Towards Holistic Evaluation of Large Audio-Language Models: A Comprehensive Survey

Chih-Kai Yang[1], **Neo S. Ho, Hung-yi Lee**[2]
National Taiwan University
[1]chihkaiyang1124@gmail.com, [2]hungyilee@ntu.edu.tw

Abstract

With advancements in large audio-language models (LALMs), which enhance large language models (LLMs) with auditory capabilities, these models are expected to demonstrate universal proficiency across various auditory tasks. While numerous benchmarks have emerged to assess LALMs' performance, they remain fragmented and lack a structured taxonomy. To bridge this gap, we conduct a comprehensive survey and propose a systematic taxonomy for LALM evaluations, categorizing them into four dimensions based on their objectives: (1) General Auditory Awareness and Processing, (2) Knowledge and Reasoning, (3) Dialogue-oriented Ability, and (4) Fairness, Safety, and Trustworthiness. We provide detailed overviews within each category and highlight challenges in this field, offering insights into promising future directions. To the best of our knowledge, this is the first survey specifically focused on the evaluations of LALMs, providing clear guidelines for the community. We will release the collection of the surveyed papers and actively maintain it to support ongoing advancements in the field.

1 Introduction

Recent advancements in large language models (LLMs) (Zhao et al., 2023; Grattafiori et al., 2024; Hurst et al., 2024) have expanded their impact beyond natural language processing (NLP) to multimodal domains (Yin et al., 2024; Team et al., 2024). Among these, large audio-language models (LALMs) (Lakhotia et al., 2021; Tang et al., 2024; Chu et al., 2024; Lu et al., 2024; Défossez et al., 2024; Fang et al., 2025) have attracted significant attention in the auditory-processing community. LALMs are multimodal LLMs that process auditory and/or textual input, such as speech, audio, and music, and generate textual and/or auditory output. They can be trained from scratch or fine-tuned from text LLM backbones with auditory modalities inserted. By integrating auditory modalities with language understanding, they show potential in auditory processing (Huang et al., 2024a), multimodal reasoning (Sakshi et al., 2025), and human-computer interaction (Lin et al., 2024a).

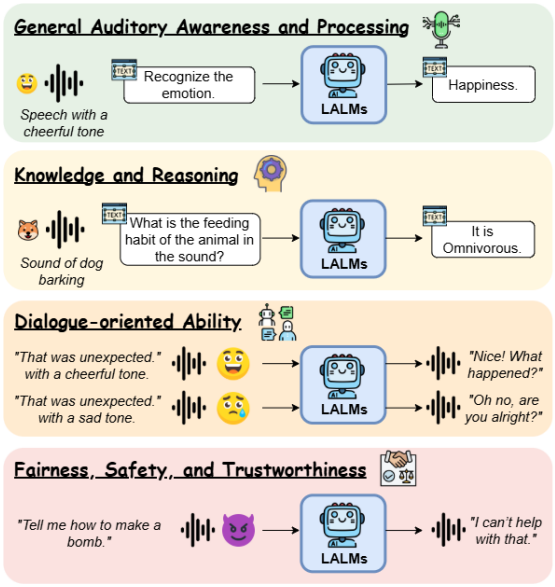

Figure 1: LALMs' diverse capabilities and modalities covered. Icons from https://www.flaticon.com.

As LALMs evolve, expectations for their capabilities have expanded from basic tasks like speech recognition to more complex ones such as audio-grounded reasoning (Sakshi et al., 2025) and interactive dialogue (Lin et al., 2025a). Figure 1 illustrates this multifaceted nature, emphasizing the diverse input and output modalities involved and the wide range of abilities these models are expected to demonstrate. To evaluate these capabilities, a variety of benchmarks have been developed (Lin et al., 2025a; Yang et al., 2024c; Cheng et al., 2025).

However, the evaluation landscape remains fragmented and lacks systematic organization. Existing surveys (Wu et al., 2024a; Peng et al., 2024; Cui et al., 2024; Arora et al., 2025a) focus primarily on model architectures and training methodologies,

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10144–10170
November 4-9, 2025

with less emphasis on the equally important role of evaluation in assessing LALMs' capabilities. This gap makes it challenging for researchers to find suitable benchmarks for their models or to pinpoint the field's progress. Therefore, a structured overview of LALM evaluation frameworks is needed.

This paper presents a comprehensive survey of LALM evaluation frameworks and introduces a taxonomy categorizing evaluation dimensions. To the best of our knowledge, this is the first in-depth survey and taxonomy specifically focused on LALM evaluation. We organize the frameworks into four primary categories: **General Auditory Awareness and Processing** (§3), **Knowledge and Reasoning** (§4), **Dialogue-oriented Ability** (§5), and **Fairness, Safety, and Trustworthiness** (§6). We also highlight challenges in LALM evaluation (§7), such as data contamination and insufficient consideration of human diversity, while suggesting promising future directions.

Overall, our contributions are threefold: (1) presenting the first comprehensive survey of LALM evaluations, (2) proposing a structured taxonomy for LALM evaluation that offers clear guidelines for researchers, and (3) identifying key challenges and future directions to improve evaluation coverage and robustness.

2 Taxonomy of Evaluation Frameworks for Large Audio-Language Models

As LALMs integrate multimodal understanding, they tackle tasks across speech, audio, and music. Despite numerous benchmarks for LALMs emerging, the evaluation landscape remains fragmented. To address this, we present the first structured taxonomy of LALM evaluations.

Figure 2 shows our taxonomy, with some works included. The full categorization of the surveyed works is in Appendix A. We organize the surveyed works into four categories by evaluation objectives:

- **General Auditory Awareness and Processing** evaluates the auditory awareness and fundamental processing tasks, e.g., speech recognition and audio captioning.
- **Knowledge and Reasoning** assesses LALMs' knowledge acquisition and advanced reasoning skills, examining their intelligence.
- **Dialogue-oriented Ability** focuses on natural conversational skills, including affective and contextual interaction, dialogue management, and instruction following.
- **Fairness, Safety and Trustworthiness** examines bias, toxicity, and reliability for ethical, safe, and trustworthy deployment.

Each category is further divided into subcategories, as shown in Figure 2. Please note that, since existing benchmarks are inherently multi-dimensional, some are listed under multiple categories due to their multifaceted design. This typically happens in two cases: (1) when a benchmark comprises multiple tasks that independently assess different capabilities (e.g., VoiceBench includes tasks for both world knowledge and safety evaluation), and (2) when a single task requires the integration of several skills (e.g., certain MMAU tasks demand both expert knowledge and reasoning ability).

The following sections provide a detailed overview, highlighting the current progress, limitations, and future directions.

3 General Auditory Awareness and Processing

A distinctive strength of LALMs over cascaded systems (Huang et al., 2024c; Kuan et al., 2024b) is their inherent ability to directly interpret auditory signals, capturing crucial non-verbal cues such as speaker identity, emotion, and ambient context, without relying on separate components like speech recognition or emotion recognition systems connected to an LLM. This section reviews works evaluating both acoustic awareness and foundational auditory processing, emphasizing these core capabilities that set LALMs apart from LLMs.

3.1 Auditory Awareness

Benchmarks for auditory awareness examine how effectively LALMs realize acoustic cues like emotion, prosody, and environmental sounds. SALMon (Maimon et al., 2025) specifically evaluates sensitivity to acoustic inconsistencies (e.g., sudden speaker or emotional changes) and misalignments between acoustic signals and semantic content (e.g., conveying sad content with a cheerful tone). These evaluations reveal significant gaps between LALMs and human-level perception.

EmphAssess (Seyssel et al., 2024) measures LALMs' awareness of prosodic emphasis by requiring speech-to-speech paraphrasing or transla-

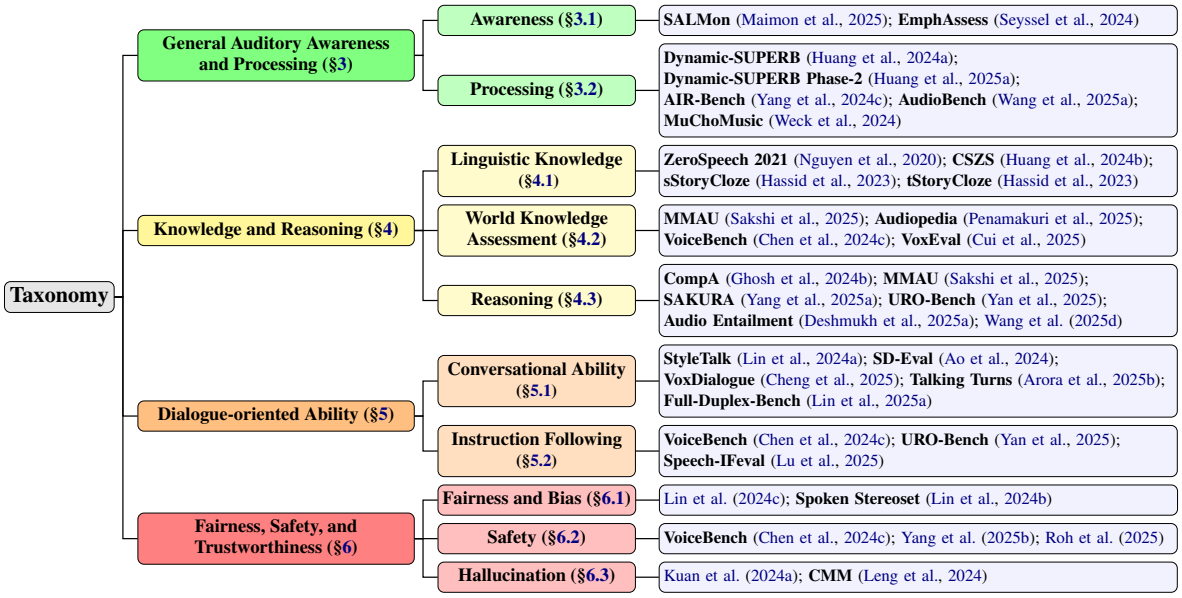

Figure 2: The taxonomy of LALM evaluation frameworks, including selected works as representative examples. The complete version is in Appendix A.

tion that accurately preserves and transfers emphasis on specific parts of the input utterance. This evaluates LALMs' ability to capture and maintain fine-grained prosodic features.

