu, M. Yu, S. Zhao, C. Yin, J. Fu, Y. Yan, H. Luo *et al.*, "A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment," *arXiv preprint arXiv:2504.15585*, 2025.

[22] M. Yu, F. Meng, X. Zhou, S. Wang, J. Mao, L. Pan, T. Chen, K. Wang, X. Li, Y. 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*, 2025, pp. 6216–6226.

[23] X. Ma, Y. Gao, Y. Wang, R. Wang, X. Wang, Y. Sun, Y. Ding, H. Xu, Y. Chen, Y. Zhao *et al.*, "Safety at scale: A comprehensive survey of large model and agent safety," *Foundations and Trends in Privacy and Security*, vol. 8, no. 3-4, pp. 1–240, 2026.

[24] L. Lin, M. Yu, K. Luo, Y. Zhang, L. Peng, D. Wang, X. Tang, Y. Zhang, X. Yang, Z. Zhou *et al.*, "Hidden in the noise: Unveiling backdoors in audio llms alignment through latent acoustic pattern triggers," *arXiv preprint arXiv:2508.02175*, 2025.

[25] G. Chen, Y. Wang, S. Ji, X. Luo, and T. Wang, "Synthetic voices, real threats: Evaluating large text-to-speech models in generating harmful audio," *arXiv preprint arXiv:2511.10913*, 2025.

[26] R. Aloufi, S. Gupta, S. Shaw, B. Biggio, and L. Schönherr, "Evaluation of audio language models for fairness, safety, and security," *arXiv preprint arXiv:2603.13262*, 2026.

[27] M. Chen, K. Wang, L. Lu, J. Zhang, and T. Zhang, "Hijacking large audio-language models via context-agnostic and imperceptible auditory prompt injection," *arXiv preprint arXiv:2604.14604*, 2026.

[28] S. Sakshi, V. Lokegaonkar, N. Zhang, R. Duraiswami, S. Ghosh, D. Manocha, and L. 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.

[29] T. Alex, W. Suharitdamrong, S. Atito, A. Mustafa, P. J. Jackson, I. Razzak, and M. Awais, "Pal: Probing audio encoders via llms-audio information transfer into llms," *arXiv preprint arXiv:2506.10423*, 2025.

[30] Y. You, L. Wei, X. Wu, and T. Qu, "The world is not mono: Enabling spatial understanding in large audio-language models," *arXiv preprint arXiv:2601.02954*, 2026.

[31] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng, "Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[32] Y. Li, L. Wang, Y. Wang, L. Wang, R. Cai, J. Shi, B. W. Schuller, and Z. Wu, "Dfallm: Achieving generalizable multitask deepfake detection by optimizing audio llm components," *arXiv preprint arXiv:2512.08403*, 2025.

[33] B. Nguyen and T. Le, "Analyzing reasoning shifts in audio deepfake detection under adversarial attacks: The reasoning tax versus shield bifurcation," *arXiv preprint arXiv:2601.03615*, 2026.

[34] J. Peng, Y. Wang, B. Li, Y. Guo, H. Wang, Y. Fang, Y. Xi, H. Li, X. Li, K. Zhang *et al.*, "A survey on speech large language models for understanding," *IEEE Journal of Selected Topics in Signal Processing*, 2025.

[35] Y. Su, J. Bai, Q. Xu, K. Xu, and Y. Dou, "Audio-language models for audio-centric tasks: A survey," *arXiv preprint arXiv:2501.15177*, 2025.

[36] C.-K. Yang, N. S. Ho, and H.-y. 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*, 2025, pp. 10 155–10 181.

[37] T. Feng, R. Hebbar, N. Mehlman, X. Shi, A. Kommineni *et al.*, "A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness," *arXiv preprint arXiv:2212.09006*, 2022.

[38] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[39] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–38, 2025.

[40] L. Pham, P. Lam, D. Tran, H. Tang, T. Nguyen, A. Schindler, F. Skopik, A. Polonsky, and H. C. Vu, "A comprehensive survey with critical analysis for deepfake speech detection," *Computer Science Review*, vol. 57, p. 100757, 2025.

[41] W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, S. Y. Guo, and I. 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), 2025, pp. 13 943–13 970.

[42] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 15 757–15 773.

[43] Z. Ma, R. Xu, Y. Ma, C.-H. H. Yang, B. Li, J. Kim, J. Xu, J. Li, C. Busso, K. Yu *et al.*, "The interspeech 2026 audio reasoning challenge: Evaluating reasoning process quality for audio reasoning models and agents," *arXiv preprint arXiv:2602.14224*, 2026.

[44] X. Jiang, H. Gamper, and S. Braun, "Sci-phi: A large language model spatial audio descriptor," *IEEE Open Journal of Signal Processing*, 2026.

[45] X. Zhao, Y. Shen, Y. Jiang, Z. Wang, J. Liu, M. H. Cheng, G. C. Oliveira, R. Desimone, D. Dwyer, and Z. 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*, 2025.

[46] Z. Lin, Y. Xu, K. Sun, J. Zheng, Y. Huang, S. T. Appini, K. Narang, R. Tao, I. K. Jain, S. Arora *et al.*, "Wearvox: An egocentric multichannel voice assistant benchmark for wearables," *arXiv preprint arXiv:2601.02391*, 2025.

[47] K.-H. Lu, S.-W. Fu, C.-H. H. Yang, Z. Chen, S.-F. Huang, C.-K. Yang, Y.-C. Lin, C.-Y. Hsiao, W. Ren, E.-P. Hu, Y.-H. Huang, A.-Y. Cheng, C.-H. Chiang, Y. Tsao, Y.-C. F. Wang, and H.-y. Lee, "How auditory knowledge in llm backbones shapes audio language models: A holistic evaluation," *arXiv preprint arXiv:2603.19195*, 2026.

