

A Survey of Audio Reasoning in Audio Foundation Models

Zhihan Guo^{1,*}, Wenqian Cui^{1,*}, Guan-Ting Lin², Daxin Tan¹, Jingyao Li¹,
Qiyong Zheng¹, Dingdong Wang¹, Jing Xiong³, Han Shi⁴, Jiaya Jia⁴, Irwin King¹

¹The Chinese University of Hong Kong ²National Taiwan University

³University of Hong Kong ⁴The Hong Kong University of Science and Technology
{zhguo22,wqcui23,king}@cse.cuhk.edu.hk

Abstract

Reasoning is now central to modern foundation models, but remains underdeveloped in the audio modality. Unlike text transcripts, audio contains not only words but also vocal cues, background events, and temporal patterns, all of which can be necessary for reliable inference. This survey presents the first dedicated synthesis of audio reasoning in multimodal foundation models. We formulate audio reasoning as reasoning-augmented generation grounded in acoustic evidence, distinguishing it from conventional audio understanding. We then organize existing work around two core settings: *Audio-to-Text reasoning*, where models infer textual answers from acoustic input, and *Audio-to-Speech reasoning*, where models reason before or during speech generation under latency constraints. Finally, we summarize evaluation practices and challenges in acoustic grounding, shortcut learning, modality hallucination, and latency-constrained interaction.

1 Introduction

Reasoning has become a central capability of modern foundation models (Wei et al., 2022; Li et al., 2025c). For instance, DeepSeek-R1 achieves superior performance on verifiable tasks such as mathematics and coding competitions (Guo et al., 2025a). In Large Language Models (LLMs), Chain-of-Thought (CoT) have substantially improved performance on complex tasks such as mathematics, coding, and logical problem solving (Snell et al., 2025; Wu et al., 2025d). In contrast, despite the rapid progress of Audio Foundation Models (AFMs), reliable audio reasoning remains far less understood (Arora et al., 2025; Sakshi et al., 2025).

Advancing audio reasoning is important for two reasons. First, audio is not merely an alternative carrier of textual content. It also encodes rich

paralinguistic and environmental signals, such as prosody, speaker state, emotion, overlap, temporal dynamics, and background events, that are often essential for reliable inference but are partially or entirely lost in text transcription (Ao et al., 2024; Guo et al., 2025b). Second, cascaded ASR-LLMs-TTS pipelines separate transcription, reasoning, and speech synthesis, often delaying responses and disrupting natural dialogue (Fang et al., 2025; Zhang et al., 2025c). This motivates audio reasoning models that operate directly on acoustic signals and generate timely spoken responses. Figure 1 highlights the rapid growth of audio reasoning studies, motivating a dedicated survey.

These gaps also expose a limitation of the existing survey literature. Recent surveys on audio large language models (Su et al., 2025), speech language models (Cui et al., 2025b; Arora et al., 2025), multimodal alignment (Wang et al., 2025e), real-time speech interaction (Wang et al., 2025d), and evaluation pipelines (Yang et al., 2025b; Peng et al., 2025) have provided valuable overviews of the broader audio foundation model landscape. However, they typically treat reasoning as an auxiliary capability rather than a central research problem, and therefore do not systematically examine how audio reasoning should be formulated, grounded, trained, and evaluated. Meanwhile, surveys on multimodal CoT reasoning discuss reasoning across modalities at a higher level (Latif et al., 2023), but do not focus on the audio-specific challenges of acoustic grounding, spoken response generation, and latency-aware interaction. As a result, the field still lacks a dedicated synthesis that centers audio reasoning itself. To the best of our knowledge, this is the first survey specifically devoted to audio reasoning.

However, incorporating reasoning into AFMs is not simply transferring CoT techniques from text or vision (Sprague et al., 2025; Gao et al., 2025). Audio reasoning raises modality-specific research questions about acoustic grounding, output modal-

*Equal contribution.

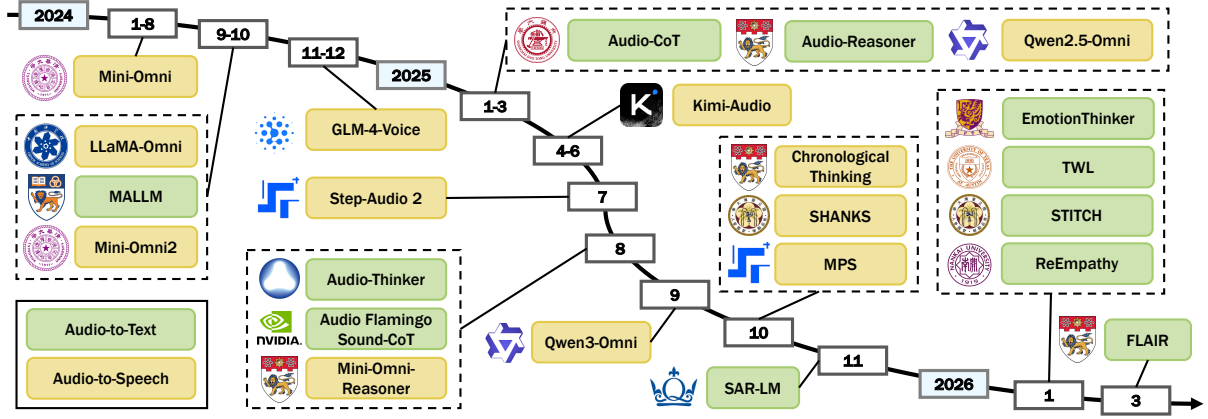


Figure 1: **Timeline of representative audio reasoning models.** Models are organized chronologically and grouped by major paradigms, including Audio-to-Text reasoning and Audio-to-Speech reasoning.

ity, and interaction latency. *One central question is how models can deeply understand and reason over acoustic information, including speech content, paralinguistic cues, background events, and temporal dynamics.* This motivates Audio-to-Text reasoning, where models infer textual answers from audio inputs while grounding their reasoning in acoustic evidence rather than relying only on transcripts (Section 3). *Another central question is how reasoning can support natural speech interaction.* This motivates Audio-to-Speech reasoning, where models must generate spoken responses while balancing reasoning depth with the latency constraints of real-time dialogue (Section 4). As shown in Figure 2, we first formalize these settings under a unified formulation of reasoning-augmented generation over acoustic evidence (Section 2), and then use this taxonomy to organize existing methods, evaluation practices, and open challenges (Sections 5 and 6). For completeness, Appendix A provides supplementary background on the architectural and foundations of audio reasoning models.

2 Notations and Problem Formulations

We introduce a compact formulation for the two settings in this survey: Audio-to-Text reasoning and Audio-to-Speech reasoning. Let $\mathbf{A} = \{\mathbf{a}_t\}_{t=1}^{T_a}$ denote the audio input, $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ denote the text instruction, $\mathbf{R} = \{\mathbf{r}_k\}_{k=1}^K$ denote the reasoning trajectory, $\mathbf{Y} = \{\mathbf{y}_m\}_{m=1}^M$ denote the text output, and $\mathbf{S} = \{\mathbf{s}_\ell\}_{\ell=1}^{L_s}$ denote the speech-token output. We use $\mathbf{O} \in \{\mathbf{Y}, \mathbf{S}\}$ as a generic output and $\mathbf{C} = (\mathbf{A}, \mathbf{X})$ as the audio-instruction context.

Audio understanding models often directly predict the output: $P(\mathbf{O} \mid \mathbf{C})$. In contrast, audio

reasoning introduces an intermediate trajectory:

$$P(\mathbf{R}, \mathbf{O} \mid \mathbf{C}) = P(\mathbf{R} \mid \mathbf{C})P(\mathbf{O} \mid \mathbf{C}, \mathbf{R}). \quad (1)$$

This distinction separates direct audio understanding from reasoning-augmented generation grounded in acoustic evidence.

3 Audio-to-Text Reasoning

Audio-to-text reasoning aims to infer textual answers from audio inputs through intermediate reasoning grounded in acoustic evidence, including speech content, paralinguistic cues, and environmental sounds. Given audio $\mathbf{A}$ and instruction $\mathbf{X}$, the model produces a textual response $\mathbf{Y}$. While direct audio understanding estimates $P(\mathbf{Y} \mid \mathbf{C})$, reasoning models introduce a trajectory $\mathbf{R}$:

$$P(\mathbf{R}, \mathbf{Y} \mid \mathbf{C}) = P(\mathbf{R} \mid \mathbf{C})P(\mathbf{Y} \mid \mathbf{C}, \mathbf{R}). \quad (2)$$

This formulation highlights the central requirement of audio-to-text reasoning: the final answer should be derived from the instruction and acoustic information, rather than from text priors alone.