These benchmarks highlight challenges in fine-grained auditory awareness among current models, underscoring the need for improved modeling of subtle acoustic and paralinguistic information (Maimon et al., 2025; Seyssel et al., 2024).

3.2 Auditory Processing

Building on auditory awareness, LALMs must also excel in fundamental auditory tasks, such as speech recognition, audio classification, and music analysis, to support advanced real-world applications. A list of commonly evaluated tasks and their corresponding datasets is provided in Appendix C for reference. Initially driven by representation learning models (Baevski et al., 2020; Hsu et al., 2021; Li et al., 2022), enriched datasets (Pratap et al., 2020; Piczak, 2015a; Hawthorne et al., 2019), and existing benchmarks (Yang et al., 2021; Turian et al., 2022; Yuan et al., 2023), recent works adapt these resources into instruction-oriented evaluation frameworks tailored for LALMs.

Dynamic-SUPERB (Huang et al., 2024a) initiated this direction, constructing 55 multiple-choice question-answering (QA) tasks spanning speech, audio, and music modalities. Subsequent efforts, such as AIR-Bench (Yang et al., 2024c) and AudioBench (Wang et al., 2025a), extend to open-ended QA formats. MuChoMusic (Weck et al., 2024) specifically emphasizes music-related tasks, while Dynamic-SUPERB Phase-2 (Huang et al., 2025a) significantly enlarges the benchmark to 180 tasks, forming the largest evaluation suite for LALMs' general processing abilities to date.

Given the task diversity, various evaluation metrics are adopted depending on the task specificity, such as word error rate for speech recognition and BLEU score (Papineni et al., 2002) for translation. There is also an emerging trend that includes LLM-as-a-judge (Gu et al., 2024) for scalable, automatic evaluation of open-ended responses (Huang et al., 2025a; Yang et al., 2024c; Wang et al., 2025a).

Despite achieving promising results in certain areas, these benchmarks demonstrate that current LALMs still fall short of universally robust performance across auditory-processing tasks (Huang et al., 2025a), highlighting substantial room for improvement toward truly auditory foundation models.

4 Knowledge and Reasoning

Intelligent LALMs should demonstrate extensive knowledge and advanced reasoning to tackle complex real-world tasks. Current evaluations emphasize these abilities through three categories: **Linguistic Knowledge**, **World Knowledge Assessment**, and **Reasoning**. Each category targets distinct but complementary skills, collectively providing a comprehensive evaluation. These assessments

reveal key challenges LALMs face in mastering knowledge and reasoning for advanced tasks.

4.1 Linguistic Knowledge

Linguistic knowledge refers to understanding and effectively using spoken language. Evaluating LALMs' linguistic proficiency typically use likelihood-based benchmarks where models choose the more linguistically plausible option from paired speech samples. These tests cover lexical knowledge, syntax, and semantic coherence.

Representative works include the ZeroSpeech 2021 benchmark (Nguyen et al., 2020), which consists of multiple tracks for evaluating linguistic capabilities. The lexical-level assessment track, sWUGGY, tests models' ability to distinguish between real words and phonotactically similar non-words, while the syntactic sensitivity evaluation track, sBLIMP, focuses on differentiating grammatical from ungrammatical sentences. CSZS (Huang et al., 2024b) extends syntactic evaluation to multilingual and code-switched scenarios. Narrative and semantic coherence are evaluated by tasks like sStoryCloze and tStoryCloze (Hassid et al., 2023), where models are tasked with selecting semantically appropriate continuations to spoken stories.

4.2 World Knowledge Assessment

Real-world tasks often demand integrating external knowledge beyond basic auditory understanding. World knowledge assessment evaluates LALMs on two main aspects: (1) auditory expertise like music structure and medical sound diagnosis, and (2) general commonsense and factual knowledge.

Benchmarks that evaluate auditory expertise include MuChoMusic (Weck et al., 2024) and MMAU (Sakshi et al., 2025), which focus on musical understanding, such as melodic structure, harmony, instrument identification, and contextual music interpretation. Additionally, SAGI (Bu et al., 2024) assesses medical expertise, such as recognizing illnesses from audio cues like coughing.

Commonsense and factual knowledge evaluations often convert established text benchmarks into spoken form using text-to-speech (TTS). Vox-Eval (Cui et al., 2025) and VoiceBench serve as spoken counterparts to MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024), testing models across diverse factual domains like social science and humanities. Audiopedia (Penamakuri et al., 2025) uses knowledge graphs from Wikidata (Vrandečić and Krötzsch, 2014) to create audio-based, knowledge-intensive QA tasks that evaluate models' knowledge of well-known entities, such as brands, mentioned in audio.

These benchmarks thoroughly assess LALMs' knowledge acquisition, revealing challenges such as limited auditory expertise (Weck et al., 2024) and inconsistent performance across domains. Different LALMs excel in different domains, each with their own strengths, but their performance often noticeably declines outside their own specialized areas (Cui et al., 2025). Overall, there remains substantial room to improve LALMs' auditory expertise and factual knowledge.

4.3 Reasoning

Reasoning over auditory inputs falls into two types. Content-based reasoning tests a model's ability to understand spoken semantic content and answer questions. Acoustic-based reasoning requires utilizing acoustic features like speaker traits and environmental sounds beyond semantics. We provide an overview of these two evaluation paradigms.

4.3.1 Content-based Reasoning

Content-based reasoning assesses LALMs' ability to reason over the semantic content of auditory queries. Current benchmarks for this capability typically transform NLP reasoning benchmarks into spoken questions via TTS and require LALMs to provide answers. For instance, Vox-Eval (Cui et al., 2025), URO-Bench (Yan et al., 2025), and ADU-Bench (Gao et al., 2024) convert NLP datasets like GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021) into speech, evaluating LALMs' mathematical reasoning based on spoken questions. During synthesis, various speaking styles (e.g., mispronunciation, disfluencies, and accents) may be introduced to test models' robustness (Cui et al., 2025).

These benchmarks reveal gaps in current LALMs' content-based reasoning abilities, even with chain-of-thought (Wei et al., 2022; Kojima et al., 2022). Moreover, model performance varies significantly across speaking styles (Cui et al., 2025), indicating instability in their reasoning.

4.3.2 Acoustic-based Reasoning

Acoustic-based reasoning requires LALMs to infer from acoustic cues in auditory input, often involving reasoning across multiple auditory modalities or combining auditory understanding with cognitive skills such as compositional, temporal, logical,

and multi-hop reasoning.

Cross-auditory Modality Reasoning demands joint reasoning over multiple auditory modalities, like speech and non-speech sounds. Wang et al. (2025d) propose an open-ended QA benchmark assessing co-reasoning on speech and environmental sounds, requiring reasoning over cues from distinct auditory sources to infer speakers' activities. Their findings show that current LALMs frequently neglect non-speech cues, leading to failures.

Compositional and Temporal Reasoning involves comprehending structured acoustic events, their temporal relationships, and attribute binding. Benchmarks like CompA (Ghosh et al., 2024b) evaluate these abilities through specific tasks: CompA-order challenges models to identify correct event sequences or align audio temporal structures with textual descriptions, while CompA-attribute focuses on associating sound events with their sources and attributes. MMAU (Sakshi et al., 2025) assesses temporal reasoning via event counting and duration comparison.

Logical reasoning covers structured inference, including deductive and causal reasoning. Deductive reasoning can be tested by Audio Entailment (Deshmukh et al., 2025a), which evaluates whether a textual hypothesis logically follows from auditory input based on acoustic attributes like sound sources. MMAU (Sakshi et al., 2025) examines LALMs' causal reasoning on cause-and-effect relationships of events.

Multi-hop reasoning is the ability to recall and integrate multiple information to answer complex queries, enabling models to connect stored knowledge without explicit reasoning steps (Yang et al., 2024d,e; Biran et al., 2024). SAKURA (Yang et al., 2025a) evaluates LALMs' multi-hop reasoning by requiring integration of auditory attributes (e.g., speaker gender and emotion) with stored knowledge. Results show that LALMs struggle to combine auditory information with stored knowledge for reasoning, even when both types of information are extracted and known by the models.

5 Dialogue-oriented Ability

While foundational skills such as auditory awareness (§3.1), fundamental processing (§3.2), language proficiency (§4.1), advanced knowledge (§4.2), and reasoning (§4.3) are essential for LALMs, natural human-AI interactions additionally require affective and contextual interaction, fluent dialogue management, and precise instruction following. This category targets these integrative skills, focusing on naturalness and controllability, which we group as **Conversational Ability** and **Instruction Following**.

5.1 Conversational Ability

Effective conversational ability in LALMs relies on generating contextually appropriate responses and smoothly managing dialogues in real time. Current evaluations address this via two complementary frameworks: affective and contextual interaction, and full-duplex dialogue management.

5.1.1 Affective and Contextual Interaction

Evaluations of affective and contextual interaction typically adopt half-duplex settings, focusing on fully turn-by-turn conversations without speaker overlaps. These benchmarks emphasize LALMs' ability to respond using both content and non-content cues such as emotional tone, speaking style, and speaker traits. StyleTalk (Lin et al., 2024a) presents models with a dialogue history and the user's current speech segment, intentionally leaving the user's intent underspecified when relying solely on the content. Consequently, models are required to leverage paralinguistic cues to respond appropriately. Subsequent works, such as SD-Eval (Ao et al., 2024) and VoxDialogue (Cheng et al., 2025), broaden the evaluation by incorporating more acoustic and contextual variables, including speaker age, accent, and environmental conditions. These benchmarks combine objective metrics (e.g., ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005)), LLM-based judgment (Gu et al., 2024), and human evaluation for comprehensive assessment.

