The Long Arc of Audio: A Comprehensive Survey of Long-Context Spoken Language Models

Anonymous ACL submission

Abstract

Spoken Language Models (SLMs) are rapidly emerging as universal speech processing systems, yet their ability to handle long-context audio remains limited. Audio is temporally dense and encodes rich semantic, paralinguistic, and acoustic information, making long-range modeling particularly challenging. This survey examines long-context spoken language modeling across two settings distinguished by where long context arises: within a single turn, covering long-form audio understanding, generation, and multi-audio reasoning; and across turns in multi-turn spoken dialogue. We review representative models, benchmarks, and potential technical solutions, and discuss open challenges and promising future directions.

1 Introduction

Spoken language models (SLMs) are emerging as general-purpose systems for speech understanding and generation, supporting tasks such as automatic speech recognition, speech translation, and open-ended spoken dialogue within a unified framework (Chu et al., 2023; Tang et al., 2024; Kong et al., 2024). Following Arora et al. (2025a), we define an SLM as a multimodal large language model (LLM) that takes audio or text as input and produces audio or text as output, with audio involved in at least one side. Under this definition, SLMs range from audio-to-text systems such as SALMONN (Tang et al., 2024) and Qwen-Audio (Chu et al., 2023), to audio-to-audio systems capable of real-time spoken interaction such as Moshi (Défossez et al., 2024).

As SLMs move toward more realistic deployment, they increasingly face inputs whose audio context is far longer than what current systems handle well. This challenge is particularly severe because audio is inherently denser than text. Audio requires far more fine-grained representations than text to preserve the rich information it carries, such as semantic content, paralinguistic cues, and acoustic context. This density gap makes memory usage, computational cost, and modeling long-range dependencies fundamental bottlenecks for SLMs.

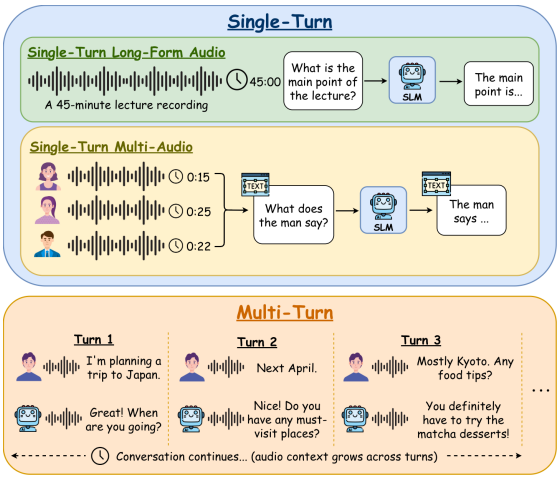

Figure 1: Single-turn and multi-turn long-context scenarios. Single-turn covers long-form audio (top) and multi-audio inputs (middle); multi-turn dialogue (bottom) accumulates audio context across turns.

Despite the severity of this challenge, long-context spoken language modeling has not been the focus of existing surveys. Prior surveys on SLMs (Latif et al., 2023; Zhang et al., 2024a; Ji et al., 2024; Cui et al., 2025; Arora et al., 2025a; Yang et al., 2025b) address the field broadly, covering model architectures, training paradigms, and evaluation across a wide range of tasks, but they do not systematically examine how SLMs cope when audio context grows long. We argue that this gap deserves dedicated attention, and that doing so requires a framework that unifies the different forms this challenge takes in practice.

This long-context challenge manifests in two broad settings, each corresponding to a different source of long audio context (Figure 1). The first arises within a single turn, where the model must process a large volume of audio at once. This can take two forms: a single continuous record-