3.1 Audio-to-Text Reasoning Methods

Existing methods mainly adapt CoT reasoning to Large Audio Language Models (LALMs) through three learning paradigms: inference-time prompting, supervised fine-tuning (SFT), and reinforcement learning (RL).

3.1.1 Inference-Time CoT Reasoning Methods

Inference-time CoT methods elicit intermediate reasoning during decoding without updating model parameters. Audio-CoT (Ma et al., 2025a) provides an early systematic study, comparing *Manual-CoT* with hand-written exemplars, *Zero-Shot-CoT*

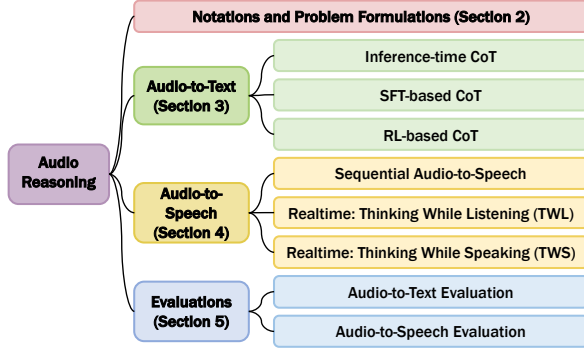


Figure 2: A compact taxonomy of audio reasoning.

with task-agnostic prompts such as “Let’s think step by step”, and *Desp-CoT*, which first generates an audio description before reasoning. The study shows that explicit reasoning can improve easy and medium-difficulty tasks, but may confuse models on harder problems. SAR-LM (Taheri et al., 2025) follows a related symbolic reasoning direction by converting complex audio signals into human-readable textual descriptions before downstream reasoning.

3.1.2 SFT-Based CoT Reasoning Methods

SFT-based CoT methods train LALMs on curated datasets that pair audio inputs with multi-step reasoning chains and final answers. Their effectiveness is largely determined by data construction, since models directly learn from the provided reasoning sequences. For example, Audio-Reasoner (Xie et al., 2025a) constructs CoTA, a large-scale dataset whose chains include planning, captioning, reasoning, and summarization, guiding models to first extract relevant acoustic information and then derive answers from it. Audio Flamingo Sound-CoT (Kong et al., 2025) further argues that reasoning chains generated only from text-based LLMs (TLMs) and audio captions can be noisy or hallucinated, and instead uses TLM-LALM collaboration to decompose questions and verify audio-grounded intermediate answers. In its BFS-style construction, the TLM decomposes the original question into sub-questions and the LALM answers each sub-question from the audio; in its DFS-style construction, the TLM and LALM interact over multiple rounds until the reasoning chain reaches a confident conclusion. This collaboration makes the generated CoT less dependent on caption-only surrogates and better aligned with acoustic evidence.

Overall, SFT-based CoT methods are driven by the quality and structure of annotated reasoning

Table 1: Types of Rewards used by RL-Based Audio-to-Text Reasoning Methods. Acc: Accuracy; Con: Consistency; Fmt: Format; Len: Length; Qual: Quality.

Models	Acc	Con	Fmt	Len	Qual
Step-Audio 2	✓	✗	✗	✓	✗
Audio-Thinker	✓	✓	✓	✗	✓
CESAR	✓	✓	✓	✓	✗
Omni-AutoThink	✓	✗	✓	✗	✗
Covo-Audio	✓	✓	✓	✗	✓
R1-AQA	✓	✗	✓	✗	✗
Omni-R1	✓	✗	✗	✗	✗
Sheng et al.	✓	✗	✗	✓	✗
Wijngaard et al.	✓	✗	✓	✗	✗
SARI	✓	✗	✓	✗	✗
Omni-CLST	✓	✗	✓	✗	✗
AudioMCQ	✓	✗	✗	✗	✗
AudSemThinker	✓	✗	✓	✓	✗
Step-Audio R1	✓	✗	✓	✗	✗
Audio-DeepThinker	✓	✓	✓	✗	✓

chains; additional datasets and generation pipelines are discussed in Section B. However, even well-designed SFT data may teach models to imitate reasoning patterns rather than internalize generalizable logic, motivating RL-based approaches that optimize reasoning behavior more directly.

3.1.3 RL-Based CoT Reasoning Methods

RL-based CoT methods optimize reasoning behavior more directly, often after SFT, by rewarding correct, faithful, and useful intermediate reasoning. Their motivation is inherited from RL for verifiable reasoning in text LLMs, but audio introduces additional challenges because the input is continuous, temporally structured, and only partially recoverable from transcripts. Early studies show both promise and limitations. R1-AQA (Li et al., 2025a) reports that RL can outperform SFT on audio reasoning, but also finds that while CoT tokens are helpful at inference time, they may fail to help under GRPO training. Omni-R1 (Rouditchenko et al., 2025) further shows that many MMAU (Sakshi et al., 2024b) audio QA examples can be solved from text alone, and that text-only RL fine-tuning can improve audio QA accuracy. These findings suggest that audio reasoning requires task designs and training objectives that explicitly test reliance on acoustic evidence rather than simply transferring text-domain RL recipes.

Recent RL methods therefore focus on three design issues. First, **difficulty-aware training** adjusts optimization according to problem difficulty.

One line of work estimates difficulty with external or backbone models and then applies curriculum learning, such as first training on easier examples before expanding to the full data distribution (Wijngaard et al., 2025b; Wen et al., 2025). Another line treats easy and hard examples differently during RL: Omni-AutoThink (Yang et al., 2025c) and Omni-CLST (Zhao et al., 2025) reduce or remove reasoning-token losses for easy questions, while Sheng et al. (Sheng et al., 2025) use attention entropy over audio tokens as a difficulty signal and reward longer reasoning on harder examples. This reflects a common intuition: unnecessary CoT may add noise for simple audio questions, whereas difficult questions require explicit decomposition.

Second, **audio-grounded training** selects or emphasizes examples whose answers depend on acoustic information. AudSemThinker (Wijngaard et al., 2025a) extracts auditory semantics such as who, what, how, when, and where before reasoning, encouraging the model to organize relevant acoustic cues. AudioMCQ (He et al., 2025) categorizes audio reasoning problems into two groups based on whether the model requires audio information to answer the questions correctly. They conduct experiments by replacing the original audio with 30 seconds of silence, proposing that RL training should only be applied to problems that necessitate audio. Step-Audio-R1 (Tian et al., 2025) similarly anchors RL in acoustic evidence through iterative rollouts, selecting trajectories that explicitly analyze acoustic features. This direction is especially important because benchmark artifacts can otherwise reward text-prior reasoning rather than genuine listening.

Third, **reward design** typically combines complementary objectives for final-answer accuracy, reasoning consistency, output format, reasoning length, and reasoning quality. Accuracy rewards check whether the final answer is correct; consistency rewards align the reasoning trajectory with the answer; format rewards enforce parsable structures such as <think> and <answer>; length rewards prevent both reasoning collapse and excessive verbosity; and quality rewards evaluate whether intermediate steps are logically valid and acoustically justified. Step-Audio 2 (Wu et al., 2025a), Audio-Thinker (Wu et al., 2025c), Audio-DeepThinker (He et al., 2026), CESAR (Fan et al., 2025), and Omni-AutoThink (Yang et al., 2025c) instantiate different combinations of these reward types, as summarized in Table 1.

Beyond the standard setting, some works extend audio-to-text reasoning to more complex tasks. SoundMind (Diao et al., 2025) studies Audio Logical Reasoning, where models judge whether spoken premises entail a conclusion. This setting requires formal deductive reasoning directly over speech audio rather than simply answering content questions. MALLM (Chen et al., 2024a) extends the standard audio-to-text reasoning to the multi-audio scenario, where the models are required to reason over multiple separate audio clips to solve the problem.