While these benchmarks rely on static data, Li et al. (2025) proposes an interactive framework inspired by Chatbot Arena (Chiang et al., 2024), where real users converse with models on topics of their choice and provide pairwise model preferences, enabling dynamic, user-centered evaluation.

5.1.2 Full-duplex Dialogue Management

Full-duplex evaluation examines LALMs in real-time, dynamic dialogues with complex behaviors like turn-taking (Duncan, 1972; Gravano and Hirschberg, 2011), backchanneling (Schegloff, 1982), and speaker interruptions and overlaps (Gravano and Hirschberg, 2012; Schegloff, 2000). These behaviors are detailed in Appendix F.

Representative works, such as Talking Turns (Arora et al., 2025b) and Full-Duplex-Bench (Lin et al., 2025a), commonly evaluate four key dimensions:

- **Timing for speaking up or interrupting**: Assesses LALMs' ability to distinguish meaningful pauses from turn-yielding moments, avoiding undesired interruptions and taking over turns appropriately.
- **Backchanneling**: Evaluates whether LALMs backchannel at proper moments with suitable frequency, reflecting their active listening.
- **Turn taking**: Examines whether LALMs transition smoothly between turns by recognizing boundaries, managing latency, and signaling their intent to maintain or yield the floor.
- **User interruption handling**: Assesses LALMs' handling of interruption, e.g., pausing and smoothly resuming the conversation.

Both use automatic evaluation metrics. Talking Turns uses supervised models trained on human dialogues (Godfrey et al., 1992) as a reference, while Full-Duplex-Bench uses metrics like response latency. However, these methods often rely on heuristics, which may be inaccurate in some cases.

Their results show that LALMs struggle with full-duplex management, especially with interruptions (Arora et al., 2025b) and seamless turn transitions (Lin et al., 2025a), highlighting current limitations in dynamic spoken interaction.

5.2 Instruction Following

Instruction following is the ability to follow user-specified instructions, e.g., requirements for performing particular actions, adhering to constraints, and adjusting response styles. Effective instruction following is essential for model controllability.

LALM instruction-following evaluations typically involve three approaches: (1) adding constraints to existing LALM benchmarks not originally for instruction following, (2) synthesizing LLM instruction-following benchmarks into speech, or (3) creating new dedicated datasets. For instance, Speech-IFeval (Lu et al., 2025) introduces constraints into LALM benchmarks such as Dynamic-SUPERB Phase-2 (Huang et al., 2025a); VoiceBench (Chen et al., 2024c) synthesizes IFEval (Zhou et al., 2023a), a text-based LLM instruction-following benchmark, into speech; and URO-Bench (Yan et al., 2025) creates custom evaluation datasets.

Evaluating instruction adherence helps distinguish limitations in following instructions and deficiencies in auditory understanding or knowledge. Common evaluated constraints include length (e.g., a minimum number of words), format (e.g., responses in JSON or all caps), action (e.g., chain-of-thought reasoning (Wei et al., 2022)), style (e.g., responses in a humorous tone), and content (e.g., including a specific word). During evaluation, instruction-following rates, i.e., the frequency with which instructions are correctly followed, are measured with rule-based (Zhou et al., 2023a) or LLM-as-a-judge methods (Gu et al., 2024).

Benchmark results reveal significant gaps in LALMs compared to their LLM backbones in instruction following (Lu et al., 2025), indicating catastrophic forgetting when adapting LLMs to auditory modalities.

6 Fairness, Safety, and Trustworthiness

Despite the advancements of LALMs, their real-world deployment may pose social risks, such as perpetuating biases, generating harmful content, or spreading misinformation, if not properly evaluated and regulated. Therefore, fairness, safety, and trustworthiness must be thoroughly assessed. This section reviews works that quantify these risks to ensure the responsible and ethical use of LALMs.

6.1 Fairness and Bias

Fairness and bias are key ethical concerns for LALMs, ensuring they do not reinforce societal inequalities, discrimination, stereotypes, or biases. Such issues can be triggered by either the speech content or its non-content acoustic cues. For example, content-triggered bias may arise when LALMs translate occupation-related terms in the speech content into stereotypical gendered terms, independent of acoustic characteristics. In contrast, acoustic-triggered bias may arise when vocal cues lead the model to associate a speaker's gender with certain occupations.

Lin et al. (2024c) quantifies LALMs' content-triggered gender biases via four tasks: speech-to-text translation, coreference resolution, sentence continuation, and question answering. In each task, gender biases and stereotypes are measured based on the models' responses.

Conversely, Spoken Stereoset (Lin et al., 2024b)

assesses acoustic-triggered bias on speakers' gender and age. The authors sampled sentences from NLP datasets like Stereoset (Nadeem et al., 2021) and BBQ (Parrish et al., 2022), which were then rewritten in the first-person perspective with explicit gender or age indicators (e.g., "mother") removed to ensure bias would be triggered by speaker characteristics rather than content. The modified sentences were synthesized into speech using TTS with voices of different genders and ages. These spoken sentences served as the context, and LALMs were tasked with selecting continuations from options that were stereotypical, anti-stereotypical, or unrelated to the context.

These works highlight LALMs' social biases, which may be inherited from their training data or LLM backbones. Additionally, since social biases are multifaceted, current benchmarks cannot include all possible societal factors, emphasizing the need for further research into both model development and benchmarks to enhance fairness.

6.2 Safety

Unlike fairness and bias, which expose societal prejudices in LALMs, safety concerns focus on preventing harmful or unsafe outputs that may negatively impact individuals or society, including user discomfort or illegal activities. Current studies typically use NLP datasets with malicious queries and convert them into speech via TTS. For example, VoiceBench (Chen et al., 2024c) and Roh et al. (2025) synthesize datasets like AdvBench (Zou et al., 2023) into spoken queries, evaluating LALMs on their ability to reject them.

During evaluation, jailbreaking techniques may be employed to test models' resistance to adversarial inputs. These include modifying speech content by inserting fictional scenarios (Shen et al., 2024) and applying auditory manipulations such as silence (Yang et al., 2025b), noise (Yang et al., 2025b; Xiao et al., 2025), accents (Roh et al., 2025; Xiao et al., 2025), and audio edits (Xiao et al., 2025; Gupta et al., 2025). Ideally, LALMs should remain robust to adversarially modified inputs and consistently reject malicious requests.

However, evaluations show that LALMs often accept malicious spoken inputs even when they can refuse similar textual ones (Chen et al., 2024c). Moreover, LALMs show considerable safety degradation compared to their LLM backbones (Yang et al., 2025b). Several jailbreaking methods can easily bypass these models (Roh et al., 2025; Xiao et al., 2025), highlighting the need for better multimodal safety alignment.

6.3 Hallucination

Hallucination occurs when a model generates non-factual or unsupported outputs, reducing reliability and misleading users. In LALMs, hallucinations can originate from both auditory and textual modalities. While textual hallucinations can be assessed with NLP benchmarks (Li et al., 2023; Chen et al., 2024a; Bang et al., 2025), we focus on auditory-induced hallucinations.

Kuan et al. (2024a) explores LALMs' object hallucination, where the models falsely identify objects or events absent from the auditory input. They evaluate this via two tasks: a discriminative task where LALMs determine whether a specified object exists in the audio, and a generative task where LALMs generate captions describing the audio. These captions are then evaluated for accuracy in reflecting the actual content of the audio. Despite generating accurate captions, LALMs struggle with object identification in the discriminative task, revealing challenges in object hallucination for question-answering tasks.

Leng et al. (2024) further analyzes object hallucination using the CMM benchmark, showing that overrepresented objects or events in the training data can lead LALMs to incorrectly predict their presence, even when they are absent. Additionally, the frequent co-occurrence of objects and events during training exacerbates these hallucinations.

These works highlight hallucination challenges in LALMs and call for improved training, modeling, and data handling to enhance trustworthiness.

7 Challenges and Future Directions

7.1 Data Leakage and Contamination

Creating and curating high-quality auditory data is far more difficult than for text. Consequently, many LALM benchmarks rely on existing auditory corpora (Panayotov et al., 2015a; Kim et al., 2019; Gemmeke et al., 2017) rather than collecting new data. This raises concerns about data leakage, since models may have seen these datasets during training (Deng et al., 2024; Zhou et al., 2023b; Jacovi et al., 2023), undermining evaluation reliability. The risk grows when large-scale web-crawled data (Radford et al., 2023; He et al., 2024) are used for training without rigorous filtering.

Thus, alongside creating or collecting custom

data, developing methods to detect and mitigate contamination (Golchin and Surdeanu, 2024; Samuel et al., 2025) will be a crucial direction for more reliable LALM evaluations.

7.2 Inclusive Evaluation Across Linguistic, Cultural, and Communication Diversity

While current benchmarks cover major languages like English and Mandarin (Huang et al., 2025a; Yan et al., 2025), many overlook crucial aspects such as low-resource languages (Magueresse et al., 2020) and code-switching (Doğruöz et al., 2021; Sitaram et al., 2019). Although these have been explored in traditional speech technologies (Khare et al., 2021; Bhogale et al., 2024; Liu et al., 2024; Yang et al., 2024b), they remain underexamined in LALMs. This limited coverage fails to capture the full linguistic diversity of human communication, as different languages possess unique characteristics (Evans and Levinson, 2009; Bickel, 2014).

Cultural factors, shaped by historical and social contexts, influence dimensions like moral norms (Graham et al., 2016; Saucier, 2018) and are essential for evaluation. As LALMs extend to diverse cultures (Yang et al., 2024a; Wang et al., 2025b), evaluation frameworks must also expand.

Along with language and culture, communication patterns also matter. While some work covers speech variations like accents, underrepresented groups such as people with speech disorders (e.g., dysarthria (Kent et al., 1999; Kim et al., 2008)) are often overlooked, as current LALMs have limited familiarity with their unique speech patterns, which affects fair and accurate understanding.