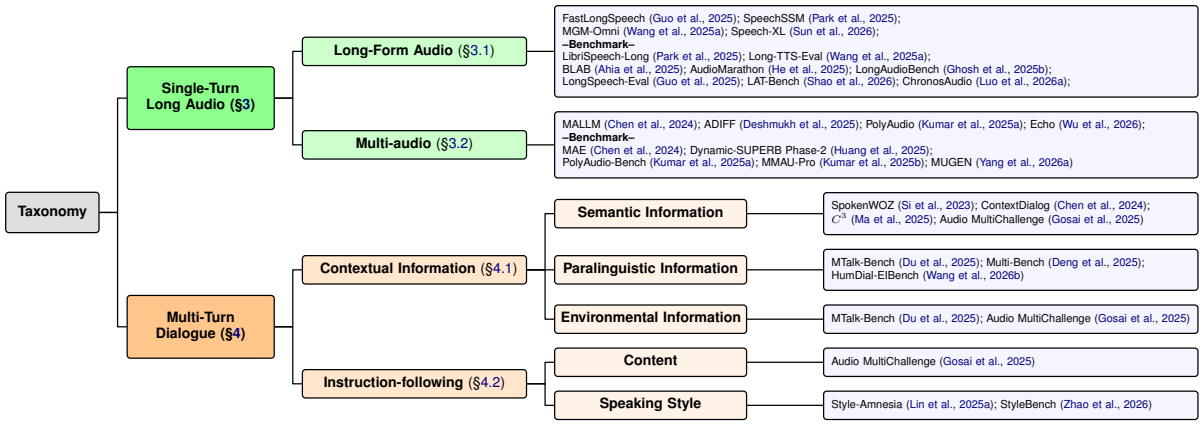

Figure 2: A taxonomy of long-context SLMs, benchmarks, and analyses. We only select representative examples here. Some are listed under multiple categories due to their multifaceted design.

ing that is itself long, such as a lecture, a podcast, or a meeting recording, or multiple independent audio segments presented together, whose aggregate duration amounts to a substantial context. The second setting arises across turns: when audio context accumulates over the course of a multi-turn conversation, the model must maintain and recall information over time. These settings are therefore not arbitrary groupings, but an exhaustive decomposition of the sources of long audio context.

In this survey, we bring these lines of work together under a single framework. Our contributions are threefold. First, we introduce a *unified taxonomy* that organizes long-context spoken language modeling into single-turn and multi-turn settings, further distinguishing single-recording and multi-audio cases within the former (Figure 2). Second, we summarize representative models, datasets, and evaluation benchmarks for each setting. Third, we review potential technical solutions across the modeling pipeline, covering low-bitrate tokenization, token compression, state-space architectures, and memory and retrieval mechanisms, and identify open challenges in this field.

2 Background: From Text LLMs to Spoken Language Models

Most modern SLMs are built on pre-trained text LLM backbones augmented with audio-processing modules, representing audio either as continuous acoustic embeddings produced by an audio encoder and projected into the LLM embedding space (Tang et al., 2024; Xu et al., 2025b; Fang et al., 2025; Lu et al., 2026; Ghosh et al., 2025a), or as discrete codec tokens processed autoregressively (Défossez et al., 2024). Their long-context capabilities are thus largely inherited from the underlying text LLMs. We review this paradigm in Appendix B.

While this paradigm transfers many strengths of text-only LLMs to the audio domain, it also introduces challenges unique to audio. The first is sequence length: audio encoders such as Whisper (Radford et al., 2023) produce features at 50 Hz, and even aggressively compressed representations, such as the Mimi codec in Moshi (Défossez et al., 2024) and the AuT representation in Qwen3-Omni (Xu et al., 2025b), still operate at 12.5 Hz, compared to only 3–4 text tokens for the semantic content of one second of speech (Défossez et al., 2024). A fixed context window thus spans far less content than in text. Second, each audio segment encodes not only semantic contents but also speaker identity, emotion, prosody, and acoustic environment. Preserving all of these jointly over long context is harder than retaining text semantics alone, and the burden grows with sequence length.

These challenges manifest differently depending on how long context arises, and we organize this survey accordingly. §3 covers single-turn settings, where audio context arises either from a single continuous recording or across multiple independent segments; §4 examines multi-turn spoken interaction; §5 surveys potential technical solutions; and §6 discusses future directions.

3 Single-Turn Long Audio Context

This section focuses on settings where a single turn presents a large volume of audio context, which we define as exceeding 30 seconds, the cap of the audio encoders underlying most SLMs (Radford et al., 2023). Such context can arise in two ways: either from a *single* continuous recording that is itself long, or from *multiple* independent audio segments whose aggregate duration amounts to a substantial

context. We address each in turn.

3.1 Long-Form Audio

We first consider the case where a single continuous recording is itself long.

Models. Work on long-form audio divides into comprehension and generation tracks, targeting understanding and production of such audio, respectively. On the comprehension side, a growing set of SLMs extend the supported input length from seconds to several minutes, including the Audio Flamingo series (Ghosh et al., 2025b,a, 2026) and the Qwen-Omni series (Xu et al., 2025a,b). Others introduce dedicated sequence compression modules to shorten the token stream prior to or within the LLM, including FastLongSpeech (Guo et al., 2025) and Speech-XL (Sun et al., 2026); we discuss such techniques in §5.1.2. Long-form generation, by contrast, has attracted considerably fewer systems, each adopting a distinct architecture. SpeechSSM (Park et al., 2025) is a textless, state-space spoken LM that produces continuations of up to 16 minutes from short prompts within a single decoding session at constant memory, while MGM-Omni (Wang et al., 2025a) builds on an omni LLM with a brain-mouth dual-track design, coupling a multimodal LLM with a separate SpeechLM to support understanding beyond 60 minutes and generation longer than 10 minutes.