3.1.4 Discussion: Is CoT Really Helping?

Despite the adoption of CoT techniques across inference-time, SFT-based, and RL-based paradigms, a closer examination of the empirical findings reveals a more nuanced picture. While CoT reasoning has proven transformative in the text domain, its benefits in audio reasoning are neither universal nor guaranteed. Several studies raise questions about when, and whether, intermediate reasoning chains genuinely contribute to performance gains—or whether they introduce new failure modes unique to the audio modality. For instance, Audio-CoT (Ma et al., 2025a) demonstrates that explicit step-by-step reasoning can paradoxically hurt performance on harder problems, while R1-AQA (Li et al., 2025a) finds that CoT tokens that are beneficial at inference time fail to improve—and may even hinder performance under GRPO training. Most strikingly, Omni-R1 (Rouditchenko et al., 2025) reveals that a substantial portion of audio QA benchmarks can be solved by a model that never listens to the audio at all. Collectively, these anomalies indicate that discovering how to effectively and universally leverage CoT to drive genuine audio reasoning abilities remains a critical open challenge.

4 Audio-to-Speech Reasoning

Integrating reasoning capabilities into end-to-end SLMs fundamentally changes how latency, context, and logical depth are jointly managed, in contrast to traditional cascaded ASR–LLM–TTS pipelines. In this survey, we adopt the terminology “Audio-to-Speech” to precisely delineate the scope of “Audio-in, Speech-out” interaction. We use “Audio-in” rather than “Speech-in” because modern acoustic encoders can process not only spoken utterances but also environmental sounds and paralinguistic acoustic cues, making the input modality broader than speech alone. Regarding the output modality,

we strictly focus on “Speech-out” systems designed for spoken conversational interaction¹.

4.1 Sequential Audio-to-Speech Reasoning

Early end-to-end Audio-to-Speech models, such as Qwen2.5-Omni (Xu et al., 2025a), Kimi-Audio (Ding et al., 2025), GLM-4-Voice (Zeng et al., 2024), Llama-omni (Fang et al., 2025), Mini-Omni (Xie and Wu, 2024a), Mini-Omni 2 (Xie and Wu, 2024b), and SLAM-Omni (Chen et al., 2025), mainly inherit reasoning abilities from text-based LLM backbones. Although they can reason over audio inputs, their performance is often limited by the text-audio modality gap (Cui et al., 2026) and the lack of audio-specific reasoning supervision. Recent systems, including Qwen3-Omni (Xu et al., 2025b), Qwen3.5-Omni (Team, 2026), Step-Audio 2 (Wu et al., 2025a), Mimo-Audio (Xiaomi, 2025), and OpenS2S (Wang et al., 2025a), begin to incorporate audio-related reasoning data or explicit thinking stages to improve acoustic understanding and spoken response generation.

We refer to this paradigm as *sequential* reasoning because the model first completes an internal reasoning process and then generates speech. Formally, the generation process can be factorized as:

$$P(\mathbf{R}, \mathbf{S} \mid \mathbf{C}) = P(\mathbf{R} \mid \mathbf{C})P(\mathbf{S} \mid \mathbf{R}, \mathbf{C}). \quad (3)$$

This design preserves the logical integrity of the response, but it also delays speech generation, making latency a major bottleneck for spoken interaction and motivating the real-time reasoning paradigms discussed next. More details are provided in Appendix C.

4.2 Realtime Audio-to-Speech Reasoning

Unlike the non-streaming formulations in Section 4.1, real-time Audio-to-Speech reasoning requires input perception and output generation to proceed incrementally and concurrently, imposing strict causal and temporal constraints on the reasoning process. Let the streaming audio input be defined as: $\mathbf{A}^{\text{str}} = \{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(C_s)}\}$, and let the generated spoken output be a sequence of chunks: $\mathbf{S}^{\text{str}} = \{\mathbf{S}_1, \mathbf{S}_2, \dots\}$. The sequential formulation in Eq. (3) is insufficient for these systems, because the model cannot wait for the full input $\mathbf{A}$ or the complete reasoning sequence $\mathbf{R}$. Instead, it

must dynamically balance reasoning depth against strict response latency at each incremental step.

The architectural strategies for incorporating reasoning into real-time audio-in speech-out models can be broadly classified into two categories based on *when* the reasoning computation occurs relative to the user’s input and the system’s output: reasoning that overlaps with the listening window (Section 4.2.1), and reasoning that overlaps with the speaking window (Section 4.2.2).

4.2.1 Category 1: Thinking While Listening (Reasoning during Input)

Methods in this category share a core premise: the duration of the user’s speech provides a natural time window that can be actively repurposed for computation. Instead of treating the listening phase as passive transcription followed by a delayed deliberation step, these systems reason simultaneously while the user is still speaking. Consequently, by the time the user finishes their utterance, the model is already prepared to respond.

Formally, this concurrent process is modeled as: $P(\mathbf{R}_t \mid \mathbf{A}_{\leq t}^{\text{str}}, \mathbf{X}, \mathbf{R}_{< t})$, where $\mathbf{A}_{\leq t}^{\text{str}} = \{\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(t)}\}$. This formulation ensures that the reasoning state at any time t relies strictly on the audio prefix observed up to that moment.

Existing approaches in this category differ across three key design choices. The first is the *reasoning format*—how the internal thoughts are represented, such as through free-form text tokens, structured nodes, or latent states. The second is the *trigger condition*—whether the reasoning runs continuously as the audio streams in, or is completeness-gated by the detection of a full phrase. The third is the *update granularity*—how frequently the reasoning state is refreshed, which can be processed per-chunk, per-segment, or per-step.

SHANKS (Chiang et al., 2025) performs chunk-level free-form reasoning and uses the updated state to support barge-in decisions. Shih et al. (Shih et al., 2026) trigger reasoning based on an entropy-based Question Completeness score (ζ), avoiding both end-of-turn waiting and redundant per-chunk reasoning. Chronological Thinking (Wu et al., 2025b) represents reasoning as typed nodes aligned with semantic segments, so the model can halt reasoning and respond once the user stops speaking. FLAIR (Wu et al., 2026b) replaces explicit CoT tokens with latent hidden-state updates supervised by a non-causal expert, reducing annotation needs and inference latency at the cost of interpretability.

¹We exclude systems targeting general audio synthesis or music generation, whose objectives (acoustic fidelity, musicality) differ fundamentally from reasoning-driven conversational interaction.

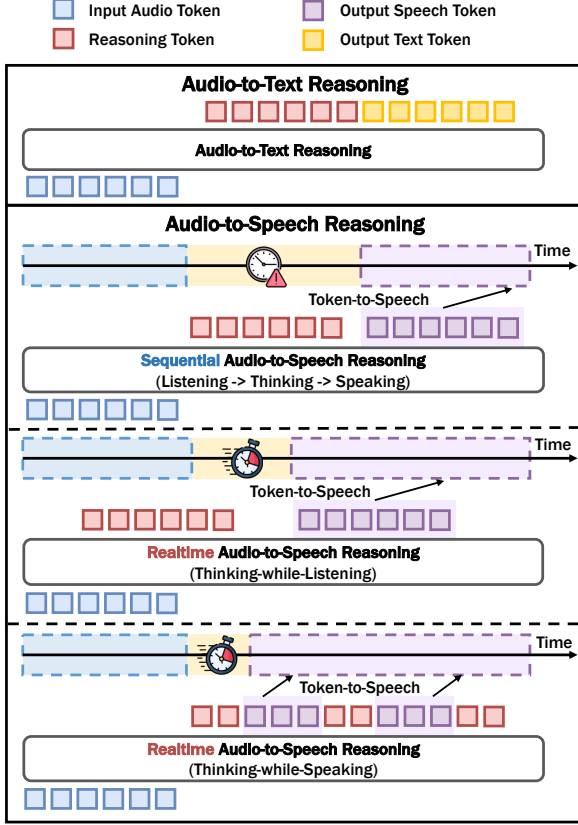


Figure 3: **Overview of major audio reasoning paradigms.** It contrasts text-output reasoning, sequential and real-time speech-output reasoning.

4.2.2 Category 2: Thinking While Speaking (Reasoning during Output)

The second category addresses latency during the generation stage. Rather than borrowing time from the user’s listening window, these systems take advantage of a hardware-level speed asymmetry: GPUs can generate tokens much faster than the real-time rate at which audio is played back. This speed difference creates a time buffer during playback, allowing the system to pre-compute subsequent reasoning and speech tokens simultaneously. The model jointly predicts the current speech chunk and the next reasoning step: $P(S_t, R_{t+1} \mid C, S_{<t}, R_{\leq t})$. Here, S_t denotes the t -th outgoing speech chunk and R_t denotes the reasoning produced in the t -th computation window. Notably, the generation is conditioned on the fully received audio A rather than a streaming prefix $A_{<t}^{\text{str}}$. This reflects the current paradigm of most models in this category, which still operate sequentially during the listening phase—waiting for the user’s utterance to conclude before initiating the concurrent thinking-and-speaking output phase.