To develop fair and broadly applicable LALMs, future evaluations should carefully consider linguistic, cultural, and communicative diversity.

7.3 Safety Evaluation Unique to Auditory Modalities

Current LALM safety evaluations (§6.2) mainly target harmful content in model outputs, often overlooking risks inherent to auditory modalities. Auditory cues such as tone, emotion, and voice quality can also influence user experience and raise concerns if uncontrolled. For instance, even harmless content can discomfort users if spoken harshly or sarcastically, and the presence of annoying noises can also cause irritation. Thus, safety should cover auditory comfort, not just content harmlessness.

Most benchmarks focus on content toxicity but seldom assess auditory-specific safety. Addressing these issues is vital for applications like voice assistants (Pias et al., 2024; Mari et al., 2024), where vocal manner greatly affects user trust and comfort. Future work should jointly consider vocal tone, noise, and other paralinguistic factors to ensure safe, user-friendly interactions.

7.4 Unified Evaluation of Harmlessness and Helpfulness

Harmlessness and helpfulness in LALMs refer to safety and fairness, and the ability to assist users, respectively. Ideally, these two properties should be enhanced together; however, in practice, they often conflict (Bai et al., 2022). For example, a model that always refuses to answer is safe but unhelpful, as it fails to assist users. A recent study (Lin et al., 2025b) shows that post-training aimed at enhancing harmlessness can reduce helpfulness, causing models to reject queries even when no safety or privacy issues exist. This tension highlights the need for a unified evaluation framework that considers both aspects simultaneously.

Existing harmlessness benchmarks (§6) rarely include helpfulness, limiting understanding of their trade-offs and offering limited guidance for balancing them effectively. Thus, developing a joint evaluation framework is a key future direction.

7.5 Personalization Evaluation

Personalization enables models to adapt to individual users by incorporating private information like users' voices and preferences, supporting applications such as personalized voice assistants.

While traditional speech technologies have explored personalization (Lee et al., 2024; Joseph and Baby, 2024), it remains underdeveloped for LALMs. Unlike recent progress in LLM personalization (Tan et al., 2024, 2025; Zhang et al., 2024), LALM personalization is more complex due to the auditory dimension: LALMs must adapt to user-specific knowledge, as text LLMs do, but also become familiar with users' voice characteristics and speaking habits, and adjust their own speaking style to match user preferences. Such complexity necessitates the development of specialized evaluations to fully assess LALM personalization, making it a valuable area for future investigation.

8 Conclusion

Holistic evaluation of LALMs is as crucial as modeling and training in advancing the field. This survey reviews existing evaluation frameworks and

proposes a taxonomy categorizing current progress into four important research areas, reflecting the diverse expectations of LALM capabilities. We present a thorough overview of the literature, highlighting challenges and future directions, such as data contamination, inclusivity, auditory-specific safety, and personalization. We hope this survey provides clear guidelines for researchers and stimulates further advancements in LALM evaluation.

Limitations

We acknowledge a few limitations in this paper. First, the scope of our taxonomy is based on existing evaluation frameworks and benchmarks, meaning it does not cover all possible real-world auditory tasks. The auditory modalities are inherently complex, with a wide range of tasks and applications that cannot be exhaustively covered. As LALMs continue to evolve, new capabilities and applications will emerge, leading to growing expectations for these models. Consequently, the evaluation landscape will likely expand and shift, requiring our taxonomy to be updated and adapted to include these new tasks and applications. We will continue to follow the advancements in this field and adjust our taxonomy accordingly to reflect these developments.

Second, this survey primarily focuses on current benchmarks used to evaluate LALMs' performance across various aspects. As a result, it does not put much emphasis on more basic or traditional evaluation methods, such as subjective assessments of speech generation quality (e.g., Mean Opinion Score), which are commonly used to evaluate model-generated audio. While these methods are valuable in certain applications, they fall outside the scope of this paper, which aims to provide a comprehensive overview of more advanced and specialized benchmarks.

Acknowledgments

We thank the reviewers for their valuable feedback, which have helped us improve the paper. To enhance the comprehensiveness of our survey, we have added further discussion in the appendix.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. SD-eval: A benchmark dataset for spoken dialogue understanding beyond words. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025a. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. 2025b. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. In *The Thirteenth International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Kaushal Santosh Bhogale, Deovrat Mehendale, Niharika Parasa, Sathish Kumar Reddy G, Tahir Javed, Pratyush Kumar, and Mitesh M. Khapra. 2024. Empowering low-resource language asr via large-scale pseudo labeling. In *Interspeech 2024*, pages 2519–2523.

Balthasar Bickel. 2014. Linguistic diversity and universals. *The Cambridge handbook of linguistic anthropology*, pages 101–124.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, Miami, Florida, USA. Association for Computational Linguistics.

Fan Bu, Yuhao Zhang, Xidong Wang, Benyou Wang, Qun Liu, and Haizhou Li. 2024. Roadmap towards superhuman speech understanding using large language models. *arXiv preprint arXiv:2410.13268*.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji, Yueru He, Jimin Huang, Zining Zhu, Qianqian Xie, Xiao-yang Liu, Koduvayur Subbalakshmi, and 1 others. 2025. Finaudio: A benchmark for audio large language models in financial applications. *arXiv preprint arXiv:2503.20990*.

Kedi Chen, Qin Chen, Jie Zhou, He Yishen, and Liang He. 2024a. DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9057–9079, Miami, Florida, USA. Association for Computational Linguistics.

Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D'Haro, Robby Tan, and Haizhou Li. 2024b. Beyond single-audio: Advancing multi-audio processing in audio large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10917–10930.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024c. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. 2025. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech 2018*, pages 1086–1090.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025. Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. *arXiv preprint arXiv:2501.04962*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. Fma: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference*.

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

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. 2025a. Audio entailment: Assessing deductive reasoning for audio understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23769–23777.

Soham Deshmukh, Shuo Han, Rita Singh, and Bhiksha Raj. 2025b. ADIFF: Explaining audio difference using natural language. In *The Thirteenth International Conference on Learning Representations*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

A Seza Doğruöz, Sunayana Sitaram, Barbara Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

Starkey Duncan and Donald W Fiske. 2015. *Face-to-face interaction: Research, methods, and theory*. Routledge.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with WaveNet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1068–1077. PMLR.

Nicholas Evans and Stephen C Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448.

Y. Fan, J.W. Kang, L.T. Li, K.C. Li, H.L. Chen, S.T. Cheng, P.Y. Zhang, Z.Y. Zhou, Y.Q. Cai, and D. Wang. 2020. Cn-celeb: A challenging chinese speaker recognition dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7604–7608.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. LLaMA-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*.

Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *INTERSPEECH*, pages 889–892.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2024. Benchmarking open-ended audio dialogue understanding for large audio-language models. *arXiv preprint arXiv:2412.05167*.

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

Felix Gervits and Matthias Scheutz. 2018. Towards a conversation-analytic taxonomy of speech overlap. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024b. Compa: Addressing the gap in compositional reasoning in audio-language models. In *The Twelfth International Conference on Learning Representations*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Julia A Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts. *Journal of pragmatics*, 14(6):883–903.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and 1 others. 2024a. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2024b. Listen, think, and understand. In *International Conference on Learning Representations*.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE.

Jesse Graham, Peter Meindl, Erica Beall, Kate M Johnson, and Li Zhang. 2016. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8:125–130.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Agustín Gravano and Julia Hirschberg. 2012. A corpus-based study of interruptions in spoken dialogue. In *Interspeech 2012*, pages 855–858.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Isha Gupta, David Khachaturov, and Robert Mullins. 2025. " i am bad": Interpreting stealthy, universal and robust audio jailbreaks in audio-language models. *arXiv preprint arXiv:2502.00718*.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, and 1 others. 2023. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36:63483–63501.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Chien-yu Huang, Wei-Chih Chen, Shu wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, and 57 others. 2025a. Dynamic-SUPERB phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations*.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-Yi Lee. 2024a. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140.

Chien-yu Huang, Min-Han Shih, Ke-Han Lu, Chi-Yuan Hsiao, and Hung-yi Lee. 2025b. Speechcaps: Advancing instruction-based universal speech models with multi-talker speaking style captioning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kuan-Po Huang, Chih-Kai Yang, Yu-Kuan Fu, Ewan Dunbar, and Hung-Yi Lee. 2024b. Zero resource code-switched speech benchmark using speech utterance pairs for multiple spoken languages. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10006–10010.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024c. Audiogpt: Understanding and generating

speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. 2024. Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084.

Gail Jefferson. 1986. Notes on'latency'in overlap onset. *Human studies*, pages 153–183.

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

George Joseph and Arun Baby. 2024. Speaker personalization for automatic speech recognition using weight-decomposed low-rank adaptation. In *Proc. Interspeech 2024*, pages 2875–2879.

Mintong Kang, Chejian Xu, and Bo Li. 2025. Advwave: Stealthy adversarial jailbreak attack against large audio-language models. In *The Thirteenth International Conference on Learning Representations*.

Ray D Kent, Gary Weismer, Jane F Kent, Houri K Vorperian, and Joseph R Duffy. 1999. Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of communication disorders*, 32(3):141–186.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Proc. Interspeech 2021*, pages 1529–1533.

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

Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon R Gunderson, Thomas S Huang, Kenneth L Watkin, Simone Frame, and 1 others. 2008. Dysarthric speech database for universal access research. In *Interspeech*, volume 2008, pages 1741–1744.

Heeseung Kim, Che Hyun Lee, Sangkwon Park, Jiheum Yeom, Nohil Park, Sangwon Yu, and Sungroh Yoon. 2025. Does your voice assistant remember? analyzing conversational context recall and utilization in voice interaction models. *arXiv preprint arXiv:2502.19759*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. 2024a. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. In *Interspeech 2024*, pages 4144–4148.