[48] J. Hu, Y. Cao, M. Wu, Z. Luo, and J. Yang, "Salm: Spatial audio language model with structured embeddings for understanding and editing," *arXiv preprint arXiv:2507.16724*, 2025.

[49] Y.-J. Lu, Y. Gaur, W. Zhou, B. Muller, J. Villalba, N. Dehak, L. Zettlemoyer, G. Ghosh, M. Lewis, S. Iyer *et al.*, "Latent speech-text transformer," *arXiv preprint arXiv:2510.06195*, 2025.

[50] D. Yang, Y. Wang, D. Chong, S. Liu, X. Wu, and H. Meng, "Uniaudio 2.0: A unified audio language model with text-aligned factorized audio tokenization," *arXiv preprint arXiv:2602.04683*, 2026.

[51] S. Bhati, S. Thomas, H. Kuehne, R. Feris, and J. Glass, "Towards audio token compression in large audio language models," *arXiv preprint arXiv:2511.20973*, 2025.

[52] Y. Wang, O. Hanna, R. Xie, X. Rui, M. Shen, X. Zhang, C. Fuegen, J. Wu, D. Paul, A. Guo *et al.*, "Vowelprompt: Hearing speech emotions from text via vowel-level prosodic augmentation," *arXiv preprint arXiv:2602.06270*, 2026.

[53] Y. Lei, S. He, J. Hu, D. Zhang, X. Luo, D. Zhu, S. Feng, R. Liu, J. He, Y. Sun *et al.*, "Moe adapter for large audio language models: Sparsity, disentanglement, and gradient-conflict-free," *arXiv preprint arXiv:2601.02967*, 2026.

[54] M. Gibier, R. Duroselle, P. Serrano, O. Boeffard, and J.-F. Bonastre, "Segmentwise pruning in audio-language models," *arXiv preprint arXiv:2511.14293*, 2025.

[55] S. BN, A. M. Sherrill, J. Alaparthi, D. Mattioli, R. I. Arriaga, C. W. Wiese, and S. Abdullah, "Fine-tuning large audio-language models with lora for precise temporal localization of prolonged exposure therapy elements," *arXiv preprint arXiv:2506.09707*, 2025.

[56] K. Luo, L. Lin, Y. Zhang, M. Aloqaily, D. Wang, Z. Zhou, J. Zhang, K. Wang, L. Sun, and Q. Wen, "Chronosaudio: A comprehensive long-audio benchmark for evaluating audio-large language models," *arXiv preprint arXiv:2601.04876*, 2026.

[57] Y. Chaichana, P. Taveekitworachai, W. Sirichotedumrong, P. Manakul, and K. 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)*, 2026, pp. 6046–6066.

[58] H. Wang, Y. Li, S. Ma, H. Liu, and X. Wang, "Listening between the frames: Bridging temporal gaps in large audio-language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 31, 2026, pp. 26 233–26 241.

[59] J. Hu, Z. Li, B. Qi, G. Liu, and P. Wang, "End-to-end contrastive language-speech pretraining model for long-form spoken question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 37, 2026, pp. 31 041–31 049.

[60] D. Zhang, G. Wang, J. Xue, K. Fang, L. Zhao, R. Ma, S. Ren, S. Liu, T. Guo, W. Zhuang *et al.*, "Mimo-audio: Audio language models are few-shot learners," *arXiv preprint arXiv:2512.23808*, 2025.

[61] J. Wang, Z. Ma, Z. Luo, T. Wang, M. Ge, X. Wang, and L. Wang, "Pay more attention to audio: Mitigating imbalance of cross-modal attention in large audio language models," *arXiv preprint arXiv:2509.18816*, 2025.

[62] H. He, X. Du, R. Sun, Z. Dai, Y. Xiao, M. Yang, J. Zhou, X. Li, Z. Liu, Z. Liang *et al.*, "Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models," *arXiv preprint arXiv:2509.21060*, 2025.

[63] P. Grinberg and H. Shahmohammadi, "Alarm: Audio-language alignment for reasoning models," *arXiv preprint arXiv:2603.09556*, 2026.

[64] Q. Wang, X. Jiang, L. He, J. Wu, and N. Mesgarani, "Sightsound-r1: Cross-modal reasoning distillation from vision to audio language models," *arXiv preprint arXiv:2509.15661*, 2025.

[65] J. Hu, D. Zhu, X. Luo, D. Zhang, S. He, Y. Lei, H. Zheng, S. Feng, J. He, Y. Sun *et al.*, "Cord: Bridging the audio-text reasoning gap via weighted on-policy cross-modal distillation," *arXiv preprint arXiv:2601.16547*, 2026.

[66] Q. Yang, B. Zhao, Z. Kang, X. Li, Y. He, C. Liu, X. Zhang, X. Qu, J. Peng, and J. Wang, "Attention-weighted centered kernel alignment for knowledge distillation in large audio-language models applied to speech emotion recognition," *arXiv preprint arXiv:2602.01547*, 2026.

[67] J. Zhao, C. Li, J. Zhao, R. Chen, D. Yu, M. D. Plumbley, and W. Wang, "Feedback-driven retrieval-augmented audio generation with large audio language models," *arXiv preprint arXiv:2511.01091*, 2025.