The index offset between S_t and R_{t+1} captures the asynchronous pipeline: within a single time window, the model emits S_t while concurrently computing R_{t+1} for use in subsequent steps.

Current approaches within this category generally differ in two main aspects. The first is *structural organization*—how reasoning and speech are integrated. Models may alternate reasoning and speech blocks at the chunk level, interleave both token types within a single sequence, or utilize two separate modules running in parallel. The second aspect is the *reasoning objective*—what the system is actually “thinking” about. This reasoning typically focuses either on solving an underlying logical problem or self-reflecting on the quality of the response being generated.

STITCH (Chiang et al., 2026) alternates unspoken reasoning chunks and spoken response chunks, using playback time to compute the next reasoning step. Mini-Omni-Reasoner (Xie et al., 2025b) performs token-level interleaving of silent reasoning and spoken response tokens through a Thinker-Talker architecture. Mind-Paced Speaking (MPS) (Wu et al., 2026a) decouples reasoning and articulation into parallel modules, enabling speech to start before the full reasoning chain is complete. ReEmpathy (Jia et al., 2026) extends this paradigm to reflective reasoning, using self-critiques to refine the empathetic quality of subsequent speech.

4.3 Training Data for Spoken Reasoning

Table 3 summarizes the training data used across the surveyed models, with “(S)” denoting TTS-synthesized versions of originally text-based datasets. Two patterns stand out:

Reliance on synthesized text benchmarks. Native spoken reasoning corpora remain scarce. Current work therefore predominantly adapts text-based benchmarks (e.g., GSM8K, MultiWOZ, TriviaQA) into spoken form via TTS synthesis, with reasoning supervision distilled from strong text LLMs. This adaptation pipeline is shared across nearly all surveyed methods, and explains why *mathematical and logical reasoning* dominates current evaluation—its sequential derivation aligns naturally with discrete reasoning-token generation and inherits a rich supply of text CoT data. *Paralinguistic and empathetic reasoning*, targeted by ReEmpathy, is a smaller but emerging direction where suitable text sources are less readily available; FLAIR similarly extends spoken reasoning to open-domain QA via synthetic dialogues, indicat-

ing that data adaptation strategies are beginning to diversify beyond math benchmarks.

Annotation strategies diverge by architecture. Although the underlying audio data are largely shared, how reasoning supervision is attached to it differs: structured node parsing (Chronological Thinking), text-audio interleaving at chunk or token granularity (STITCH, Mini-Omni-Reasoner), and reflection-token annotation (ReEmpathy). FLAIR is a notable exception that eliminates explicit reasoning annotations altogether—its latent targets are produced on-the-fly by a non-causal expert via ELBO optimization, sidestepping the dataset construction burden the other methods share.

4.4 Discussion: Reasoning with Latency Constraints

While all surveyed methods aim to reduce Time-To-First-Token (TTFT), they exhibit distinct performance profiles depending on how reasoning and audio generation are temporally synchronized.

Thinking While Listening eliminates wait time after the user finishes speaking, but is fundamentally bounded by the duration of the user’s utterance and, more critically, operates on incomplete information. A short query that requires deep reasoning may leave insufficient computation budget, and reasoning initiated too early risks committing to an incorrect interpretation if the user’s final words alter the underlying condition. Latent-reasoning variants such as FLAIR partially mitigate this by allowing smoother revision through a continuous hidden state, at the cost of interpretability.

Thinking While Speaking is bounded instead by the ratio of reasoning throughput to audio playback speed. Models like STITCH (Chiang et al., 2026) and Mini-Omni-Reasoner (Xie et al., 2025b) hide inference latency by completing reasoning for the $(N+1)$ -th chunk before playback of the N -th chunk finishes. When reasoning chains exceed the playback buffer, however, audio generation may stall, producing perceptible pauses or disfluencies. Since neither paradigm is universally optimal, a natural direction is a hybrid architecture with a *reasoning scheduler* that dynamically selects the appropriate strategy—e.g., based on query completeness or whether the reasoning budget fits within the playback buffer—and handles graceful recovery from premature reasoning. A complementary direction, outside the scope of this survey, is to offload reasoning to an external text-based back-end running asynchronously alongside the speech model

(e.g., (Kuroki et al., 2025; Chien et al., 2026)); such designs could in principle be combined with the internal reasoning paradigms discussed above.

5 Evaluations

Evaluating audio reasoning requires assessing not only final-answer correctness, but also whether the reasoning is grounded in acoustic evidence. We summarize existing evaluation practices for the two settings studied in this survey. Table 4 presents an overview of recent audio reasoning benchmarks.

5.1 Audio-to-Text Evaluation

Audio-to-Text evaluation asks whether a model can produce textual answers from audio inputs while using the acoustic evidence. Existing benchmarks can be broadly divided into two categories:

Content-based reasoning includes tasks where inference depends on linguistic content, such as spoken QA (Chen et al., 2024b), dialogue understanding (Gosai et al., 2025; Zhang et al., 2025a), and instruction following (Lu et al., 2025; Wang et al., 2025c). The model needs to recover meaning and reason over it to produce correct responses. Instruction following in speech is a form of content-based reasoning, as it requires understanding and executing semantic intent. Recent settings (Lu et al., 2025; Wang et al., 2025c; Chen et al., 2024b) introduce content-preserving acoustic variations, such as emotion, speaking rate, or background noise. These aim not to test reasoning over acoustic cues, but to evaluate whether content-based reasoning remains stable under such perturbations.

Acoustic-based reasoning involves tasks where correct inference depends on information not reducible to text, including prosody, emotion, speaker traits, emphasis, hesitation, and other paralinguistic signals (Sakshi et al., 2024a; Wang et al., 2025b; Ma et al., 2025b). Text transcripts alone are insufficient, and the model must use acoustic evidence. Thus, acoustic-based reasoning more directly tests whether models leverage speech-specific information, rather than relying on implicit text-only processing. This distinction is key to understanding whether systems truly reason over speech or mainly over textual abstractions. Detailed descriptions of the evaluation metrics are provided in Appendix D.

5.2 Audio-to-Speech Evaluation

Audio-to-Speech evaluation inherits the goals of Audio-to-Text evaluation, but adds requirements

specific to spoken interaction. The evaluation goes further to encompass interaction ability, response alignment, and real-time behavior under streaming constraints that are unique to spoken dialogue systems. First, *dialogue-oriented ability* measures coherence across turns, contextual adaptation, pragmatic cues, turn-taking, overlap, and interruptions. Second, *speech quality* evaluates naturalness, intelligibility, prosody, fluency, and absence of audio artifacts. Third, *response alignment* assesses whether speech matches the semantic content, dialogue history, intended emotion, and paralinguistic expression. Fourth, *real-time interaction* measures latency, streaming stability, and incremental generation. Finally, *subjective evaluation*, including MOS and human or model-based judgments, captures perceptual quality, emotional appropriateness, coherence, and conversational effectiveness.

6 Challenges and Future Directions

Reliability of Synthesized Reasoning Data. Audio reasoning is constrained by the lack of high-quality, audio-grounded reasoning data. Existing methods often reduce annotation costs by using text-only LLMs to synthesize reasoning chains from transcripts, captions, or sound event tags. Yet these textual proxies may fail to capture fine-grained acoustic cues, making the resulting reasoning weakly grounded in the original audio. Scalable construction of faithful acoustic reasoning data remains a key bottleneck.

Modality Hallucination and Shortcut Learning. Current LALMs often inherit reasoning abilities from text-dominant backbones, creating a mismatch between linguistic reasoning and acoustic grounding. This can lead to shortcut learning or modality hallucination, where models rely on audio captions, transcripts, or intermediate text representations instead of the continuous acoustic signal (Tian et al., 2025). Consequently, these models sometimes bypass the auditory input entirely, relying excessively on the strong semantic priors of their language backbone to infer answers. Reducing such text-surrogate reasoning and anchoring inference in raw acoustic cues is essential for faithful and robust audio reasoning.

The Accuracy vs. Latency Trade-Off. Audio-to-Speech reasoning faces an trade-off between reasoning accuracy and interaction latency. Longer reasoning chains can improve deduction but delay

spoken responses, disrupting natural dialogue. Recent paradigms such as “thinking while listening” and “thinking while speaking” reduce this delay by overlapping reasoning with user input or system output. However, streaming methods lag behind offline or text-based reasoning systems in accuracy and deduction quality (Chiang et al., 2026, 2025). Closing this gap while preserving low-latency, full-duplex interaction remains a key challenge.