Benchmarks. For comprehension, benchmarks span the few-minute scale (Ghosh et al., 2025b; Guo et al., 2025; He et al., 2025), the tens-of-minutes scale (Luo et al., 2026a; Yang et al., 2026b; Shao et al., 2026), and the longest setting in BLAB (Ahia et al., 2025) with 51-minute-average clips; for generation, LibriSpeech-Long (Park et al., 2025) and Long-TTS-Eval (Wang et al., 2025a) provide multi-minute reference evaluation. The reported failure modes converge across both tracks: performance degrades as audio length grows (Ahia et al., 2025; He et al., 2025; Luo et al., 2026a), which ChronosAudio (Luo et al., 2026a) attributes to attention diffusion and the loss of temporal locality and finds that current mitigations recover only about half of the lost performance; models additionally rely on text prompts over audio content (Ahia et al., 2025) and specialize on individual tasks at the expense of others when trained jointly (Yang et al., 2026b), while on the generation side coherence collapses past tens of seconds, with timbre consistency degrading over multi-minute outputs (Park et al., 2025; Wang et al., 2025a). Evaluation coverage also skews semantic (transcription, summarization, captioning, question answering). Temporal grounding is addressed by BLAB and LAT-Bench, but paralinguistic dimensions remain largely clip-level, leaving fine-grained prosody, emotion trajectory, and acoustic scene change essentially unevaluated.

3.2 Multi-Audio Reasoning

Long audio context need not arise from a single recording. When multiple independent audio segments are presented within a single turn, their aggregate duration can still amount to a substantial audio context, yet these segments share no inherent structure: they may differ in speaker, content, acoustic environment, or recording condition. Jointly attending to several such sources thus constitutes a distinct challenge from the single-recording case. We survey the current state of model development and evaluation for this setting below.

Models. Training approaches fall into two paradigms. The first uses *discriminative and comparative* objectives, where models learn to describe differences between audio pairs or sets, as in MALLM (Chen et al., 2024), ADIFF (Deshmukh et al., 2025), and PolyAudio (Kumar et al., 2025a). The second uses *interleaved input*, where multiple audio segments can appear at arbitrary positions alongside text rather than following a presumed modality order, as in Audio Flamingo Next (Ghosh et al., 2026) and Echo (Wu et al., 2026). Several works find that mixing single-audio data into multi-audio training improves stability (Chen et al., 2024; Kuan and Lee, 2025). Training is often capped at a small number of inputs (commonly two), though some scale further, with PolyAudio reaching up to five; generalization beyond this nonetheless remains limited, with performance degrading as the number of audios at inference grows.

Benchmarks. Existing benchmarks cover three complementary aspects. Some adapt traditional *auditory processing* tasks into instruction-following format (e.g., MAE (Chen et al., 2024), Dynamic-SUPERB Phase-2 (Huang et al., 2025), MMAU-Pro (Kumar et al., 2025b)), covering intent identification, sound event detection, and speaker verification, while ADIFF (Deshmukh et al., 2025) takes a more open-ended approach via audio difference captioning. *Cross-audio reasoning* benchmarks (e.g., MMAU-Pro, PolyAudio-Bench (Kumar et al., 2025a), MUGEN (Yang et al., 2026a))

require comparing or aggregating semantic content, temporal relations, speaker traits, prosody, and emotion across clips. MAE additionally includes *dialogue tasks* conditioned on multiple speech segments. Most benchmarks limit concurrent inputs to two or three; PolyAudio-Bench and MUGEN scale to five or six. Yang et al. (2026a) shows that performance degrades as the number of audios increases, suggesting that aggregating information across multiple sources remains a bottleneck.