[68] J. Shi, H. Du, Y. A. Hong, and Y. Gao, "Emo-tta: Improving test-time adaptation of audio-language models for speech emotion recognition," *arXiv preprint arXiv:2509.25495*, 2025.

[69] S. Ji, Y. Chen, M. Fang, J. Zuo, J. Lu, H. Wang, Z. Jiang, L. Zhou, S. Liu, X. Cheng, X. Yang, Z. Wang, Q. Yang, J. Li, Y. Jiang, J. He, Y. Chu, J. Xu, and Z. Zhao, "Wavchat: A survey of spoken dialogue models," *arXiv preprint arXiv:2411.13577*, 2024.

[70] Y. Chen and H. Yu, "From turn-taking to synchronous dialogue: A survey of full-duplex spoken language models," *arXiv preprint arXiv:2509.14515*, 2025.

[71] X. Zhang, Y. Chen, S. Hu, X. Han, Z. Xu, Y. Xu, W. Zhao, M. Sun, and Z. Liu, "Beyond the turn-based game: Enabling real-time conversations with duplex models," *Conference on Empirical Methods in Natural Language Processing*, pp. 11 543–11 557, 2024.

[72] W. Yu, S. Wang, X. Yang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, Y. Wang, and C. Zhang, "Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation," *arXiv preprint arXiv:2411.18138*, 2024.

[73] K. Hu, E. Hosseini-Asl, C. Chen, E. Casanova, S. Ghosh, P. Żelasko, Z. Chen, J. Li, J. Balam, and B. Ginsburg, "Efficient and direct duplex modeling for speech-to-speech language model," *arXiv preprint arXiv:2505.15670*, 2025.

[74] Z. Liu, Y. Duan, M. Wang, P. Feng, H. Zhang, X. Xing, Y. Shan, H. Zhu, Y. Dai, C. Lu, X. Qiu, L. Xie, L. Wang, N. Yan, Z. Zheng, Z. Ma, K. Yu, and X. Chen, "X-talk: On the underestimated potential of modular speech-to-speech dialogue system," *arXiv preprint arXiv:2512.18706*, 2025.

[75] R. Yan, W. Chen, Z. Liu, Z. Ma, H. Lin, H. Wen, H. Xie, J. Wu, Y. Liang, Y. Zhao *et al.*, "Soulx-duplug: Plug-and-play streaming state prediction module for realtime full-duplex speech conversation," *arXiv preprint arXiv:2603.14877*, 2026.

[76] K.-W. Chang, W.-C. Chen, E.-P. Hu, H.-y. Lee, and J. Glass, "Tico: Time-controllable training for spoken dialogue models," *arXiv preprint arXiv:2603.22267*, 2026.

[77] C.-Y. Hsiao, K.-H. Lu, Y.-K. Fu, G.-T. Lin, H.-T. Hung, and H.-y. Lee, "Aspirin: Action space projection for interactivity-optimized reinforcement learning in full-duplex speech language models," *arXiv preprint arXiv:2604.10065*, 2026.

[78] Y. Yao, X. Li, X. Jiang, X. Fang, N. Yu, W. Ma, A. Sun, and Y. Wang, "Flm-audio: Natural monologues improves native full-duplex chatbots via dual training," *arXiv preprint arXiv:2509.02521*, 2025.

[79] C.-M. Chien, M. Orsini, E. Kharitonov, N. Zeghidour, K. Livescu, and A. Défossez, "Moshirag: Asynchronous knowledge retrieval for full-duplex speech language models," *arXiv preprint arXiv:2604.12928*, 2026.

[80] D. Wu, T. Zhang, Y. Li, H. Liu, C. Chen, E. S. Chng, and Y. Bengio, "The silent thought: Modeling internal cognition in full-duplex

spoken dialogue models via latent reasoning," *arXiv preprint arXiv:2603.17837*, 2026.

[81] N. Kuzmin, T. Zhong, J. Deng, Y. Zhu, T. Tsoi, T. Cao, S. Lui, K. A. Lee, and E. S. Chng, "Privacy-preserving end-to-end full-duplex speech dialogue models," *arXiv preprint arXiv:2603.08179*, 2026.

[82] W. Wang, C. Li, L. Zhang, Y. Zhao, Y. Zou, H. Li, M. Cui, H. Zhang, K. Wei, L. Xu *et al.*, "Covo-audio technical report," *arXiv preprint arXiv:2602.09823*, 2026.

[83] Z. Xiong, Y. Cai, Z. Li, J. Yuan, and Y. Wang, "Thinking with sound: Audio chain-of-thought enables multimodal reasoning in large audio-language models," *arXiv preprint arXiv:2509.21749*, 2025.

[84] D. Wu, X. Zhang, D. Yang, J. Yao, L. Chen, Q. Liu, S. Zhao, C. Ma, Y. Kang, and Y. Zhou, "Echo: Towards advanced audio comprehension via audio-interleaved reasoning," *arXiv preprint arXiv:2602.11909*, 2026.

[85] L.-L. Ieong, C.-C. Chen, C.-K. Yang, Y.-H. Huang, A.-Y. Cheng, and H.-y. Lee, "Nudging hidden states: Training-free model steering for chain-of-thought reasoning in large audio-language models," *arXiv preprint arXiv:2603.14636*, 2026.

[86] Y.-J. Shih, D. Raj, C. Wu, W. Zhou, S. Bong, Y. Gaur, J. Mahadeokar, O. Kalinli, and M. Seltzer, "Can speech llms think while listening?" *arXiv preprint arXiv:2510.07497*, 2025.