Long-context Audio Reasoning. Current methods in audio reasoning have primarily demonstrated efficacy on relatively short, well-defined audio clips. However, scaling these reasoning capabilities to process continuous, long-form audio streams—such as meetings, podcasts, or environmental recordings lasting several minutes—introduces significant computational and cognitive uncertainties. Processing to such inputs is challenging because uncompressed, temporally dense acoustic features create prohibitively long sequences, and it is unclear whether models can preserve precise perception and context over extended durations. Multi-hop reasoning over long audio, where events separated by minutes must be recalled and integrated, therefore requires dedicated benchmarks and more efficient architectures.

Shifting Toward Foundational Pre-Training for Reasoning. Most audio reasoning methods rely on post-training after basic modality alignment, typically through SFT or RL. This raises questions about the foundational limits of the model capabilities. Future work should explore reasoning-oriented pre-training or mid-training objectives, such as large-scale audio-text prediction and acoustic grounding. Whether these objectives can make reasoning more native to the audio modality and improve subsequent post-training remains open.

7 Conclusion

This survey reviews audio reasoning in multi-modal foundation models, framing it as reasoning-augmented generation grounded in acoustic evidence. We organize recent work into Audio-to-Text and Audio-to-Speech reasoning, and summarize evaluation practices and key challenges, including data scarcity, shortcut learning, modality hallucination, latency constraints, and long-context reasoning. It offers a concise roadmap for faithful and efficient audio-grounded reasoning systems.

Limitations

This survey intentionally focuses on Audio-to-Text and Audio-to-Speech reasoning, the two most representative settings in current audio reasoning research. Due to conference space constraints, related directions such as Audio-Visual reasoning, Agentic Audio Reasoning, and broader architectural foundations are discussed selectively or deferred to the appendix, and can be further expanded in a longer version. Since existing studies also differ substantially in datasets, tasks, and evaluation protocols, we emphasize conceptual organization and methodological comparison rather than exhaustive empirical ranking. This focused scope enables a clearer synthesis of the emerging field.

Acknowledgments

The research presented in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 2300246, RGC C1043-24G), (CUHK 14203425, RGC GRF 2151317), and CUHK 7010870.

References

- 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. *Advances in Neural Information Processing Systems*, 37:56898–56918.
- Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.
- Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning (ICML)*, pages 5178–5193. PMLR.
- Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, Kai Yu, Yuxuan Hu, Jinyu Li, Yan Lu, Shujie Liu, and Xie Chen. 2025. *SLAM-omni: Timbre-controllable voice interaction system with single-stage training*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2262–2282, Vienna, Austria. Association for Computational Linguistics.
- Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D’Haro, Robby T Tan, and Haizhou Li. 2024a. 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. 2024b. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.
- Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. 2025. Shanks: Simultaneous hearing and thinking for spoken language models. *arXiv preprint arXiv:2510.06917*.
- Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie LIU, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang. 2026. *STITCH: Simultaneous thinking and talking with chunked reasoning for spoken language models*. In *The Fourteenth International Conference on Learning Representations*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. <https://lmsys.org/blog/2023-03-30-vicuna/>. Accessed: 2023-03-30.
- Chung-Ming Chien, Manu Orsini, Eugene Kharitonov, Neil Zeghidour, Karen Livescu, and Alexandre Défossez. 2026. *Moshirag: Asynchronous knowledge retrieval for full-duplex speech language models*. Preprint, arXiv:2604.12928.
- Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025a. Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16735–16753.
- Wenqian Cui, Xiao-Hui Li, Daxin Tan, Qiyong Zheng, and Irwin King. 2026. *Minimizing modality gap from the input side: Your speech llm can be a prosody-aware text llm*. *arXiv preprint arXiv:2605.05927*.
- Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025b. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.
- Alexandre Défossez and 1 others. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00080*.

- Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiye Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. 2025. Soundmind: RL-incentivized logic reasoning for audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 528–540.
- Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.
- Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. 2025. Incentivizing consistent, effective and scalable reasoning capability in audio llms via reasoning process rewards. *arXiv preprint arXiv:2510.20867*.
- Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*, volume 2025, pages 57607–57624.
- Qingkai Fang, Shoutao Niu, Rui Zhou, Zhengrui Lin, Minpeng Chen, and Yang Feng. 2024. LLaMA-Omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.
- Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4763–4784.
- Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. In *Proc. Interspeech 2021*, pages 571–575.
- Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and 1 others. 2025. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. *arXiv preprint arXiv:2512.14865*.
- 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*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. 2025b. Recent advances in discrete speech tokens: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, and 1 others. 2025. Measuring audio’s impact on correctness: Audio-contribution-aware post-training of large audio language models. *arXiv preprint arXiv:2509.21060*.
- Xiang He, Chenxing Li, Jinting Wang, Yan Rong, Tianxin Xie, Wenfu Wang, Li Liu, and Dong Yu. 2026. Audio-deepthinker: Progressive reasoning-aware reinforcement learning for high-quality chain-of-thought emergence in audio language models. *arXiv preprint arXiv:2604.18187*.
- 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.
- Dhruv Jain, Harshit Shukla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Voiceagentbench: Are voice assistants ready for agentic tasks? *arXiv preprint arXiv:2510.07978*.
- Yuhang Jia, Pei Liu, Haoqin Sun, Jiaming Zhou, Xuxin Cheng, Cao Liu, Ke Zeng, Xunliang Cai, and Yong Qin. 2026. Reflecting twice before speaking with empathy: Self-reflective alternating inference for empathy-aware end-to-end spoken dialogue. *arXiv preprint arXiv:2601.18281*.
- Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Zhifeng Kong, Arushi Goel, Joao Felipe Santos, Sreyan Ghosh, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2025. Audio flamingo sound-cot technical report: Improving chain-of-thought reasoning in sound understanding. *arXiv preprint arXiv:2508.11818*.
- Sonal Kumar, Sreyan Ghosh, Yueqian Lin, Yiran Chen, Ramani Duraiswami, and Dinesh Manocha. 2025a. Polyaudio: Advancing multi-audio analysis & reasoning in large audio language models.
- Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeon-gon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, and 1 others. 2025b. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*.
- So Kuroki, Yotaro Kubo, Takuya Akiba, and Yujin Tang. 2025. Kame: Tandem architecture for enhancing knowledge in real-time speech-to-speech conversational ai. *arXiv preprint arXiv:2510.02327*.

- Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, and 1 others. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.
- Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025a. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.
- Hui Li, Changhao Jiang, Hongyu Wang, Ming Zhang, Jiajun Sun, Zhixiong Yang, Yifei Cao, Shihan Dou, Xiaoran Fan, Baoyu Fan, and 1 others. 2025b. Mdar: A multi-scene dynamic audio reasoning benchmark. *arXiv preprint arXiv:2509.22461*.
- Longhao Li, Hongjie Chen, Zehan Li, Qihan Hu, Jian Kang, Jie Li, Lei Xie, and Yongxiang Li. 2026a. Audio-cogito: Towards deep audio reasoning in large audio language models. *arXiv preprint arXiv:2604.12527*.
- Yangzhuo Li, Shengpeng Ji, Yifu Chen, Tianle Liang, Haorong Ying, Yule Wang, Junbo Li, Jun Fang, and Zhou Zhao. 2026b. Wavbench: Benchmarking reasoning, colloquialism, and paralinguistics for end-to-end spoken dialogue models. *arXiv preprint arXiv:2602.12135*.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025c. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.
- Guan-Ting Lin, Chen Chen, Zhehuai Chen, and Hung-yi Lee. 2026a. Full-duplex-bench-v3: Benchmarking tool use for full-duplex voice agents under real-world disfluency. *arXiv preprint arXiv:2604.04847*.
- Jingru Lin, Chen Zhang, Tianrui Wang, and Haizhou Li. 2026b. Audiorag: A challenging benchmark for audio reasoning and information retrieval. *arXiv preprint arXiv:2602.10656*.
- 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. *arXiv preprint arXiv:2505.19037*.
- Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025a. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.
- Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2025b. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.
- Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, and 1 others. 2025. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Preprint*, arXiv:2212.04356.
- 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 (ICML)*, pages 28492–28518. PMLR.
- Soham Ray, Keshav Dhandhanian, Victor Barres, and Karthik Narasimhan. 2026. τ -voice: Benchmarking full-duplex voice agents on real-world domains. *arXiv preprint arXiv:2603.13686*.
- Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. Omni-r1: Do you really need audio to fine-tune your audio llm? *arXiv preprint arXiv:2505.09439*.
- S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024a. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.
- Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024b. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.
- Sakshi 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 *International Conference on Learning Representations*, volume 2025, pages 84929–84964.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Zhichao Sheng, Shilin Zhou, Chen Gong, and Zhenghua Li. 2025. Think smart, not hard: Difficulty adaptive reasoning for large audio language models. *arXiv preprint arXiv:2509.21960*.