4 Multi-turn Dialogue

Extended audio context also arises in multi-turn dialogue, where each turn adds a new audio segment and total duration compounds over the conversation. Multi-turn capability in current SLMs mostly emerges from training on conversational data rather than dedicated architectural design: Audio Flamingo Next (Ghosh et al., 2026) and StepAudio 2 (Wu et al., 2025) focus on curating multi-turn spoken dialogue corpora for training, while SLAM-Omni (Chen et al., 2025a) represents dialogue history as text to avoid prohibitively long audio sequences, but this discards paralinguistic and acoustic information whose long-range retention is itself at issue. Given this sparse modeling landscape, where dedicated architectural design for multi-turn spoken interaction remains largely unexplored, the question is therefore how well current systems cope as context grows, which is primarily addressed through evaluation. We organize this section around two challenges benchmarks reveal: retention of contextual information from earlier turns (§4.1), and consistent instruction-following across content and expressive style (§4.2).

4.1 Contextual Information Loss in Multi-turn Dialogue

Performance degradation over longer dialogues is well-documented in text-based LLMs, where models lose track of earlier information and fail to integrate facts scattered across turns (Kwan et al., 2024; Han et al., 2025; Qamar et al., 2025; Li et al., 2025c; Laban et al., 2025). These challenges are amplified in speech, where temporally dense audio consumes context windows far more rapidly than text. We organize the discussion around three categories of contextual information: semantic, paralinguistic, and environmental.

Semantic Information. The core challenge is long-range memory: models must retain and recall factual content such as user preferences scattered across many turns. SpokenWOZ (Si et al., 2023) is among the earliest works to demonstrate this difficulty in task-oriented dialogues. ContextDialog (Kim et al., 2025a) shows that SLMs memorize substantially less than their text counterparts, with retrieval augmentation providing only limited relief. Audio MultiChallenge (Gosai et al., 2025) corroborates these findings using human-recorded in-the-wild conversations, further showing that total audio duration is a more reliable predictor of performance degradation than turn count alone, suggesting that spoken context density is the primary bottleneck.

Paralinguistic Information. Beyond content, models must perceive *how* things are said and respond accordingly. MTalk-Bench (Du et al., 2025) finds that SLMs outperform humans on semantic tasks but underperform on paralinguistic cue perception, highlighting the relative difficulty of the latter. Multi-Bench (Deng et al., 2025) directly links failure to missing multi-turn context integration. HumDial-EIBench (Wang et al., 2026b) further exposes a systematic *text-dominance bias* by requiring models to track emotional trajectories and infer implicit causal factors across turns, finding that models treat audio primarily as a transcription proxy rather than an independent modality.

Environmental Information. Ambient sounds also carry information that should inform later responses. For example, if rain sounds appear in an earlier turn, a later request for clothing recommendations should factor in the implied weather. Yet this dimension is the least studied and hardest. MTalk-Bench (Du et al., 2025) mixes environmental sounds with speech, while Audio MultiChallenge (Gosai et al., 2025) uses in-the-wild recordings for greater realism. Both converge on the same finding: SLMs largely fail to perceive ambient sounds, performing far worse on acoustic cue integration than on semantic recall, with the bottleneck in perception rather than reasoning.

4.2 Instruction-Following

Beyond recalling previously revealed information, a further challenge in multi-turn spoken interaction is whether SLMs can *consistently* follow user-specified instructions throughout a dialogue. In spoken scenarios, such instructions can shape model behavior along two orthogonal dimensions: the *content* of responses and the *speaking style* through which responses are delivered.

Content. User instructions may impose constraints on the content or format of model responses. Audio MultiChallenge (Gosai et al., 2025) distinguishes two types: *fixed* instructions that persist across all turns (e.g., maintaining a persona), and *conditional* instructions that are triggered only upon a specific condition (e.g., switching format after a designated phrase). Empirical results show that the latter poses a substantially greater challenge, suggesting that deferred constraint application is a distinct and more demanding sub-problem.

Speaking Style. A complementary challenge is whether paralinguistic styles such as emotion, accent, volume, and speaking rate can be sustained via user instructions across turns. Lin et al. (2025a) identify *style amnesia*: SLMs comply with a style instruction in the first turn, but adherence degrades over subsequent turns. StyleBench (Zhao et al., 2026) examines a more demanding variant, *incremental* style modulation, where models must progressively amplify or diminish dimensions such as emotion, speed, volume, and pitch in response to user requests. This requires awareness of prior expressive output, and current SLMs fall substantially short, revealing a gap between single-turn compliance and dynamic multi-turn stylistic adaptation.