[87] J. Fan, R. Ren, J. Li, R. Pandey, P. G. Shivakumar, I. Bulyko, A. Gandhe, G. Liu, and Y. Gu, "Incentivizing consistent, effective and scalable reasoning capability in audio llms via reasoning process rewards," *arXiv preprint arXiv:2510.20867*, 2025.

[88] P. Li, B. Zhao, Z. Kang, J. Peng, X. Qu, Y. He, and J. Wang, "Emo-rl: Emotion-rule-based reinforcement learning enhanced audio-language model for generalized speech emotion recognition," *arXiv preprint arXiv:2509.15654*, 2025.

[89] S. Wu, C. Li, W. Wang, H. Zhang, H. Wang, M. Yu, and D. Yu, "Audio-thinker: Guiding audio language model when and how to think via reinforcement learning," *arXiv preprint arXiv:2508.08039*, 2025.

[90] X. Diao, C. Zhang, K. Kong, W. Wu, C. Ma, Z. Ouyang, P. Qing, S. Vosoughi, and J. Gui, "Soundmind: Rl-incentivized logic reasoning for audio-language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 528–540.

[91] Z. Sheng, S. Zhou, C. Gong, and Z. Li, "Think smart, not hard: Difficulty adaptive reasoning for large audio language models," *arXiv preprint arXiv:2509.21960*, 2025.

[92] H. Jia, W. Li, J. Wu, X. Yu, Y. Gao, J. Cheng, X. Tang, F. Xia, and T. Dang, "Decoding ambiguous emotions with test-time scaling in audio-language models," *arXiv preprint arXiv:2602.03873*, 2026.

[93] G. Wijngaard, E. Formisano, M. Dumontier, and J. Jitsev, "Audiotoolagent: An agentic framework for audio-language models," *arXiv preprint arXiv:2510.02995*, 2025.

[94] X. Zhou, J. Lian, H. Hong, X. Yang, and G. Anumanchipalli, "Speech world model: Causal state-action planning with explicit reasoning for speech," *arXiv preprint arXiv:2512.05933*, 2025.

[95] D. Huang, Y. Lv, R. Xiong, C. Jin, and X. Peng, "When tone and words disagree: Towards robust speech emotion recognition under acoustic-semantic conflict," *arXiv preprint arXiv:2601.04564*, 2026.

[96] G.-T. Lin, J. Lian, T. Li, Q. Wang, G. Anumanchipalli, A. H. Liu, and H.-y. Lee, "Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities," *arXiv preprint arXiv:2503.04721*, 2025.

[97] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed *et al.*, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.

[98] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[99] A. 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," Ph.D. dissertation, University of Massachusetts Dartmouth, 2025.

[100] E. Nachmani, A. Levkovitch, R. Hirsch, J. Salazar, C. Asawaroengchai, S. Mariooryad, E. Rivlin, R. Skerry-Ryan, and M. Tadmor Ramanovich, "Spoken question answering and speech continuation using spectrogram-powered llm," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 51 883–51 898.

[101] S. Liu, A. S. Hussain, C. Sun, and Y. 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)*. IEEE, 2024, pp. 286–290.

[102] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, "Joint audio and speech understanding," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[103] M. Wang, W. Han, I. Shafran, Z. Wu, C.-C. Chiu, Y. Cao, N. Chen, Y. Zhang, H. Soltau, P. 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)*. IEEE, 2023, pp. 1–8.

[104] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma *et al.*, "Lauragpt: Listen, attend, understand, and regenerate audio with gpt," *arXiv preprint arXiv:2310.04673*, 2023.

[105] G.-T. Lin, P. G. Shivakumar, A. Gandhe, C.-H. H. Yang, Y. Gu, S. Ghosh, A. Stolcke, H.-y. Lee, and I. Bulyko, "Paralinguistics-enhanced large language modeling of spoken dialogue," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 316–10 320.

[106] H. Xue, Y. Liang, B. Mu, S. Zhang, M. Chen, Q. Chen, and L. Xie, "E-chat: Emotion-sensitive spoken dialogue system with large language models," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 586–590.

[107] D. Zhang, X. Zhang, J. Zhan, S. Li, Y. Zhou, and X. Qiu, "Speechgpt-gen: Scaling chain-of-information speech generation," *arXiv preprint arXiv:2401.13527*, 2024.

[108] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," *arXiv preprint arXiv:2402.01831*, 2024.

[109] G.-T. Lin, C.-H. Chiang, and H.-y. 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)*, 2024, pp. 6626–6642.

[110] T. A. Nguyen, B. Muller, B. Yu, M. R. Costa-Jussa, M. Elbayad, S. Popuri, C. Ropers, P.-A. Duquenne, R. Algayres, R. Mavlyutov *et al.*, "Spirit-lm: Interleaved spoken and written language model," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 30–52, 2025.

[111] Z. Li, J. E. Smerdon, R. Seager, N. Siegert, and J. S. Mankin, "Emergent trends complicate the interpretation of the united states drought monitor (usdm)," *AGU Advances*, vol. 5, no. 2, p. e2023AV001070, 2024.

[112] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran *et al.*, "Wavllm: Towards robust and adaptive speech large language model," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 4552–4572.

[113] N. Das, S. Dingliwal, S. Ronanki, R. Paturi, Z. Huang, P. Mathur, J. Yuan, D. Bekal, X. Niu, S. M. Jayanthi *et al.*, "Speechverse: A large-scale generalizable audio language model," *arXiv preprint arXiv:2405.08295*, 2024.

[114] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. 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*, 2024, pp. 6288–6313.