Chun-Yi Kuan and Hung-yi Lee. 2025. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chun-Yi Kuan, Chih-Kai Yang, Wei-Ping Huang, Ke-Han Lu, and Hung-yi Lee. 2024b. Speech-copilot: Leveraging large language models for speech processing via task decomposition, modularization, and program generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1060–1067. IEEE.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models. In *Proc. Interspeech 2023*, pages 4588–4592.

Chae-Won Lee, Jae-Hong Lee, and Joon-Hyuk Chang. 2024. Language model personalization for speech recognition: A clustered federated learning approach with adaptive weight average. *IEEE Signal Processing Letters*, 31:2710–2714.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of multi-modalities: Evaluating hallucinations of large

multimodal models across language, visual, and audio. *arXiv*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Minzhi Li, William Barr Held, Michael J Ryan, Kunat Pipatanakul, Potsawee Manakul, Hao Zhu, and Diyi Yang. 2025. Mind the gap! static and interactive evaluations of large audio models. *arXiv preprint arXiv:2502.15919*.

Mohan Li, Cong-Thanh Do, Simon Keizer, Youmna Farag, Svetlana Stoyanchev, and Rama Doddipatla. 2024. Whisma: A speech-llm to perform zero-shot spoken language understanding. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1115–1122. IEEE.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao MA, Chenghua Lin, Xingran Chen, Anton Ragni, Hanzhi Yin, Zhijie Hu, Haoyu He, and 1 others. 2022. Mapmusic2vec: A simple and effective baseline for self-supervised music audio representation learning. In *Ismir 2022 Hybrid Conference*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-Yi Lee. 2024a. 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)*, pages 6626–6642.

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

Yi-Cheng Lin, Wei-Chih Chen, and Hung-yi Lee. 2024b. Spoken stereoset: on evaluating social bias toward speaker in speech large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 871–878. IEEE.

Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung-yi Lee. 2024c. Listen and speak fairly: a study on semantic gender bias in speech integrated large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 439–446. IEEE.

Yu-Xiang Lin, Chih-Kai Yang, Wei-Chih Chen, Chen-An Li, Chien-yu Huang, Xuanjun Chen, and Hung-yi Lee. 2025b. A preliminary exploration with gpt-4o voice mode. *arXiv preprint arXiv:2502.09940*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

Hexin Liu, Leibny Paola Garcia, Xiangyu Zhang, Andy WH Khong, and Sanjeev Khudanpur. 2024. Enhancing code-switching speech recognition with interactive language biases. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10886–10890. IEEE.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024. Desta: Enhancing speech language models through descriptive speech-text alignment. In *Proc. Interspeech 2024*, pages 4159–4163.

Ke-Han Lu, Chun-Yi Kuan, and Hung-yi Lee. 2025. Speech-ifeval: Evaluating instruction-following and quantifying catastrophic forgetting in speech-aware language models. *Interspeech 2025*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Gallil Maimon, Amit Roth, and Yossi Adi. 2025. Salmon: A suite for acoustic language model evaluation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, and 1 others. 2023. The song describer dataset: a corpus of audio captions for music-and-language evaluation. In *Workshop on Machine Learning for Audio, Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems.

Alex Mari, Andreina Mandelli, and René Algesheimer. 2024. Empathic voice assistants: Enhancing consumer responses in voice commerce. *Journal of Business Research*, 175:114566.

Jan Melechovsky, Abhinaba Roy, and Dorien Herremans. 2024. Midicaps: A large-scale midi dataset with text captions. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, pages 858–865. ISMIR.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James F Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

James D Orcutt and Lynn Kenneth Harvey. 1985. Deviance, rule-breaking and male dominance in conversation. *Symbolic Interaction*, 8(1):15–32.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015a. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015b. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Prabhat Pandey, Rupak Vignesh Swaminathan, KV Girish, Arunasish Sen, Jian Xie, Grant P Strimel, and Andreas Schwarz. 2025. Sift-50m: A large-scale multilingual dataset for speech instruction fine-tuning. *arXiv preprint arXiv:2504.09081*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Abhirama Subramanyam Penamakuri, Kiran Chhatre, and Akshat Jain. 2025. Audiopedia: Audio qa with knowledge. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2024. A survey on speech large language models. *arXiv preprint arXiv:2410.18908*.

Sabid Bin Habib Pias, Alicia Freel, Ran Huang, Donald Williamson, Minjeong Kim, and Apu Kapadia. 2024. Building trust through voice: How vocal tone impacts user perception of attractiveness of voice assistants. *arXiv preprint arXiv:2409.18941*.

Karol J. Piczak. 2015a. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Karol J Piczak. 2015b. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Proc. Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. 2025. NatureLM-audio: an audio-language foundation model for bioacoustics. In *The Thirteenth International Conference on Learning Representations*.

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

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 4521–4525.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Vinay Samuel, Yue Zhou, and Henry Peng Zou. 2025. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5058–5070.

Gerard Saucier. 2018. Culture, morality and individual differences: comparability and incomparability across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1744):20170170.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71(93).

Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63.

Maureen Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2024. Emphassess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 495–507.

Xinyue Shen, Yixin Wu, Michael Backes, and Yang Zhang. 2024. Voice jailbreak attacks against gpt-4o. *arXiv preprint arXiv:2405.19103*.

Jack Sidnell. 2007. Comparative studies in conversation analysis. *Annu. Rev. Anthropol.*, 36(1):229–244.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, and 1 others. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. *arXiv preprint arXiv:2502.20616*.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6459–6475.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, and 1 others. 2022. Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR.

George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, and 1 others. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 6:15.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2025a. Audiobench: A universal benchmark for audio large language models. *NAACL*.

Bin Wang, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F Chen, and AiTi Aw. 2025b. Advancing singlish understanding: Bridging the gap with datasets and multimodal models. *arXiv preprint arXiv:2501.01034*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech 2021*, pages 2247–2251.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025c. Qualispeech: A speech quality assessment dataset with natural language reasoning and descriptions. *arXiv preprint arXiv:2503.20290*.

Yingzhi Wang, Pooneh Mousavi, Artem Ploujnikov, and Mirco Ravanelli. 2025d. What are they doing? joint audio-speech co-reasoning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, György Fazekas, and Dmitry Bogdanov.

2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Victor Junqiu Wei, Weicheng Wang, Di Jiang, Yuanfeng Song, and Lu Wang. 2024. Asr-ec benchmark: Evaluating large language models on chinese asr error correction. *arXiv preprint arXiv:2412.03075*.

Candace West. 1979. Against our will: Male interruptions of females in cross-sex conversation. *Annals of the New York Academy of Sciences*.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee. 2024a. Towards audio language modeling–an overview. *arXiv preprint arXiv:2402.13236*.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark Hasegawa-Johnson, and Mari Ostendorf. 2024b. Just asr+ llm? a study on speech large language models' ability to identify and understand speaker in spoken dialogue. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1137–1143. IEEE.

Erjia Xiao, Hao Cheng, Jing Shao, Jinhao Duan, Kaidi Xu, Le Yang, Jindong Gu, and Renjing Xu. 2025. Tune in, act up: Exploring the impact of audio modality-specific edits on large audio language models in jailbreak. *arXiv preprint arXiv:2501.13772*.

Liumeng Xue, Ziya Zhou, Jiahao Pan, Zixuan Li, Shuai Fan, Yinghao Ma, Sitong Cheng, Dongchao Yang, Haohan Guo, Yujia Xiao, and 1 others. 2025. Audio-flan: A preliminary release. *arXiv preprint arXiv:2502.16584*.

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

Chih-Kai Yang, Yu-Kuan Fu, Chen-An Li, Yi-Cheng Lin, Yu-Xiang Lin, Wei-Chih Chen, Ho Lam Chung, Chun-Yi Kuan, Wei-Ping Huang, Ke-Han Lu, and 1 others. 2024a. Building a taiwanese mandarin spoken language model: A first attempt. *arXiv preprint arXiv:2411.07111*.

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

Chih-Kai Yang, Kuan-Po Huang, Ke-Han Lu, Chun-Yi Kuan, Chi-Yuan Hsiao, and Hung-Yi Lee. 2024b. Investigating zero-shot generalizability on mandarin-english code-switched asr and speech-to-text translation of recent foundation models with self-supervision and weak supervision. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 540–544.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. 2025b. Audio is the achilles' heel: Red teaming audio large multimodal models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9292–9306, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. Superb: Speech processing universal performance benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024d. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. 2024e. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? *arXiv preprint arXiv:2411.16679*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, and 1 others. 2023. Marble: Music audio representation benchmark for universal evaluation. *Advances in Neural Information Processing Systems*, 36:39626–39647.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. 2025. Are you really listening? boosting perceptual awareness in music-qa benchmarks. *arXiv preprint arXiv:2504.00369*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Proc. Interspeech 2019*, pages 1526–1530.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Mengjie Zhao, Zhi Zhong, Zhuoyuan Mao, Shiqi Yang, Wei-Hsiang Liao, Shusuke Takahashi, Hiromi Wakaki, and Yuki Mitsufuji. 2024. Openmu: Your swiss army knife for music understanding. *arXiv preprint arXiv:2410.15573*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023a. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023b. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

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

A Detailed Categorization of the Surveyed Papers

The complete categorization of the surveyed papers, based on the proposed taxonomy (§2), is presented in Figure 3. Please note that widely used corpora for fundamental auditory processing tasks, such as speech recognition and audio captioning, are excluded from this categorization due to the extremely large number of such resources. Including them would make the figure overly detailed and cumbersome. For reference, we provide examples of these fundamental tasks and their corresponding resources in Appendix C.

From Figure 3, it is evident that the current focus of LALM evaluations predominantly centers on auditory processing tasks (§3.2), underscoring their importance to the community. While these tasks are valuable, they should not be seen as the sole consideration when evaluating models for real-world applications. A more diverse and comprehensive evaluation scope is crucial to ensure a fuller understanding of their potential and shortcomings.