5 Potential Solutions for Long Context Spoken Language Modeling

Building on the challenges identified in §3 and §4, this section surveys potential solutions for long-context spoken language modeling. Although long-form, multi-audio, and multi-turn scenarios differ, they share three architectural bottlenecks: long audio sequences, the quadratic cost of self-attention, and fixed context-window limits. Since these bottlenecks are not setting-specific, the solutions can in principle apply across all settings, though many remain to be systematically validated in long-context scenarios. We organize them into input-level sequence reduction (§5.1), efficient architectures for long-sequence computation (§5.2), and memory or retrieval mechanisms for extending usable context (§5.3), as illustrated in Figure 3.

5.1 Input Processing

Input processing constitutes the most immediate bottleneck, as Transformers scale quadratically with input length. The principal strategy is to compress audio or increase per-unit information density prior to LLM ingestion.

5.1.1 Audio Representation and Tokenization

The encoding of raw audio determines the initial sequence length. Existing approaches fall into two paradigms, continuous representations and discrete tokenization, which jointly shape the length and expressiveness of the LLM input.

Continuous Representation. The continuous paradigm leverages pretrained audio encoders, such as Whisper (Radford et al., 2023), Wav2vec 2.0 (Baevski et al., 2020), WavLM (Chen et al., 2022b), and BEATs (Chen et al., 2023a), to extract dense audio features. However, these encoders operate at high frame rates of around 50 Hz, meaning that a 5-minute recording produces approximately 15,000 features.

Discrete Tokenization. The discrete paradigm quantizes speech via neural audio codecs (Défossez et al., 2023), typically through Residual Vector Quantization (RVQ) with around eight codebooks as in SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2023), with subsequent work disentangling semantic and acoustic information across codebooks, e.g., SpeechTokenizer (Zhang et al., 2024b) and Mimi (Défossez et al., 2024). However, multiple codebooks at fixed frame rates yield substantially longer sequences: at EnCodec's 75 Hz frame rate, eight codebooks amount to 180,000 tokens for a 5-minute clip. Mimi lowers the frame rate from 75 Hz to 12.5 Hz, yet a 5-minute audio clip still results in approximately 30,000 tokens.

Similar to continuous representations, the resulting sequence length scales linearly with audio duration and quickly becomes a major bottleneck for long-context modeling, motivating the compression techniques discussed next.

5.1.2 Sequence Compression

Downstream of the tokenizer, sequence compression further shortens the speech token stream via merging, dropping, or downsampling at the token or KV cache level. We organize existing approaches by increasing selectivity.

Fixed-Rate Downsampling. The simplest strategy uniformly downsamples encoder output via pooling or strided convolutions. For instance, LLaMA-Omni (Fang et al., 2025) concatenates every five Whisper frames, and Freeze-Omni (Wang et al., 2025b) stacks downsampling convolutions before LLM prefill.

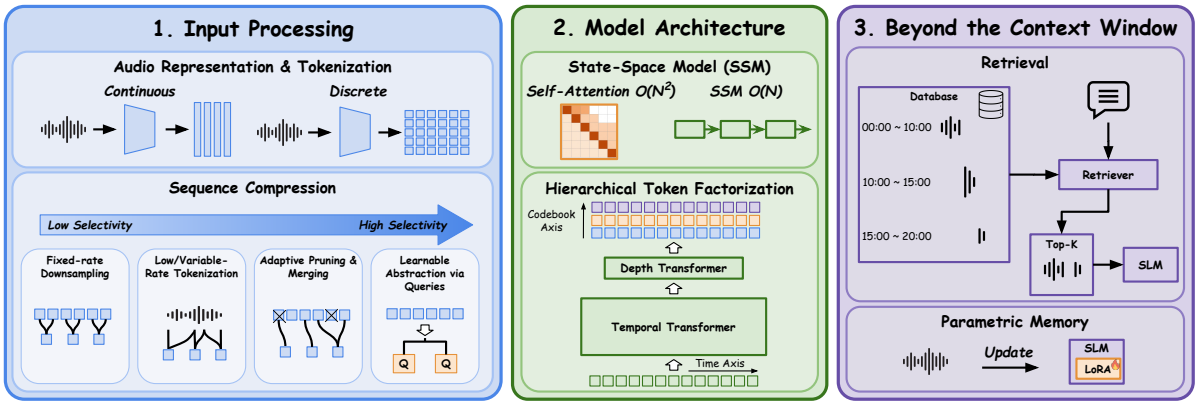

Figure 3: Overview of the potential solution.