- Yi-Jen Shih, Desh Raj, Chunyang Wu, Wei Zhou, SK Bong, Yashesh Gaur, Jay Mahadeokar, Ozlem Kalinli, and Mike Seltzer. 2026. [Can speech LLMs think while listening?](#) In *The Fourteenth International Conference on Learning Representations*.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Zayne Sprague, Fangcong Yin, Juan Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, volume 2025, pages 94118–94162.
- Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*.
- Termeh Taheri, Yinghao Ma, and Emmanouil Benetos. 2025. Sar-lm: Symbolic audio reasoning with large language models. *arXiv preprint arXiv:2511.06483*.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Qwen Team. 2026. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.
- Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, and 1 others. 2025. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2024. [Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing](#). *Preprint*, arXiv:2309.00916.
- Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and 1 others. 2025a. Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 906–917.
- Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. 2026. Emotionthinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. *arXiv preprint arXiv:2601.15668*.
- Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025b. [MMSU: A massive multi-task spoken language understanding and reasoning benchmark](#). *CoRR*, abs/2506.04779.
- Dingdong Wang, Jin Xu, Ruihang Chu, Zhifang Guo, Xiong Wang, Jincenzi Wu, Dongchao Yang, Shengpeng Ji, and Junyang Lin. 2025c. Insertor: Speech instruction following with unsupervised interleaved pre-training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18024–18046.
- Siyin Wang, Zengrui Jin, Changli Tang, Qiujia Li, Bo Li, Chen Chen, Yuchen Hu, Wenyi Yu, Yixuan Li, Jimin Zhuang, and 1 others. 2025d. Towards general auditory intelligence: Large multimodal models for machine listening and speaking. *arXiv preprint arXiv:2511.01299*.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025e. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.
- Chengwei Wei, Bin Wang, Jung-jae Kim, and Nancy F Chen. 2025. Towards spoken mathematical reasoning: Benchmarking speech-based models over multi-faceted math problems. *arXiv preprint arXiv:2505.15000*.
- 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.
- Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. 2025. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*.
- Gijs Wijnngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. 2025a. Audsemthinker: Enhancing audio-language models through reasoning over semantics of sound. *arXiv preprint arXiv:2505.14142*.
- Gijs Wijnngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. 2025b. Data-balanced curriculum learning for audio question answering. *arXiv preprint arXiv:2507.06815*.
- Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025a. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

- Donghang Wu, Haoyang Zhang, Chen Chen, Tianyu Zhang, Fei Tian, Xuerui Yang, Gang Yu, Hexin Liu, Nana Hou, Yuchen Hu, and 1 others. 2025b. Chronological thinking in full-duplex spoken dialogue language models. *arXiv preprint arXiv:2510.05150*.
- Donghang Wu, Haoyang Zhang, Jun Chen, Xiangyu, Zhang, Hexin Liu, Eng Siong Chng, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2026a. [Mind-paced speaking: A dual-brain approach to real-time reasoning in spoken language models](#). *Preprint*, arXiv:2510.09592.
- Donghang Wu, Tianyu Zhang, Yuxin Li, Hexin Liu, Chen Chen, Eng Siong Chng, and Yoshua Bengio. 2026b. The silent thought: Modeling internal cognition in full-duplex spoken dialogue models via latent reasoning. *arXiv preprint arXiv:2603.17837*.
- Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. 2025c. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025d. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations*.
- LLM-Core-Team Xiaomi. 2025. [Mimo-audio: Audio language models are few-shot learners](#).
- Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025a. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.
- Zhifei Xie, Ziyang Ma, Zihang Liu, Kaiyu Pang, Hongyu Li, Jialin Zhang, Yue Liao, Deheng Ye, Chunyan Miao, and Shuicheng Yan. 2025b. Mini-omni-reasoner: Token-level thinking-in-speaking in large speech models. *arXiv preprint arXiv:2508.15827*.
- Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.
- Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. [Qwen2.5-omni technical report](#). *Preprint*, arXiv:2503.20215.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.
- Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- 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. *arXiv preprint arXiv:2505.13237*.
- Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. 2025b. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10181.
- Dongchao Yang, Songxiang Liu, Disong Wang, Yuanyuan Wang, Guanglu Wan, and Helen Meng. 2025c. Omni-autothink: Adaptive multimodal reasoning via reinforcement learning. *arXiv preprint arXiv:2512.03783*.
- Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024b. 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.
- Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speech gpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.
- He Zhang, Wenqian Cui, Haoning Xu, Xiaohui Li, Lei Zhu, Haoli Bai, Shaohua Ma, and Irwin King. 2025a. Mtr-duplexbench: Towards a comprehensive evaluation of multi-round conversations for full-duplex speech language models. *arXiv preprint arXiv:2511.10262*.
- Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. 2025b. Wildspeech-bench: Benchmarking end-to-end speechllms in the wild. *arXiv preprint arXiv:2506.21875*.

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

Jinghua Zhao, Hang Su, Lichun Fan, Zhenbo Luo, Hui Wang, Haoqin Sun, and Yong Qin. 2025. Omni-clst: Error-aware curriculum learning with guided selective chain-of-thought for audio question answering. *arXiv preprint arXiv:2509.12275*.

A Foundations for Audio Reasoning Models

Audio reasoning fundamentally requires systems capable of deciphering the complex, continuous nature of sound and translating these acoustic signals into logical comprehension or expressive generation. The foundation of these advanced cognitive abilities is predicated upon two critical dimensions—This section discusses the 1) *underlying model architectures* designed to effectively bridge audio signals with language modeling, alongside the 2) *training technologies* necessary to cultivate the reasoning abilities.

A.1 Underlying Model Architectures

A.1.1 Large Audio Language Models (LALMs)

LALMs aim to enable the audio comprehension abilities to the LLMs. They take audio and optional text instructions as input and generate text as output. Their dominant architecture is an encoder–projector–LLM pipeline, where an audio encoder first extracts continuous representations from the input audio, a modality projector maps these representations into the embedding space of a text LLM, and the LLM then performs autoregressive text generation conditioned on the projected audio features and optional textual instructions. In LALMs, audio serves as an external input modality aligned with the textual reasoning space of the LLMs. This design enables efficient reuse of pretrained audio encoders (Radford et al., 2023; Hsu et al., 2021; Chen et al., 2023; Gong et al., 2021; Radford et al., 2022) and text LLM backbones (Touvron et al., 2023; Grattafiori et al., 2024; Yang et al., 2024a; Chiang et al., 2023), making LALMs modular and extensible.

A.1.2 Spoken Language Models (SLMs)

SLMs are foundation models built for spoken dialogue interactions. Similar to LALMs, SLMs can

take audio and optional text as inputs, but their outputs are speech representations or waveforms rather than text. Architecturally, they can follow either the encoder–projector–LLM–decoder design that resembles LALMs but adds a speech generation module (Zeng et al., 2024; Fang et al., 2024) after the LLM, or a more unified token-based design (Zhang et al., 2023; Défossez et al., 2024), in which continuous speech is first converted into discrete speech tokens and both input and output speech are modeled autoregressively in a shared token space together with text. The latter formulation is more end-to-end, but it also introduces challenges such as long token sequences, multi-codebook token structures, and waveform reconstruction from generated tokens. In short, the main difference is that LALMs align audio to text for text generation, whereas SLMs must additionally support speech generation, requiring an explicit mechanism for producing and decoding speech representations.

A.2 Training Technologies

While architectural designs provide the structural foundation for multimodal perception, the emergence of complex reasoning capabilities relies fundamentally on specific training paradigms. In this part, we explore the underlying mechanisms of the core training paradigms that empower audio models with reasoning capabilities.

A.2.1 Cross-Modal Alignment Pre-training

Before an LLM can reason over audio, it must first learn to associate acoustic patterns with linguistic meaning (Wang et al., 2024). The goal of cross-modal alignment pre-training is therefore to bridge the semantic gap between the audio modality and the LLM’s native language space. During this stage, the LLM backbone is often frozen, or only lightly updated, to preserve its linguistic and reasoning capabilities, while the bridging components, such as the modality projector in LALMs or the embedding layers for discrete audio tokens in SLMs, are optimized (Tang et al., 2024; Zhang et al., 2023). This stage typically relies on large-scale audio-text data, often instantiated through tasks such as ASR and audio captioning. The audio context may refer either to projected continuous features or to discrete audio tokens, and the model is usually trained with the standard autoregressive cross-entropy loss.