[115] K. An, Q. Chen, C. Deng, Z. Du, C. Gao, Z. Gao, Y. Gu, T. He, H. Hu, K. Hu *et al.*, "Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms," *arXiv preprint arXiv:2407.04051*, 2024.

[116] Z. Xie and C. Wu, "Mini-omni: Language models can hear, talk while thinking in streaming," *arXiv preprint arXiv:2408.16725*, 2024.

[117] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[118] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "Llama-omni: Seamless speech interaction with large language models," *arXiv preprint arXiv:2409.06666*, 2024.

[119] Z. Meng, Q. Wang, W. Cui, Y. Zhang, B. Wu, I. King, L. Chen, and P. 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.

[120] Q. Zhang, L. Cheng, C. Deng, Q. Chen, W. Wang, S. Zheng, J. Liu, H. Yu, C.-H. Tan, Z. 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)*, 2025, pp. 14 570–14 580.

[121] X. Zhang, X. Lyu, Z. Du, Q. Chen, D. Zhang, H. Hu, C. Tan, T. Zhao, Y. Wang, B. Zhang *et al.*, "Intrinsicvoice: Empowering llms with intrinsic real-time voice interaction abilities," *arXiv preprint arXiv:2410.08035*, 2024.

[122] W. Held, Y. Zhang, M. Li, W. Shi, M. J. Ryan, and D. 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)*, 2025, pp. 7876–7891.

[123] X. Wang, Y. Li, C. Fu, Y. Shen, L. Xie, K. Li, X. Sun, and L. Ma, "Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm," *arXiv preprint arXiv:2411.00774*, 2024.

[124] A. Zeng, Z. Du, M. Liu, K. Wang, S. Jiang, L. Zhao, Y. Dong, and J. Tang, "Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot," *arXiv preprint arXiv:2412.02612*, 2024.

[125] S. Zhao, T. Guo, C. Wen, B. Xiang, and W. Zou, "Ke-omni-r: Achieving advanced audio reasoning with a concise 50-words think process," https://github.com/shuaijiang/Ke-Omni-R, 2025, gitHub repository.

[126] Y. He, Z. Liu, S. Sun, B. Wang, W. Zhang, X. Zou, N. F. Chen, and A. T. Aw, "Meralion-audiollm: Technical report," *arXiv e-prints*, pp. arXiv–2412, 2024.

[127] F. Tian, X. T. Zhang, Y. Zhang, H. Zhang, Y. Li, D. Liu, Y. Deng, D. Wu, J. Chen, L. Zhao *et al.*, "Step-audio-r1 technical report," *arXiv preprint arXiv:2511.15848*, 2025.

[128] T. Li, J. Liu, T. Zhang, Y. Fang, D. Pan, M. Wang, Z. Liang, Z. Li, M. Lin, G. Dong *et al.*, "Baichuan-audio: A unified framework for end-to-end speech interaction," *arXiv preprint arXiv:2502.17239*, 2025.

[129] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," *arXiv preprint arXiv:2503.03983*, 2025.

[130] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang *et al.*, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[131] Z. Long, Y. Shen, C. Fu, H. Gao, L. Li, P. Chen, M. Zhang, H. Shao, J. Li, J. Peng *et al.*, "Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model," *arXiv preprint arXiv:2505.03739*, 2025.

[132] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle *et al.*, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *arXiv preprint arXiv:2507.08128*, 2025.

[133] K.-H. Lu, Z. Chen, S.-W. Fu, C.-H. H. Yang, S.-F. Huang, C.-K. Yang, C.-E. Yu, C.-W. Chen, W.-C. Chen, C.-y. 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*, 2026.

[134] J. Chen, Y. Hu, J. Li, K. Li, K. Liu, W. Li, X. Li, Z. Li, F. Shen, X. Tang *et al.*, "Fireredchat: A pluggable, full-duplex voice interaction system with cascaded and semi-cascaded implementations," *arXiv preprint arXiv:2509.06502*, 2025.

[135] G. K. Kumar, R. Saraf, L. Lepauloux, A. Muneer, B. Mokeddem, and H. Hacid, "Competitive audio-language models with data-efficient single-stage training on public data," *arXiv preprint arXiv:2509.07526*, 2025.

[136] C. Yan, B. Wu, P. Yang, P. Tan, G. Hu, L. Xie, Y. Zhang, F. Tian, X. Yang, X. Zhang *et al.*, "Step-audio-editx technical report," *arXiv preprint arXiv:2511.03601*, 2025.

[137] C. Liu, M. Aljunied, G. Chen, H. P. Chan, W. Xu, Y. Rong, and W. Zhang, "Seallms-audio: Large audio-language models for southeast asia," *arXiv preprint arXiv:2511.01670*, 2025.

[138] T. F. Team, Q. Chen, L. Cheng, C. Deng, X. Li, J. Liu, C.-H. Tan, W. Wang, J. Xu, J. Ye *et al.*, "Fun-audio-chat technical report," *arXiv preprint arXiv:2512.20156*, 2025.

[139] D. Zhang, Y. Lei, J. Hu, S. He, S. Deng, X. Luo, D. Zhu, S. Feng, R. Liu, J. He *et al.*, "Eureka-audio: Triggering audio intelligence in compact language models," *arXiv preprint arXiv:2602.13954*, 2026.

[140] J. Chen, Z. Guo, J. Chun, P. Wang, A. Perrault, and M. Elsner, "Do audio llms really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance," *arXiv preprint arXiv:2510.10444*, 2025.

[141] C. Wang, G. Deng, X. Yang, H. Qiu, and T. 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*, 2025, pp. 4878–4888.