Low-Frame-Rate Speech Tokenization. A complementary line compresses at the tokenizer itself, allowing the SLM to operate on shorter sequences end-to-end. LFSC (Casanova et al., 2025) reaches 21.5 Hz via finite scalar quantization; DualCodec (Li et al., 2025a) attains 12.5 Hz by coupling SSL semantics with waveform reconstruction; TaDiCodec (Wang et al., 2025c) further pushes a single-codebook diffusion tokenizer to 6.25 Hz; and syllable-level tokenization (Lee et al., 2026b) compresses to 4–5 Hz.

Variable-Frame-Rate Speech Tokenization. Since speech is non-uniform in information density, variable-rate tokenizers allocate tokens adaptively to content. CodecSlime (Wang et al., 2026a) merges temporally similar frames from a fixed-rate codec; VARSTok (Zheng et al., 2026) applies density-peak clustering with implicit duration coding; FlexiCodec (Li et al., 2025b) refines this merging to reach lower rates; and DyCAST (Della Libera et al., 2026) employs character-level soft alignment to produce semantically aligned tokens. TASTE (Tseng et al., 2026) aligns speech tokens with text via cross-attention, achieving a low, dynamic frame rate.

Adaptive Token Pruning and Merging. Adaptive methods score and retain informative tokens or merge redundant ones. SpeechPrune (Lin et al., 2025b) adopts a training-free two-phase procedure based on speech-text similarity and attention scores, remaining robust at 80% pruning; HeadRouter (He et al., 2026) exploits the semantic-versus-acoustic selectivity of attention heads via per-sample routing, surpassing unpruned baselines while discarding 30% of tokens. Complementary analyses further show that cosine-similarity merging aggregates adjacent tokens into word- or phrase-level units with minimal performance loss (Xiang et al., 2026).

Learnable Abstraction via Queries. Learnable modules abstract acoustic content into compact low-rate token sequences via trainable queries. SALMONN (Tang et al., 2024) applies a window-level Q-Former and reducing the 50 Hz encoder output to roughly 3 tokens/sec. MMS-LLaMA (Yeo et al., 2025) allocates queries dynamically by speech rate, achieving 86% reduction at 3.5 tokens/sec. HFQ-Former (Lee et al., 2026a) compresses hierarchically across three temporal scales, reaching around 93% reduction at 1.67 tokens/sec.

Content-Aware Dynamic Compression. Linguistic content can also guide compression. FastLongSpeech (Guo et al., 2025) adopts a CTC-based iterative fusion strategy that estimates text density and merges representations accordingly, coupled with dynamic-ratio training. Compressed-to-fine modeling (Liu et al., 2025) preserves fine-grained local codec tokens while compressing long-range context into coarse summaries.

5.2 Model Architecture

5.2.1 State-Space Models

Even after compression, multi-minute audio remains prohibitive for self-attention due to its quadratic complexity. Selective State-Space Models (SSMs), notably Mamba (Gu and Dao, 2023) and Mamba-2 (Dao and Gu, 2024), achieve linear-time training and constant-memory inference through input-dependent recurrence over a fixed-size hidden state. These properties align well with speech: long-form decoding avoids the unbounded growth of the KV cache in Transformers.

Existing audio SSM work falls into three integration patterns, distinguished by the extent to which the Transformer architecture is retained.

Pure SSM Backbones. Pure SSM backbones can match Transformers on audio understanding while scaling more favorably with sequence length (Bhati et al., 2024), and recent work extends this to audio-language modeling with a Mamba-2 backbone (Lee et al., 2026c). These designs fully exploit the linear-time and constant-memory profile of state-space recurrence, but forgo the precise token-level recall that attention provides.

Hybrid Designs. A second line interleaves attention layers with state-space blocks, retaining attention's local precision while delegating long-range propagation to selective state spaces. SpeechSSM (Park et al., 2025), introduced in §3.1, exemplifies this approach, sustaining coherent generation for up to 16 minutes at constant memory cost.

Attention-Augmented SSM. A third, less explored direction keeps an SSM as the primary backbone but attaches narrow attention modules for task-driven retrieval over the recurrent state, appealing when the recurrent state already encodes long-range context and attention is needed only for selective lookup. For instance, SQuBa (Eom et al., 2025) applies querying attention over a Mamba context for long-speech summarization.

5.2.2 Hierarchical Token Factorization

Multi-codebook tokenizers such as Mimi (Défossez et al., 2024) pose a challenge for long-context modeling. With eight codebooks at a frame rate of 12.5 Hz, naively flattening them into a single sequence yields 100 tokens per second of audio. This expands the sequence length by 8× and shrinks the audio that fits within a fixed context window. Moshi addresses this by decomposing generation across two Transformers. A large *Temporal Transformer* processes the sequence at the frame rate of 12.5 Hz. At each step, it produces a context embedding that captures long-range dependencies across time. Conditioned on this embedding, a smaller *Depth Transformer* then autoregressively generates the eight codebook tokens within that frame. Since the Temporal Transformer operates on one embedding per frame rather than the flattened sequence, it still runs at the 12.5 Hz frame rate. This makes multi-codebook tokenization more efficient for long-context spoken language modeling.