Table 2: Summary of training datasets and construction methods for Audio-to-Text Reasoning.

Model	Source Data	CoT Construction	Training	Scale
Audio-CoT	MMAU (eval only)	Prompt-based (no training)	—	—
SAR-LM	MMAU, MMAR (eval only)	Symbolic features (no training)	—	—
Audio-Reasoner	AudioSet, AudioCaps, + 7 others	Gemini (structured 4-part)	SFT	1.2M
AF Sound-CoT	AudioSkills, Clotho-AQA, + 10 others	LLM-ALM collaborative (BFS/DFS)	SFT	1.24M
Audio-Cogito	AudioSet, Clotho, AudioCaps, + 7 others	Self-distillation (Qwen3-OmniThinking)	SFT	545k
R1-AQA	AVQA	None (direct answer)	GRPO	38k
Omni-R1	AVQA, VGGSound	ChatGPT-generated QA	GRPO	38k–54k
SARI	AudioSet, MusicBench, MELD, AVQA	Qwen2.5-72B (structured + unstructured)	SFT+GRPO	42k
Audio-Thinker	AVQA	Qwen2.5-72B (structured 4-part)	SFT+GRPO	40k
AudioMCQ	7 datasets (Clotho, AudioCaps, etc.)	Qwen3-235B (3-stage + simplified)	SFT+GRPO	571k
AudSemThinker	YouTube closed captions	Qwen2.5-72B (2/3-phase semantic)	SFT+GRPO	213k
Omni-CLST	AVQA, AudSemThinker data	Reused (guided dropout)	SFT+GRPO	249k
Omni-AutoThink	Audio-Reasoner, + multi-modal	Hierarchical difficulty calibration	SFT+GRPO	571k
Sheng et al.	AVQA	—	GRPO	38k
Wijngaard et al.	AVQA, ClothoAQA, + 4 others	Gemini2.5-Pro (cold-start)	SFT+GRPO	64k
CESAR	AVQA (augmented)	—	GRPO	38k+
Step-Audio 2	AudioSet, AudioCaps, + in-house	LLM from mixing recipes	SFT+RL	Large
Step-Audio-R1	ASR, paralinguistic, + in-house	Self-distillation (MGRD)	SFT+RLVR	5M
EmotionThinker	IEMOCAP, MELD, Espresso, + 2 others	GPT-4o (prosody-enriched)	SFT+GRPO	35k
SoundMind	LogiQA 2.0-NLI (TTS synthesis)	DeepSeek-R1	SFT+RL	6.4k
PolyAudio	AudioSet, LibriSpeech, + 5 others	LLM (multi-audio tasks)	SFT+GRPO	160k
Audio-DeepThinker	AVQA, AudioSet, MagnaTagATune, + others	DeepSeek V3.1	GRPO	69k

A.2.2 Post-training

While cross-modal alignment pre-training establishes a fundamental mapping between acoustic representations and linguistic semantics, this baseline alignment is typically insufficient for robust instruction following, multi-step reasoning, or reliable user interaction. Consequently, the model must undergo post-training—a suite of optimization procedures designed to refine the behavioral priors of the aligned system and elicit the higher-order reasoning capabilities required for complex multimodal tasks.

Post-training typically proceeds in two primary phases: Supervised Fine-Tuning (SFT) and Alignment Optimization. During SFT, the model is trained on curated instruction-response pairs—often enriched with CoT annotations—to learn the structural logic of following complex audio-related commands. While SFT provides a foundation through demonstration, subsequent alignment via Preference Optimization or Reinforcement Learning (RL), such as PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), further refines the model’s behavior. These methods optimize for higher-level objectives such as factual accuracy, reasoning consistency, and faithfulness to the acoustic evidence, ensuring the model’s outputs align with human intent and logical rigor.

B Dataset and Pre-processing for Audio-to-Text Reasoning

The development of audio-to-text reasoning capabilities critically depends on the construction of high-quality training data. Unlike standard audio understanding tasks that rely on simple annotations such as sound event labels or captions, reasoning-oriented training requires explicit intermediate reasoning chains that teach models to decompose complex problems step by step. This subsection surveys the data construction and preprocessing strategies employed across the audio-to-text reasoning literature, covering data sources and reasoning chain construction.

B.1 Data Sources and QA Pair Construction

The majority of existing works repurpose established audio corpora rather than collecting new recordings. AVQA is the most widely adopted source, used by R1-AQA (Li et al., 2025a), Audio-Thinker (Wu et al., 2025c), SARI (Wen et al., 2025), Omni-CLST (Zhao et al., 2025), CESAR (Fan et al., 2025), Omni-R1 (Rouditchenko et al., 2025), and Sheng et al. (Sheng et al., 2025), who extract audio tracks from the original audio-visual dataset and adapt question text by replacing visual references (e.g., “video”) with auditory ones (e.g., “audio”). Other commonly leveraged datasets include AudioSet, AudioCaps, Clotho, MusicBench, and VGGSound, which provide diverse audio content span-

Table 3: Comparison of Datasets and Tasks in Realtime Audio-to-Speech Reasoning Models. “(S)” indicates TTS-synthesized versions of originally text-based datasets.

Model	Domain	Source Data
<i>Category 1: Thinking While Listening</i>		
Chronological Thinking (Wu et al., 2025b)	Task-oriented Dialogue, Open-domain QA	GenConv(S), SpokenWOZ(S), LlamaQ(S)
SHANKS (Chiang et al., 2025)	Math Reasoning	Tulu3-Persona-Math-Grade(S)
Shih et al. (Shih et al., 2026)	Common Sense & Math Reasoning	COT COLLECTION (Kim et al., 2023)(S)
FLAIR (Wu et al., 2026b)	Open-domain QA	Synthetic dialogues (GPT-OSS/Qwen2.5/Llama3.1)
<i>Category 2: Thinking While Speaking</i>		
MPS (Wu et al., 2026a)	Math Reasoning, Complex Conversation	Not disclosed
STITCH (Chiang et al., 2026)	Math Reasoning, Knowledge QA	VoiceAssistant400K, Tulu-3-Persona(S), Natural Questions(S), TriviaQA(S)
Mini-Omni-Reasoner (Xie et al., 2025b)	Math Reasoning	SPOKEN-MATH-PROBLEMS-3M
ReEmpathy (Jia et al., 2026)	Empathetic Dialogue & Reflection	EmotionTalk, OpenS2S subset

ning environmental sounds, music, and speech. AudioMCQ (He et al., 2025) aggregates seven source datasets covering speech (47%), sound (39.1%), music (8.1%), and temporal reasoning (5.8%), constructing 571k multiple-choice QA samples. AudSemThinker (Wijngaard et al., 2025a) constructs an entirely new corpus by mining YouTube closed captions (213k samples) and processing them through an ensemble of nine specialized audio, image, and video analysis models.

Two paradigms dominate QA pair construction. The first directly reuses existing QA annotations with minimal modification, as in R1-AQA (Li et al., 2025a) and Audio-Thinker (Wu et al., 2025c). The second employs LLMs to synthesize new questions and answer choices from audio captions: Omni-R1 (Rouditchenko et al., 2025) uses ChatGPT to generate multiple-choice questions from multi-perspective audio captions, SARI (Wen et al., 2025) uses Qwen2.5-72B, and AudioMCQ (He et al., 2025) uses Qwen3-235B. Audio-Reasoner (Xie et al., 2025a) adopts a progressive difficulty strategy, generating three questions per audio clip at increasing complexity levels—from factual queries to multi-step reasoning problems—using Google Gemini. For multi-audio scenarios, PolyAudio (Kumar et al., 2025a) constructs 110k QA pairs spanning 11 multi-audio reasoning tasks (e.g., cross-clip comparison, chronological ordering, multi-audio entailment) using LLM-generated task-specific prompts.

B.2 Reasoning Chain Construction

The construction of intermediate reasoning chains—critical for SFT-based CoT training—follows several distinct strategies.