Auditory Tasks	Common Datasets
Audio Tasks	
Audio Captioning	AudioCaps (Kim et al., 2019) Clotho (Drossos et al., 2020)
Audio Classification	ESC-50 (Piczak, 2015b) AudioSet (Gemmeke et al., 2017)
Vocal Sound Classification	VocalSound (Gong et al., 2022)
Speech Tasks	
Automatic Speech Recognition	LibriSpeech (Panayotov et al., 2015b) AISHELL-1 (Bu et al., 2017) Common Voice (Ardila et al., 2020)
Speaker Identification	VoxCeleb2 (Chung et al., 2018) CN-Celeb (Fan et al., 2020)
Text-to-Speech	LJSpeech (Ito and Johnson, 2017) VCTK (Veaux et al., 2017) LibriTTS (Zen et al., 2019)
Speech Emotion Recognition	IEMOCAP (Busso et al., 2008) CREMA-D (Cao et al., 2014)
Language Identification	VoxLingua107 (Valk and Alumäe, 2021) FLEURS (Conneau et al., 2023)
Speech Translation	CoVoST 2 (Wang et al., 2021) MuST-C (Di Gangi et al., 2019)
Speech Diarization	LibriMix (Cosentino et al., 2020)
Keyword Spotting	Speech Command (Warden, 2018)
Music Tasks	
Music Captioning Text-to-Music	MusicCaps (Agostinelli et al., 2023) Song Describer Dataset (Manco et al., 2023) MidiCaps (Melechovsky et al., 2024)
Music Transcription	MAESTRO (Hawthorne et al., 2019)
Instrument Classification	NSynth (Engel et al., 2017)
Genre Classification	FMA (Defferrard et al., 2017) GTZAN (Tzanetakis and Cook, 2002)

Table 1: Commonly used datasets for various auditory tasks. This overview covers key tasks in audio, speech, and music processing and the datasets that are widely adopted in academic and industrial research.

B Brief Summary for Benchmarks Discussed in the Main Text

In this section, we summarize the key features and evaluation metrics of the benchmarks presented in Figure 2, aiming to guide researchers in selecting benchmarks suitable for their own use cases. The details are provided in Table 5 and Table 6.

C Examples of General Auditory Processing Tasks and Resources

Table 1 lists representative auditory processing tasks and their associated resources. As foundational components of auditory processing, these

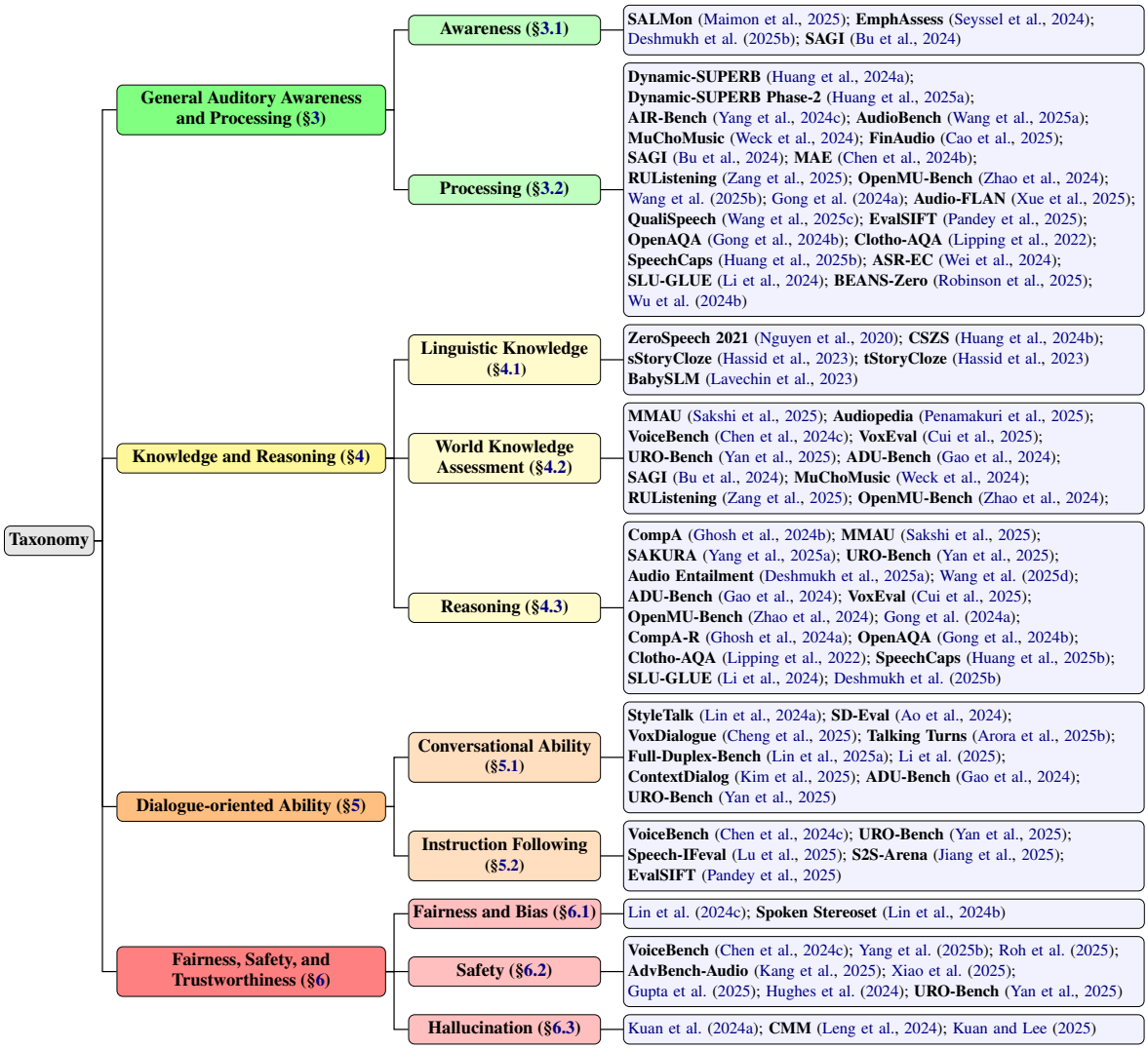

Figure 3: The complete categorization of the surveyed papers based on the proposed taxonomy.

tasks are well-suited for adaptation in LALM evaluation, as discussed in (§3.2).

D Overlap between Common Corpora and Existing Benchmarks

In Table 1, we list several corpora that are commonly used in the community. However, as discussed in (§7.1), evaluation benchmarks may face risks of data leakage and contamination if they heavily rely on these existing resources. To provide a quantitative view of this issue, we report detailed statistics on the number of benchmarks discussed in the main text (Figure 2) that make use of the datasets in Table 1.

From Table 2, we observe that certain corpora are used with particularly high frequency, which reinforces concerns about data contamination, especially when web-scraped data are incorporated into training without rigorous filtering. We emphasize that this issue should be taken seriously.

E Utilization of Synthetic Auditory Data in Benchmarks in the Main Text

Table 3 summarizes whether the benchmarks in Figure 2 use real or synthetic auditory data. Synthetic audio is mainly employed to (1) generate data difficult to obtain in real settings, such as controlled stress or specific sound effects, and (2) verbalize task instructions or dialogues consistently.

F Dynamics in Full-Duplex Dialogues

In this section, we briefly introduce the dynamics discussed in (§5.1.2). **Turn-taking** (Sacks et al., 1974) is a fundamental aspect of conversational organization, where speakers alternate turns to speak, ensuring only one person talks at a time. This process is complex, involving various behaviors that help facilitate smooth transitions between speakers.

Dataset	# of Benchmarks Using the Dataset
AudioCaps	**6**
Clotho	**5**
ESC-50	3
AudioSet	**8**
VocalSound	2
LibriSpeech	**7**
Common Voice	**8**
VoxCeleb (1&2)	3
LJSpeech	3
VCTK	4
LibriTTS	2
IEMOCAP	4
CREMA-D	2
VoxLingua107	1
CoVoST 2	2
LibriMix	1
Speech Command	2
MusicCaps	4
Song Describer Dataset	3
MAESTRO	1
NSynth	2
FMA	2

Table 2: Number of benchmarks in Figure 2 that use the datasets in Table 1. Datasets used by more than five benchmarks are highlighted in bold, while those not used by any benchmark are omitted.

For example, speakers often signal the end of their turn through clear cues, allowing the listener to recognize when they are yielding the floor (Duncan, 1972; Duncan and Fiske, 2015). Furthermore, turn-taking conventions may be shaped by cultural factors (Sidnell, 2007), which influence how and when speakers take their turns due to linguistic and social differences. Understanding and modeling these behaviors are essential steps toward achieving natural and effective communication in both human-human and human-AI interactions.

Backchanneling involves the listener's use of phatic expressions that signal active listening and attentiveness to the speaker (Fujie et al., 2005). These verbal cues, such as "yeah," "I see," or "uh-huh," along with non-verbal cues like nodding, serve as feedback, showing sympathy, agreement, or understanding. By offering such responses, listeners help maintain the flow of conversation without interrupting the speaker. This behavior not only fosters a sense of connection but also enhances the speaker's feeling of being heard and understood, contributing to a more interactive and supportive dialogue. As such, backchanneling plays a crucial role in sustaining conversation dynamics and promoting positive communicative exchanges.

Speaker overlap refers to the simultaneous speech of multiple speakers, while **speaker interruption** occurs when one speaker interjects during another's turn, which breaks the turn-taking principles (Gravano and Hirschberg, 2012). These phenomena are complex: they can be competitive, reflecting hostility or dominance (West, 1979; Orcutt and Harvey, 1985), or they can be neutral or supportive, helping to maintain and coordinate the flow of dialogue (Goldberg, 1990; Jefferson, 1986; Gervits and Scheutz, 2018). Despite their varying forms, both overlap and interruption are natural components of human conversation.