5.3 Beyond the Context Window

A complementary line of work avoids relying solely on larger context windows by preserving past information outside the active context or in compact memory representations, thereby reducing the need to attend over the entire history. Existing methods can be broadly divided into *retrieval-based* and *parametric-memory* approaches.

Retrieval. One family of methods externalizes audio into a queryable repository, retrieving only query-relevant segments to relieve the SLM context-length limit. In spoken dialogue, WavRAG (Chen et al., 2025b) and SpeechRAG (Min et al., 2025) bypass ASR by directly indexing and retrieving audio. Stream RAG (Arora et al., 2025b) performs retrieval in parallel with user speech, while MoshiRAG (Chien et al., 2026) retrieves asynchronously after response onset to reduce latency.

For long-form audio understanding, CLSR (Hu et al., 2026) uses a contrastive language-speech retriever to identify question-relevant segments for spoken QA. LongAudio-RAG (Vakada et al., 2026) indexes long audio as timestamped event records in an SQL database and retrieves only the relevant events at inference. This enables the hour-scale processing scenarios discussed in §3.

Parametric Memory. A second family encodes past audio directly into the model's weights. Dynamic Parameter Memory (Mai et al., 2025) distills each utterance of a long dialogue into a temporary LoRA module that is updated online, so dialogue history is carried by adapter parameters rather than by tokens in the context window.

6 Challenges and Future Directions

The preceding sections surveyed the current landscape of long-context spoken language modeling, covering representative models, benchmarks, and potential solutions. Here we discuss key open challenges and highlight promising future directions.

6.1 Generalization over Scale

A unifying thread across all three settings is the problem of *generalization over scale*. Since models cannot be trained on every possible audio duration, number of concurrent audios, or conversation length, they must extrapolate gracefully beyond training scale. Yet as reviewed in §3–4, this generalization is systematically poor, and particularly underexplored on the generation side (§3.1).

This failure has implications beyond long-context settings. For instance, reasoning over mul-

tiple audios is structurally related to auditory in-context learning (Chen et al., 2024), in which a model extracts a pattern from audio exemplars and applies it to a new query, a capability current SLMs demonstrably lack (Yang et al., 2025c; Piao et al., 2026). Developing SLMs with robust generalization across these dimensions, on both the understanding and generation sides, is a prerequisite for general-purpose systems, and one of the most important directions for future research.

6.2 Effective Context Length in SLMs

While §6.1 addresses generalization to scales beyond training, even within claimed context limits, LLMs may not reliably utilize the full span they support. The concept of effective context length, how far into a long input a model can actually retain and use information, has proven essential for honest capability characterization in language models (Kamradt, 2023; Hsieh et al., 2024; An et al., 2025). Studies have shown that many models fail well within their nominal limits, with effective context length often falling below 50% of the claimed capacity (Hsieh et al., 2024). However, no equivalent analysis exists for long-context SLMs, and whether and how degradation varies across the context remains largely uncharacterized; see Appendix E for how it is measured and the challenges of adapting such analysis to audio.

An open question is whether different types of audio information, such as semantic content, paralinguistic cues, and environmental context, have distinct effective context lengths within the same model. If different information types decay at different rates, long-context performance may finally be bottlenecked by the information type with the shortest effective context length. Understanding and extending these information-specific limits remains an important direction for future research.

6.3 Retrieval Beyond Semantics for Long-Context Spoken Dialogue

Retrieval-augmented approaches have emerged as a promising direction for extending the effective context of SLMs beyond the model's input window, as discussed in §5.3. However, existing audio RAG systems (Chen et al., 2025b; Min et al., 2025; Arora et al., 2025b) are designed primarily around semantic matching; while a few extend to acoustic information (Vakada et al., 2026), paralinguistic attributes remain largely unaddressed.

This limitation is particularly consequential in long-context settings. As discussed in §4.1, paralinguistic signals are harder to retain than semantic content, and more likely to be lost before retrieval is even invoked. Semantic-only retrieval thus cannot recover what is already lost, and may compound the loss by surfacing context that is semantically relevant but paralinguistically mismatched. Developing retrieval models and evaluation protocols where what is retrieved depends not only on what was said but on how it was said remains an important open direction.