Single-LLM generation is the most prevalent approach, where a powerful LLM produces rea-

soning chains given audio captions, questions, and answers. Audio-Reasoner (Xie et al., 2025a) uses Gemini to generate four-component chains (planning, captioning, reasoning, and summarizing) for its 1.2M-sample CoTA dataset. SARI (Wen et al., 2025) uses Qwen2.5-72B to produce both structured (four-section) and unstructured (free-form) CoT formats, demonstrating that structured reasoning yields more robust generalization. AudioMCQ (He et al., 2025) employs Qwen3-235B to generate three-stage reasoning chains (question type analysis, audio content analysis, answer selection) along with simplified unstructured versions. Audio-Cogito (Li et al., 2026a) uses Qwen3-Omni-Thinking to generate free-form reasoning chains.

LLM-ALM collaborative generation addresses the concern that text-only LLMs may hallucinate audio-specific details. Audio Flamingo Sound-CoT (Kong et al., 2025) proposes a collaborative pipeline where a text LLM and an audio language model jointly construct reasoning chains through either parallel sub-question decomposition (BFS-style) or multi-round interactive conversation (DFS-style), ensuring that intermediate reasoning steps are grounded in actual audio content.

Iterative self-distillation enables the model to progressively improve its own reasoning data. Step-Audio-R1 (Tian et al., 2025) introduces Modality-Grounded Reasoning Distillation (MGRD), where the model generates candidate reasoning chains that are then filtered for acoustic grounding (i.e., references to perceptual features rather than text descriptions), logical coherence, and answer correctness, progressively shifting reasoning from text-surrogate to genuinely audio-grounded.

Domain-specific feature-enriched generation augments LLM-based CoT construction with spe-

Table 4: Overview of Reasoning Benchmarks for SLMs. Open and Closed denote open-ended and closed-ended evaluation types, respectively.

Benchmark	Domain	Audio Type	Eval Type	Size	#Tasks	Year
MMAU (Sakshi et al., 2024b)	General Audio Reasoning	Speech, Sound, Music	Closed	10K	27	2024
MMAU-Pro (Kumar et al., 2025b)	General Audio Reasoning	Speech, Sound, Music	Closed	0.5K	49	2025
MMAR (Ma et al., 2025b)	General Audio Reasoning	Speech, Sound, Music	Closed	1K	16	2025
MMSU (Wang et al., 2025b)	Speech-focused Reasoning	Speech	Closed	5K	47	2025
VoxEval (Cui et al., 2025a)	Speech-focused Reasoning	Speech	Closed	14K	56	2025
CMDAR (Li et al., 2025b)	Chinese Audio Reasoning	Speech, Sound	Closed, Open	3K	5	2025
Spoken-MQA (Wei et al., 2025)	Spoken Mathematical Reasoning	Speech	Closed, Open	3K	5	2025
WildSpeech-Bench (Zhang et al., 2025b)	End-to-end Interaction	Speech	Open	1.1K	5	2025
URO-Bench (Yan et al., 2025)	End-to-end Interaction	Speech	Closed, Open	5K	20	2025
VoiceAgentBench (Jain et al., 2025)	Multi-turn Tool Use & Reasoning	Speech	Closed	6K	6	2025
SAKURA (Yang et al., 2025a)	Multi-hop Audio Reasoning	Speech, Sound	Closed	4K	4	2025
WavBench (Li et al., 2026b)	End-to-end Interaction	Speech	Closed, Open	2K	5 scenarios	2026
τ -Voice (Ray et al., 2026)	End-to-end Interaction	Speech	Closed	0.3K	3 domains	2026
AudioRAG (Lin et al., 2026b)	Audio Retrieval-based Reasoning	Speech, Sound, Music	Closed	8K	3 scenarios	2026
Full-Duplex-Bench-v3 (Lin et al., 2026a)	Multi-turn Tool Use & Reasoning	Speech	Closed	0.1K	4	2026

cialized audio analysis. EmotionThinker (Wang et al., 2026) constructs EmotionCoT-35K by first extracting prosodic features—pitch, energy, speaking rate, word-level stress, and intonation patterns—via classical signal processing and specialized models, then using GPT-4o to generate reasoning traces grounded in these acoustic cues.

Table 2 summarizes the training datasets and construction methods across the surveyed audio-to-text reasoning methods.

C Details of Sequential Audio-to-Speech Reasoning

Sequential Audio-to-Speech reasoning refers to systems that perform internal reasoning before producing spoken output. A first group of end-to-end Audio-to-Speech models, including Qwen2.5-Omni (Xu et al., 2025a), Kimi-Audio (Ding et al., 2025), GLM-4-Voice (Zeng et al., 2024), Llama-omni (Fang et al., 2025), Mini-Omni (Xie and Wu, 2024a), Mini-Omni 2 (Xie and Wu, 2024b), and SLAM-Omni (Chen et al., 2025), obtain much of their reasoning ability from text-based LLM backbones. These models support audio-in speech-out interaction, but their reasoning behavior is often constrained by the modality gap between text and audio (Cui et al., 2026) and by limited audio-specific reasoning supervision.

More recent systems explicitly introduce audio-related reasoning data or thinking stages into the training or inference pipeline. Qwen3-Omni (Xu et al., 2025b), Qwen3.5-Omni (Team, 2026), Step-Audio 2 (Wu et al., 2025a), and Mimo-Audio (Xiaomi, 2025) incorporate audio-to-text reasoning data to improve the model’s understanding of complex acoustic inputs or to guide

more appropriate spoken responses. In particular, Qwen3.5-Omni (Team, 2026) uses on-policy distillation to transfer high-quality reasoning from text-conditioned outputs to audio-conditioned responses, reducing the text-audio performance gap. OpenS2S (Wang et al., 2025a) introduces an optional static thinking stage before speech synthesis, allowing the model to reason over the user’s emotional cues before generating an empathetic spoken response.

The defining property of this paradigm is its stepwise dependency: the model first derives an intermediate reasoning sequence $\mathbf{R}$ from the audio-instruction context $\mathbf{C}$, and only then generates the speech output $\mathbf{S}$. This process can be written as:

$$P(\mathbf{R}, \mathbf{S} \mid \mathbf{C}) = P(\mathbf{R} \mid \mathbf{C})P(\mathbf{S} \mid \mathbf{R}, \mathbf{C}). \quad (4)$$

Such a formulation helps preserve response coherence because speech generation is conditioned on a completed reasoning trajectory. However, it also means that perception and reasoning must finish before speech begins, which increases response latency. This latency issue is the main motivation for real-time Audio-to-Speech reasoning methods such as thinking while listening and thinking while speaking.

D Audio-to-Text Evaluation Metrics

Evaluation of audio-to-text tasks typically includes both objective metrics, which measure correctness against references, and subjective evaluation, which assesses quality in open-ended settings.

ACC-Metrics. A common approach is to use accuracy-based metrics from language model benchmarks (Sakshi et al., 2024a; Wang et al.,

2025b; Ma et al., 2025b; Sakshi et al., 2024b), especially when outputs are discrete or can be normalized into a small set. This is typical in multiple-choice QA, classification, and closed-ended reasoning. Common metrics include accuracy, F1, and, in retrieval or imbalanced settings, mean average precision (mAP). These directly measure agreement with ground truth and suit controlled reasoning evaluation. Many spoken reasoning benchmarks adopt this paradigm for clear comparison and reduced ambiguity.

Text-overlap Metrics. For free-form outputs, exact match can be too strict, as correct answers may vary in wording. Text-overlap metrics such as BLEU, ROUGE, and METEOR measure similarity between outputs and references. They are useful for structured generation tasks like summarization, short QA, or captioning, where references exist and surface similarity is informative. However, they poorly capture reasoning quality, as responses may be semantically correct but lexically different, or similar yet incorrect.

Human Evaluation. Subjective evaluation is important for open-ended tasks where multiple responses are valid or correctness depends on nuanced interpretation. Evaluators assess correctness, coherence, relevance, faithfulness to the audio, and reasoning quality. It is especially valuable for long-form responses, dialogue, and explanations, where objective metrics fall short. Despite high cost and limited scalability, it is still considered a highly reliable evaluation approach.

LLM-based Evaluation. To improve scalability, recent work uses LLMs as evaluators (Yang et al., 2024b; Chen et al., 2024b; Li et al., 2026b). The evaluator model scores or compares outputs based on criteria such as correctness, completeness, consistency with audio-derived context, or stepwise reasoning quality. This approach is attractive for open-ended reasoning due to flexibility and lower cost than human annotation. However, reliability depends on prompt design, bias, and the model’s ability to assess audio-grounded reasoning, so it is best viewed as a scalable proxy rather than a replacement for human evaluation.