G Input/Output Modalities of the Surveyed Works

Our proposed taxonomy (§2) is organized by the evaluation objectives of the surveyed works rather than by the modalities they cover. Nevertheless, modality information is essential for researchers seeking benchmarks suited to models specialized in particular modalities. Thus, we provide the input/output modality details in Tables 7, 8, 9, and 10, corresponding to the categories of General Auditory Awareness and Processing (§3), Knowledge and Reasoning (§4), Dialogue-oriented Ability (§5), and Fairness, Safety, and Trustworthiness (§6), respectively. These tables are compiled based on the original papers of the surveyed works.

Please note that due to unique evaluation designs, some benchmarks do not produce explicit "outputs" but instead rely on input likelihood comparisons or similarity measures with specific instances. This absence of outputs is clearly indicated in the tables.

H Comparison between End-to-end LALMs and Cascaded Systems

Table 4 presents a comparison between end-to-end (E2E) LALMs and cascaded systems that couple LLMs with modules such as speech recognition, evaluated across the benchmarks in Figure 2. Cascaded systems generally perform better on benchmarks emphasizing content- or semantics-based tasks, which require only limited integration of non-semantic auditory cues. These tasks align well with the strengths of cascaded pipelines. For instance, VoiceBench (Chen et al., 2024c) and VoxEval (Cui et al., 2025) primarily assess world knowledge and

reasoning abilities adapted from text-based datasets (e.g., MMLU (Hendrycks et al., 2021)), where textual representations alone are sufficient, thus giving cascaded approaches a clear advantage.

Conversely, certain benchmarks highlight the advantages of E2E LALMs. For example, SALMon (Maimon et al., 2025) requires fine-grained auditory perception that cascaded systems struggle to replicate. This suggests the importance of future benchmarks that place greater emphasis on nuanced auditory understanding and reasoning, where LALMs hold stronger potential to excel.

I Common Metrics for LALMs

As previously noted, this paper primarily focuses on benchmarks for evaluating LALMs across diverse aspects, rather than on basic or traditional evaluation metrics. In this section, we briefly introduce some of the most commonly used metrics.

To assess the generation quality and naturalness of auditory outputs, the most widely adopted measure is the Mean Opinion Score (MOS), which relies on human annotators' judgments to provide subjective quality assessments. While effective, MOS collection can be costly and time-consuming. To mitigate this, proxy models such as UTMOS (Saeki et al., 2022) have been proposed to automatically predict MOS scores, offering a more efficient alternative to direct human evaluation.

For evaluating the content of generated auditory outputs, several additional metrics are commonly used. In cases where models can produce both speech and text, Character Error Rate (CER) and Word Error Rate (WER) measure the consistency between generated textual outputs and transcriptions of generated speech (via an ASR system), thereby quantifying alignment across modalities.

Beyond surface-level alignment, other metrics assess the semantic quality of responses. Widely used measures include ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), which evaluate semantic overlap between model outputs and references. More recently, LLM-as-a-judge (Gu et al., 2024) has been increasingly adopted to provide flexible, criterion-driven evaluations tailored to researchers' specific needs.

J Information on AI Assistance

We acknowledge the assistance of GPT-4.1-mini in refining the paper and improving its clarity.

Benchmark	Real Data	Synthetic Data
SALMon (Maimon et al., 2025)	✓	✓
EmphAssess (Seyssel et al., 2024)		✓
Dynamic-SUPERB (Huang et al., 2024a)	✓	✓
Dynamic-SUPERB Phase-2 (Huang et al., 2025a)	✓	✓
AIR-Bench (Yang et al., 2024c)	✓	✓
AudioBench (Wang et al., 2025a)	✓	✓
MuChoMusic (Weck et al., 2024)	✓	
ZeroSpeech 2021 (Nguyen et al., 2020)	✓	✓
CSZS (Huang et al., 2024b)		✓
sStoryCloze & tStoryCloze (Mostafazadeh et al., 2017)		✓
MMAU (Sakshi et al., 2025)	✓	✓
Audiopedia (Penamakuri et al., 2025)		✓
VoiceBench (Chen et al., 2024c)	✓	✓
VoxEval (Cui et al., 2025)		✓
CompA (Ghosh et al., 2024b)	✓	✓
SAKURA (Yang et al., 2025a)	✓	
URO-Bench (Yan et al., 2025)	✓	✓
Audio Entailment (Deshmukh et al., 2025a)	✓	
Wang et al. (2025d)	✓	✓
StyleTalk (Lin et al., 2024a)		✓
SD-Eval (Ao et al., 2024)	✓	✓
VoxDialogue (Cheng et al., 2025)		✓
Talking Turns (Arora et al., 2025b)	✓	
Full-Duplex-Bench (Lin et al., 2025a)	✓	✓
Speech-IFEval (Lu et al., 2025)	✓	
Lin et al. (2024c)	✓	✓
Spoken Stereoset (Lin et al., 2024b)		✓
Yang et al. (2025b)		✓
Roh et al. (2025)		✓
Kuan et al. (2024a)	✓	
CMM (Leng et al., 2024)	✓	

Table 3: Statistics of the utilization of real/synthetic auditory data in benchmarks in Figure 2.

Type	Benchmark List
Cascaded Systems Outperform	Dynamic-SUPERB (Huang et al., 2024a) MMAU (Sakshi et al., 2025) VoiceBench (Chen et al., 2024c) VoxEval (Cui et al., 2025) SAKURA (Yang et al., 2025a) SD-Eval (Ao et al., 2024) Speech-IFEval (Lu et al., 2025) Yang et al. (2025b)
E2E LALMs Outperform	SALMon (Maimon et al., 2025) Dynamic-SUPERB Phase-2 (Huang et al., 2025a)[1] AIR-Bench (Yang et al., 2024c)[2] StyleTalk (Lin et al., 2024a)[3]

[1] Overall, the top-performing E2E LALMs outperform cascaded systems, except for a small number of tasks where their performance is slightly inferior.
[2] The top-performing E2E LALMs significantly outperform cascaded systems on the foundation benchmark, with only a slight drop in performance on the chat benchmark.
[3] The top-performing E2E LALMs significantly outperform cascaded systems across various metrics, except for one where they slightly underperform.

Table 4: Summary of comparisons between end-to-end (E2E) LALMs and cascaded systems, limited to benchmarks in Figure 2 whose original papers explicitly conducted and reported such comparisons.

Benchmark	Features	Metrics
General Auditory Awareness and Processing		
SALMon (Maimon et al., 2025)	Challenging benchmark for nuanced auditory awareness of semantic, paralinguistic, and acoustic information.	Likelihood-based Comparison
EmphAssess (Seyssel et al., 2024)	Benchmark testing awareness of prosodic features, requiring models to preserve them during speech-to-speech translation.	Precision, Recall, F1
Dynamic-SUPERB (Huang et al., 2024a)	The first benchmark covering audio, speech, and music with 55 multiple-choice QA tasks.	Accuracy
Dynamic-SUPERB Phase-2 (Huang et al., 2025a)	An expanded version of Dynamic-SUPERB, currently the largest benchmark in the auditory processing category, featuring 180 tasks including open-ended ones.	Accuracy, TER, MSE, KTAU, LCC, SRCC, ERR, Miss Time, WER, Sacre BLEU, MER, IoU, F1, MEDAE, Angle Diff, Abs Diff, PER, PCC, DER, CER, POS
AIR-Bench (Yang et al., 2024c)	Benchmark covering audio, speech, and music with both multiple-choice and open-ended questions.	Accuracy, LLM-as-a-judge
AudioBench (Wang et al., 2025a)	Benchmark focusing on speech and audio, comprising 8 tasks and 26 datasets.	Word Error Rate, LLM-as-a-judge, METEOR
MuChoMusic (Weck et al., 2024)	In-depth investigation of music-oriented knowledge and processing abilities.	Accuracy, Instruction-following rate
Knowledge and Reasoning		
ZeroSpeech 2021 (Nguyen et al., 2020)	Evaluating linguistic knowledge (lexical and syntactic understanding), primarily in English.	Likelihood-based Comparison, Similarity Comparison
CSZS (Huang et al., 2024b)	Assessing semantic and syntactic knowledge in code-switching scenarios (English-Mandarin, English-French, English-Spanish).	Likelihood-based Comparison
sStoryCloze & tStoryCloze (Mostafazadeh et al., 2017)	Benchmarks for semantic coherence in spoken story continuation. Focused on English.	Likelihood-based Comparison
MMAU (Sakshi et al., 2025)	Expert knowledge and advanced reasoning (e.g., temporal, causal reasoning across audio, speech, music).	Accuracy
Audiopedia (Penamakuri et al., 2025)	Knowledge-intensive QA benchmark for evaluating world knowledge of entities.	Accuracy, F1
VoiceBench (Chen et al., 2024c)	Comprehensive benchmark for world knowledge, instruction following, safety, and robustness to acoustic variations.	LLM-as-a-judge, Accuracy, Refusal Rate
VoxEval (Cui et al., 2025)	Spoken version of MMLU considering speaker, style, and audio quality variations.	Accuracy
CompA (Ghosh et al., 2024b)	Compositional reasoning of temporal order and attribute binding of sound events and captions.	Self-defined Text, Audio, and Group Scores
SAKURA (Yang et al., 2025a)	Multi-hop reasoning benchmark integrating knowledge and auditory information (gender, emotion, language, animal sounds).	Accuracy
URO-Bench (Yan et al., 2025)	Benchmark of knowledge, safety, and instruction-following (English, Chinese, code-switching).	Word Error Rate, Character Error Rate, LLM-as-a-judge, UTMOS, Latency
Audio Entailment (Deshmukh et al., 2025a)	Deductive reasoning from auditory information.	Accuracy, Precision, Recall, F1
Wang et al. (2025d)	Joint reasoning integrating acoustic and speech information.	LLM-as-a-judge

Table 5: Summary of the features and metrics of benchmarks in the **General Auditory Awareness and Processing** and the **Knowledge and Reasoning** categories in Figure 2.