[142] G. Morais and M. Fuentes, "Investigating modality contribution in audio llms for music," *arXiv preprint arXiv:2509.20641*, 2025.

[143] C.-A. Li, T.-H. Lin, and H.-y. Lee, "When silence matters: The impact of irrelevant audio on text reasoning in large audio-language models," *arXiv preprint arXiv:2510.00626*, 2025.

[144] R. Zhang, Z. Liu, L. Sun, T. Zhu, and W. Lv, "The sonar moment: Benchmarking audio-language models in audio geo-localization," *arXiv preprint arXiv:2601.03227*, 2026.

[145] L. Jain, P. Mousavi, M. Ravanelli, and C. Subakan, "Investigating faithfulness in large audio language models," *arXiv preprint arXiv:2509.22363*, 2025.

[146] Z. Ma, X. Li, Y. Song, W. Chen, C. Du, J. Wu, Y. Chen, Z. Chen, Y. Wang, Y. Wang *et al.*, "Towards reliable large audio language model," *arXiv preprint arXiv:2505.19294*, 2025.

[147] F. López, S. Kesiraju, and J. Luque, "Robustness assessment of large audio language models in multiple-choice evaluation," *arXiv preprint arXiv:2510.04584*, 2025.

[148] B. Li, W. Huang, Y. Qiu, Y. Guo, H. Wang, Z. Li, J. Peng, Z. Ma, X. Chen, and K. Yu, "Isa-bench: Benchmarking instruction sensitivity for large audio language models," *arXiv preprint arXiv:2510.23558*, 2025.

[149] V. S. Sadasivan, S. Feizi, R. Mathews, and L. Wang, "Attacker's noise can manipulate your audio-based llm in the real world," *arXiv preprint arXiv:2507.06256*, 2025.

[150] Y. Yu, H. Jin, Y. Yu, J. Zhuang, and H. Wang, "Now you hear me: Audio narrative attacks against large audio-language models," *arXiv preprint arXiv:2601.23255*, 2026.

[151] A. Fortier, T. Thebaud, J. Villalba, N. Dehak, and P. Cardinal, "Backdoor attacks against speech language models," *arXiv preprint arXiv:2510.01157*, 2025.

[152] Y. Ren, X. Xu, B. Li, S. Wang, and C. Zhang, "Can audio large language models verify speaker identity?" 2025. [Online]. Available: https://arxiv.org/abs/2509.19755

[153] Y. Xie, X. Guo, J. Zhou, T. Wang, J. Liu, R. Fu, X. Wang, H. Cheng, and L. Ye, "Interpretable all-type audio deepfake detection with audio llms via frequency-time reinforcement learning," *arXiv preprint arXiv:2601.02983*, 2026.

[154] X. Xu, Y. Ren, L. Liu, W. Wu, B. Li, C. Lu, S. Wang, and C. Zhang, "Holiantispoof: Audio llm for holistic speech anti-spoofing," *arXiv preprint arXiv:2602.04535*, 2026.

[155] L. Zhang, T. Thebaud, Z. Cai, S. Khudanpur, D. Povey, L. P. García-Perera, M. Wiesner, and N. Andrews, "Can llms help localize fake words in partially fake speech?" *arXiv preprint arXiv:2603.11205*, 2026.

[156] J. Wang, L. Lin, K. Luo, W. Wang, Y. Chen, M. Aloqaily, X. Tang, Z. Zhou, K. Wang, L. Sun *et al.*, "Hearsay benchmark: Do audio llms leak what they hear?" *arXiv preprint arXiv:2601.03783*, 2026.

[157] X. Zhan, G. Sun, J. Such, and P. Woodland, "Protecting bystander privacy via selective hearing in lalms," *arXiv preprint arXiv:2512.06380*, 2025.

[158] Z. R. Tam and Y.-N. Chen, "Medvoicebias: A controlled study of audio llm behavior in clinical decision-making," *arXiv preprint arXiv:2511.06592*, 2025.

[159] S.-L. Wei, Y.-L. Liao, Y.-H. Chang, H.-H. Huang, and H.-H. 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.

[160] Y.-X. Lin, C.-A. Li, S.-L. Wei, P.-C. Chen, H.-H. Chen, and H.-y. Lee, "Hearing the order: Investigating selection bias in large audio-language models," *arXiv preprint arXiv:2510.00628*, 2025.

[161] Z. H. Pang, X. Gao, T. Kawahara, and N. F. Chen, "Erm-minmaxgap: Benchmarking and mitigating gender bias in mul-

tilingual multimodal speech-llm emotion recognition," *arXiv preprint arXiv:2603.21050*, 2026.

[162] H. Li, C. Zhou, C. Wang, S. Liang, Y. Chen, Q. Xie, J. Ye, and J. Wu, "Stylebreak: Revealing alignment vulnerabilities in large audio-language models via style-aware audio jailbreak," *arXiv preprint arXiv:2511.10692*, 2025.

[163] B.-H. Feng, C.-F. Liu, Y.-H. L. Liang, C.-K. Yang, S.-W. Fu, Z. Chen, K.-H. Lu, S.-F. Huang, C.-H. H. Yang, Y.-C. F. Wang *et al.*, "Investigating safety vulnerabilities of large audio-language models under speaker emotional variations," *arXiv preprint arXiv:2510.16893*, 2025.

[164] J. Roh, V. Shejwalkar, and A. Houmansadr, "Multilingual and multi-accent jailbreaking of audio llms," *arXiv preprint arXiv:2504.01094*, 2025.

[165] B. Kim, H. Dingeto, T. Kwon, D. Choi, D. Lee, H. Park, J. Lee, and J. Shin, "When good sounds go adversarial: Jailbreaking audio-language models with benign inputs," *arXiv preprint arXiv:2508.03365*, 2025.

[166] I. Gupta, D. Khachaturov, and R. Mullins, "" i am bad": Interpreting stealthy, universal and robust audio jailbreaks in audio-language models," *arXiv preprint arXiv:2502.00718*, 2025.

[167] Z. Peng, Y. Liu, Z. Sun, M. Li, Z. Luo, J. Zheng, W. Dong, X. He, X. Wang, Y. Xue *et al.*, "Jalmbench: Benchmarking jailbreak vulnerabilities in audio language models," *arXiv preprint arXiv:2505.17568*, 2025.

[168] G. Chen, F. Song, Z. Zhao, X. Jia, Y. Liu, Y. Qiao, and W. Zhang, "Audiojailbreak: Jailbreak attacks against end-to-end large audio-language models," *arXiv preprint arXiv:2505.14103*, 2025.

[169] Z. Song, Q. Jiang, M. Cui, M. Li, L. Gao, Z. Zhang, Z. Xu, Y. Wang, C. Wang, G. Ouyang *et al.*, "Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models," *arXiv preprint arXiv:2505.15406*, 2025.

[170] H. Cheng, E. Xiao, J. Shao, Y. Wang, L. Yang, C. Shen, P. Torr, J. Gu, and R. Xu, "Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models," *arXiv preprint arXiv:2501.13772*, 2025.

[171] W. Jin, Y. Cao, J. Su, M. Xue, J. Hao, K. Xu, J. S. Dong, and D. Wang, "Almguard: Safety shortcuts and where to find them as guardrails for audio-language models," *arXiv preprint arXiv:2510.26096*, 2025.

[172] W. Lin, J. Li, H. Xiong, and L. Liu, "Sarsteer: Safeguarding large audio language models via safe-ablated refusal steering," *arXiv preprint arXiv:2510.17633*, 2025.

[173] H. Yang, L. Qu, E. Shareghi, and G. Haffari, "Reshaping representation space to balance the safety and over-rejection in large audio language models," *arXiv preprint arXiv:2505.19670*, 2025.

[174] K. Li, C. Shen, Y. Liu, J. Han, K. Zheng, X. Zou, Z. Wang, S. Zhang, X. Du, H. Luo *et al.*, "Audiotrust: Benchmarking the multifaceted trustworthiness of audio large language models," *arXiv preprint arXiv:2505.16211*, 2025.

[175] G. Hou, J. He, Y. Zhou, J. Guo, Y. Qiao, R. Zhang, and W. 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*, 2025, pp. 25 671–25 687.

[176] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[177] T. Huang, S. Hu, F. Ilhan, S. F. Tekin, Z. Yahn, Y. Xu, and L. Liu, "Safety tax: Safety alignment makes your large reasoning models less reasonable," *arXiv preprint arXiv:2503.00555*, 2025.

[178] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. 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)*, 2025, pp. 4297–4316.

[179] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," *arXiv preprint arXiv:2410.19168*, 2024.

[180] Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," *Transactions of the Association for Computational Linguistics*, vol. 14, pp. 378–398, 2026.

[181] R. Yan, X. Li, W. Chen, Z. Niu, C. Yang, Z. Ma, K. Yu, and X. Chen, "Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models," *arXiv preprint arXiv:2502.17810*, 2025.

[182] F. Jiang, Z. Lin, F. Bu, Y. Du, B. Wang, and H. Li, "S2s-arena, evaluating speech2speech protocols on instruction following with paralinguistic information," *arXiv preprint arXiv:2503.05085*, 2025.

[183] S. Arora, Z. Lu, C.-C. Chiu, R. Pang, and S. Watanabe, "Talking turns: Benchmarking audio foundation models on turn-taking dynamics," *arXiv preprint arXiv:2503.01174*, 2025.

[184] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong *et al.*, "Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," *arXiv preprint arXiv:2505.13032*, 2025.

[185] C.-K. Yang, N. Ho, Y.-T. Piao, and H.-y. Lee, "Sakura: On the multi-hop reasoning of large audio-language models based on speech and audio information," *arXiv preprint arXiv:2505.13237*, 2025.

[186] H. Liu, Y. Wang, Z. Cheng, H. Liu, Y. Li, Y. Hou, R. Wu, Q. Gu, Y. Wang, and Y. Wang, "Vocalbench: Benchmarking the vocal conversational abilities for speech interaction models," *arXiv preprint arXiv:2505.15727*, 2025.

[187] X. Cheng, D. Fu, C. Wen, S. Yu, Z. Wang, S. Ji, S. Arora, T. Jin, S. Watanabe, and Z. Zhao, "Aha-bench: Benchmarking audio hallucinations in large audio-language models," *Advances in Neural Information Processing Systems*, vol. 38, 2026.

[188] Y. Kim, T. Kim, W. Kang, E. Park, J. Yoon, D. Lee, X. Liu, D. McDuff, H. Lee, C. Breazeal *et al.*, "Vocalagent: Large language models for vocal health diagnostics with safety-aware evaluation," *arXiv preprint arXiv:2505.13577*, 2025.

[189] D. Wang, J. Wu, J. Li, D. Yang, X. Chen, T. Zhang, and H. Meng, "Mmsu: A massive multi-task spoken language understanding and reasoning benchmark," *arXiv preprint arXiv:2506.04779*, 2025.