Benchmark	Features	Metrics
Dialogue-oriented Ability		
StyleTalk (Lin et al., 2024a)	Earliest benchmark for affective and contextual conversational abilities of LALMs. Primarily focuses on speaker emotion and speaking styles.	BLEU, ROUGE, METEOR, BERT Score, F1, Human Evaluation (with A/B test)
SD-Eval (Ao et al., 2024)	Benchmark for conversational abilities of LALMs, covering speaker emotion, accent, age, and environmental sounds.	LLM-as-a-judge, ROUGE-L, BLEU, METEOR, BERT Score, Human Evaluation
VoxDialogue (Cheng et al., 2025)	Benchmark for conversational abilities of LALMs, further expanding the scope to 12 auditory attributes.	BLEU, ROUGE-L, METEOR, BERT Score, F1, LLM-as-a-judge
Talking Turns (Arora et al., 2025b)	Benchmark for evaluating the turn-taking dynamics of LALMs in full-duplex dialogues.	Automatic judgment with an internally trained judge model
Full-Duplex-Bench (Lin et al., 2025a)	Benchmark for pause handling, backchanneling, turn-taking, and interruption management in full-duplex dialogues.	Takeover Rate, Backchannel Frequency, Jensen-Shannon Divergence, Latency, LLM-as-a-judge
Speech-IFEval (Lu et al., 2025)	Benchmark specifically tailored for instruction following, disentangling instruction-following from speech perception. Can also analyze the degree of catastrophic forgetting in LALMs compared with their LLM backbones.	Word Error Rate, Accuracy, LLM-as-a-judge, Instruction-following Rate
Fairness, Safety, and Trustworthiness		
Lin et al. (2024c)	Quantifying the gender bias of LALMs through four tasks.	Accuracy, F1, F1 Differences, Language Modeling Score, Stereotypical Score, Idealized Context Association Tests Score, Instruction Following Rate, Bias Score
Spoken Stereoset (Lin et al., 2024b)	Evaluating social bias in LALMs, including gender and age.	Speech Language Instruction Following Score, Speech Language Modeling Score, Speech Language Bias Score
Yang et al. (2025b)	Quantifying safety issues of LALMs under jailbreaking attempts.	Attack Success Rate by Attempt, Attack Success Rate by Questions
Roh et al. (2025)	Evaluating safety alignment of LALMs under adversarial multilingual and multi-accent audio jailbreaks.	Jailbreak Success Rate, Word Error Rate, Accuracy
Kuan et al. (2024a)	Benchmark for object hallucination in LALMs, covering both discriminative and generative tasks.	Accuracy, Precision, Recall, F1, Word Error Rate, Self-defined Metrics
CMM (Leng et al., 2024)	Benchmark for quantifying object hallucination and identifying its potential causes. Includes vision modality as well.	Perception Accuracy, Hallucination Resistance

Table 6: Summary of the features and metrics of benchmarks in the **Dialogue-oriented Ability** and the **Fairness, Safety, and Trustworthiness** categories in Figure 2.

General Auditory Awareness and Processing								
Benchmark	**Input Modalities**				**Output Modalities**			
	Text	Audio	Speech	Music	Text	Audio	Speech	Music
SALMon (Maimon et al., 2025)		✓	✓		Likelihood-based evaluation. No output modality.			
Wu et al. (2024b)	✓		✓		✓			
EmphAssess (Seyssel et al., 2024)			✓				✓	
Deshmukh et al. (2025b)	✓	✓	✓		✓			
Dynamic-SUPERB (Huang et al., 2024a)	✓	✓	✓	✓	✓			
Dynamic-SUPERB Phase-2 (Huang et al., 2025a)	✓	✓	✓	✓	✓			
AIR-Bench (Yang et al., 2024c)	✓	✓	✓	✓	✓			
AudioBench (Wang et al., 2025a)	✓	✓	✓		✓			
MuChoMusic (Weck et al., 2024)	✓			✓	✓			
FinAudio (Cao et al., 2025)	✓		✓		✓			
SAGI (Bu et al., 2024)	✓	✓	✓	✓	✓			
MAE (Chen et al., 2024b)	✓	✓	✓		✓			
RUListening (Zang et al., 2025)	✓			✓	✓			
OpenMU-Bench (Zhao et al., 2024)	✓			✓	✓			
Wang et al. (2025b)	✓		✓		✓			
Gong et al. (2024a)	✓	✓	✓	✓	✓			
Audio-FLAN (Xue et al., 2025)	✓	✓	✓	✓	✓	✓	✓	✓
QualiSpeech (Wang et al., 2025c)	✓		✓		✓			
EvalSIFT (Pandey et al., 2025)	✓		✓		✓		✓	
OpenAQA (Gong et al., 2024b)	✓	✓			✓			
Clotho-AQA (Lipping et al., 2022)	✓	✓			✓			
SpeechCaps (Huang et al., 2025b)	✓		✓		✓			
ASR-EC (Wei et al., 2024)	✓		✓		✓			
SLU-GLUE (Li et al., 2024)	✓		✓		✓			
BEANS-Zero (Robinson et al., 2025)	✓			✓	✓			

Table 7: Input and output modalities of benchmarks in the **General Auditory Awareness and Processing** category shown in Figure 3.

Knowledge and Reasoning								
Benchmark	**Input Modalities**				**Output Modalities**			
	Text	Audio	Speech	Music	Text	Audio	Speech	Music
ZeroSpeech 2021 (Nguyen et al., 2020)			✓		Likelihood-based evaluation. No output modality.			
CSZS (Huang et al., 2024b)			✓		Likelihood-based evaluation. No output modality.			
sStoryCloze (Hassid et al., 2023)			✓		Likelihood-based evaluation. No output modality.			
tStoryCloze (Hassid et al., 2023)			✓		Likelihood-based evaluation. No output modality.			
BabySLM (Lavechin et al., 2023)	✓		✓		Likelihood-based evaluation. No output modality.			
CompA (Ghosh et al., 2024b)	✓	✓			Similarity-based evaluation on text and audio inputs.			
MMAU (Sakshi et al., 2025)	✓	✓	✓	✓	✓			
Audiopedia (Penamakuri et al., 2025)	✓		✓		✓			
VoiceBench (Chen et al., 2024c)	✓		✓		✓			
VoxEval (Cui et al., 2025)			✓				✓	
SAKURA (Yang et al., 2025a)	✓	✓	✓		✓			
URO-Bench (Yan et al., 2025)	✓		✓		✓		✓	
Audio Entailment (Deshmukh et al., 2025a)	✓	✓			✓			
ADU-Bench (Gao et al., 2024)			✓		✓		✓	
SAGI (Bu et al., 2024)	✓	✓	✓	✓	✓			
MuChoMusic (Weck et al., 2024)	✓			✓	✓			
RUListening (Zang et al., 2025)	✓			✓	✓			
OpenMU-Bench (Zhao et al., 2024)	✓			✓	✓			
Gong et al. (2024a)	✓	✓	✓	✓	✓			
CompA-R (Ghosh et al., 2024a)	✓	✓			✓			
OpenAQA (Gong et al., 2024b)	✓	✓			✓			
Clotho-AQA (Lipping et al., 2022)	✓	✓			✓			
SLU-GLUE (Li et al., 2024)	✓		✓		✓			
SpeechCaps (Huang et al., 2025b)	✓		✓		✓			
Wang et al. (2025d)	✓	✓	✓		✓			
Deshmukh et al. (2025b)	✓	✓	✓		✓			

Table 8: Input and output modalities of benchmarks in the **Knowledge and Reasoning** category shown in Figure 3.

Dialogue-oriented Ability								
Benchmark	**Input Modalities**				**Output Modalities**			
	Text	Audio	Speech	Music	Text	Audio	Speech	Music
StyleTalk (Lin et al., 2024a)	✓		✓				✓	
SD-Eval (Ao et al., 2024)	✓	✓	✓		✓			
VoxDialogue (Cheng et al., 2025)		✓	✓	✓	✓			
Talking Turns (Arora et al., 2025b)			✓				✓	
Full-Duplex-Bench (Lin et al., 2025a)			✓				✓	
Li et al. (2025)			✓		✓			
ContextDialog (Kim et al., 2025)			✓		✓		✓	
ADU-Bench (Gao et al., 2024)			✓		✓		✓	
VoiceBench (Chen et al., 2024c)	✓		✓		✓			
URO-Bench (Yan et al., 2025)	✓		✓		✓		✓	
Speech-IFeval (Lu et al., 2025)	✓		✓		✓			
S2S-Arena (Jiang et al., 2025)			✓				✓	
EvalSIFT (Pandey et al., 2025)	✓		✓		✓		✓	

Table 9: Input and output modalities of benchmarks in the **Dialogue-oriented Ability** category shown in Figure 3.

Fairness, Safety, and Trustworthiness								
Benchmark	**Input Modalities**				**Output Modalities**			
	Text	Audio	Speech	Music	Text	Audio	Speech	Music
Lin et al. (2024c)	✓		✓		✓			
Spoken Stereoset (Lin et al., 2024b)	✓		✓		✓			
VoiceBench (Chen et al., 2024c)	✓		✓		✓			
Yang et al. (2025b)	✓	✓	✓		✓			
Roh et al. (2025)	✓		✓		✓			
AdvBench-Audio (Kang et al., 2025)	✓		✓		✓			
Xiao et al. (2025)	✓		✓		✓			
Gupta et al. (2025)	✓		✓		✓			
Hughes et al. (2024)	✓		✓		✓			
URO-Bench (Yan et al., 2025)	✓		✓		✓		✓	
Kuan et al. (2024a)	✓	✓	✓		✓			
CMM (Leng et al., 2024)	✓	✓			✓			
Kuan and Lee (2025)	✓	✓			✓			

Table 10: Input and output modalities of benchmarks in the **Fairness, Safety, and Trustworthiness** category shown in Figure 3.