[190] Y. Hou, H. Liu, Y. Wang, Z. Cheng, R. Wu, Q. Gu, Y. Wang, and Y. Wang, "Sova-bench: Benchmarking the speech conversation ability for llm-based voice assistant," *arXiv preprint arXiv:2506.02457*, 2025.

[191] L. Zhang, J. Zhang, B. Lei, C. Wu, A. Liu, W. Jia, and X. Zhou, "Wildspeech-bench: Benchmarking end-to-end speechllms in the wild," *arXiv preprint arXiv:2506.21875*, 2025.

[192] H. Wang, L. Ma, D. Guo, X. Wang, L. Xie, J. Xu, and J. Lin, "Contextasr-bench: A massive contextual speech recognition benchmark," *arXiv preprint arXiv:2507.05727*, 2025.

[193] C. Ma, W. Tao, and S. 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*, 2025, pp. 22 789–22 807.

[194] J. Kim, H. Yun, S. H. Woo, C.-H. H. Yang, and G. Kim, "Wow-bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations," *arXiv preprint arXiv:2508.20976*, 2025.

[195] S. Kumar, Š. Sedláček, V. Lokegaonkar, F. López, W. Yu, N. Anand, H. Ryu, L. Chen, M. Plička, M. Hlaváček *et al.*, "Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence," *arXiv preprint arXiv:2508.13992*, 2025.

[196] W. Yang, Y. Li, Y. Wei, M. Fang, and L. Chen, "Speechr: A benchmark for speech reasoning in large audio-language models," *arXiv preprint arXiv:2508.02018*, 2025.

[197] K. Wang, H. Ren, Z. Lu, M. Zhan, and H. Li, "Voiceassistant-eval: Benchmarking ai assistants across listening, speaking, and viewing," *arXiv preprint arXiv:2509.22651*, 2025.

[198] B. J. Carone, I. R. Roman, and P. Ripollés, "The muse benchmark: Probing music perception and auditory relational reasoning in audio llms," *arXiv preprint arXiv:2510.19055*, 2025.

[199] P. He, Z. Wen, Y. Wang, Y. Wang, X. Liu, J. Huang, Z. Lei, Z. Gu, X. Jin, J. Yang *et al.*, "Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms," *arXiv preprint arXiv:2510.07293*, 2025.

[200] S. H. B. Satish, G. E. Henter, and É. Székely, "Do bias benchmarks generalise? evidence from voice-based evaluation of gender bias in speechllms," *arXiv preprint arXiv:2510.01254*, 2025.

[201] T. Guo, H. Chen, H. Liang, M. Qiang, B. Zeng, L. Sun, B. Cui, and W. Zhang, "Brace: A benchmark for robust audio caption quality evaluation," *arXiv preprint arXiv:2512.10403*, 2025.

[202] Y.-J. Lu, K. Gao, M. Liang, H. Wang, T. Thebaud, L. Moro-Velazquez, N. Dehak, and J. Villalba, "Spoken dialogsum: An emotion-rich conversational dataset for spoken dialogue summarization," *arXiv preprint arXiv:2512.14687*, 2025.

[203] Y. Peng, C. Cai, Z. Liu, S. Fan, S. Jiang, H. Xu, Y. Liu, Q. Chen, K. Xu, Y. Li *et al.*, "Mac-slu: Multi-intent automotive cabin spoken language understanding benchmark," *arXiv preprint arXiv:2512.01603*, 2025.
[204] C. Yang, K. Huang, L. Fan, Q. Tu, B. Jiang, D. Zhang, L. Yin, S. Li, Z. Fei, Q. Cheng *et al.*, "Wesr: Scaling and evaluating word-level event-speech recognition," *arXiv preprint arXiv:2601.04508*, 2026.
[205] Y. Zhang, L. Lin, K. Luo, S. Yan, J. Wang, Y. Guo, Y. Chen, Y. Qin, Z. Zhou, K. Wang *et al.*, "Rsa-bench: Benchmarking audio large models in real-world acoustic scenarios," *arXiv preprint arXiv:2601.10384*, 2026.
[206] Y. Wang, X. Qian, T.-H. Zhang, J. Gao, Y. Pan, X. Wang, Z. Pan, C. Wei, and Y. Wang, "Palm-bench: A comprehensive benchmark for personalized audio-language models," *arXiv preprint arXiv:2601.03531*, 2026.
[207] S. Wang, Z. Zhao, H. Yue, C. Wang, S. Wang, H. Bu, X. Xu, and L. Xie, "Humdial-eibench: A human-recorded multi-turn emotional intelligence benchmark for audio language models," 2026. [Online]. Available: https://arxiv.org/abs/2604.11594
[208] Y. Wang, H. Liu, Y. Xu, Q. Ni, L. Wang, W. Lin, K. Feng, D. Chen, X. Tan, L. Wang *et al.*, "Voxsafebench: Not just what is said, but who, how, and where," *arXiv preprint arXiv:2604.14548*, 2026.
[209] F. Zhao, Y. Chen, W. Lu, D. Zhang, X. Yue, and J. Wei, "Halluaudio: A comprehensive benchmark for hallucination detection in large audio-language models," 2026. [Online]. Available: https://arxiv.org/abs/2604.19300
[210] NWU-LIST, "Aabench: A comprehensive benchmark for evaluating audio adversarial robustness in llms," https://huggingface.co/datasets/NWULIST/AABench, 2026, accessed: 2026-05-08.
[211] ——, "Speechjailbreaker: A framework for evaluating and breaking audio-language model safety," https://github.com/NWULIST/SpeechJailbreaker, 2026, gitHub repository, Accessed: 2026-05-08.