6.4 Safety of Long-Context SLMs

As SLMs evolve toward more general and capable systems, ensuring their safety and robustness is increasingly important. Prior work has shown that SLMs are vulnerable to diverse audio-centric attack vectors absent in text (Kang et al., 2025; Feng et al., 2026a; Li et al., 2026), even within a single short utterance (see Appendix F for a detailed review).

Long-context settings compound these risks: extended audio gives attackers more room to embed harmful signals stealthily, across either conversational turns or a single long recording. Evidence from text reinforces this concern, where long-context inputs open new attack surfaces (Kim et al., 2025b; Feng et al., 2026b). Yet with the exception of Shen et al. (2024), who study a two-turn fictional persuasion setting against GPT-4o, nearly all existing work evaluates safety under a single isolated utterance of a few seconds; Ivry and Watanabe (2026) examine SLMs in multi-turn dialogues, but only as external judges rather than interactive agents. Safety in long-context SLMs thus remains largely unexplored and represents an important direction for future research.

7 Conclusion

Despite rapid progress in SLMs, handling long audio context, such as continuous long recordings, multiple audios, or extended dialogues, remains a fundamental bottleneck. This survey offers a unified perspective on the challenge, reviewing representative models, benchmarks, and potential technical solutions in the field. Building on this review, we consolidate recurring limitations of current systems and identify key open challenges and promising directions for long-context spoken language modeling. We hope it serves as a useful roadmap for future research toward truly long-context spoken language systems.

Limitations

We acknowledge certain limitations in this survey. Most notably, our review captures the state of long-context spoken language modeling at the time of writing. Given the rapid evolution of this field, new models, benchmarks, and techniques are expected to emerge after publication. As the landscape expands and shifts, our proposed taxonomy will naturally require adaptation to encompass emerging tasks and applications. We will monitor these advancements and update our framework to align with future progress.

Second, we focus specifically on long context spoken language modeling as a distinct challenge for SLMs. Broader aspects of SLM research, such as general capability modeling and evaluation, fall outside the scope of this survey and have been covered by complementary reviews (Latif et al., 2023; Zhang et al., 2024a; Ji et al., 2024; Cui et al., 2025; Arora et al., 2025a; Yang et al., 2025b; Luo et al., 2026b). We refer readers interested in these aspects to those works.

Finally, as a survey paper, our contribution lies in consolidating and synthesizing the conclusions reported across prior studies, offering a unified perspective on the field. Building on this foundation, the open challenges identified in §6 call for controlled comparisons under unified experimental settings, which we leave to future work.

Ethics Statement

This work is a survey and does not involve new data collection, human subjects, or model training; we have therefore identified no direct ethical concerns arising from the paper itself. However, we acknowledge that the long-context spoken language modeling capabilities surveyed here may be misused, for example through the analysis of long recordings captured without consent or the generation of long-form deepfake speech, in addition to the safety vulnerabilities discussed in §6.4. We do not endorse such uses, and we hope this survey encourages the community to advance long-context SLMs alongside corresponding safeguards.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text. *Preprint*, arXiv:2301.11325.

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, Siddhant Arora, Shuyue Stella Li, Vishal Puttagunta, Mofetoluwa Adeyemi, Charishma Buchireddy, Ben Walls, Noah Bennett, Shinji Watanabe, Noah A. Smith, Yulia Tsvetkov, and Sachin Kumar. 2025. Blab: Brutally long audio bench. *Preprint*, arXiv:2505.03054.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2025. Why does the effective context length of llms fall short? In *International Conference on Learning Representations*, volume 2025, pages 54791–54810.

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. *Transactions on Machine Learning Research*.

Siddhant Arora, Haidar Khan, Kai Sun, Xin Luna Dong, Sajal Choudhary, Seungwhan Moon, Xinyuan Zhang, Adithya Sagar, Surya Teja Appini, Kaushik Patnaik, Sanat Sharma, Shinji Watanabe, Anuj Kumar, Ahmed Aly, Yue Liu, Florian Metze, and Zhaojiang Lin. 2025b. Stream rag: Instant and accurate spoken dialogue systems with streaming tool usage. *Preprint*, arXiv:2510.02044.

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Saurabhchand Bhati, Yuan Gong, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. State-space large audio language models. *arXiv preprint arXiv:2411.15685*.

Edresson Casanova, Ryan Langman, Paarth Neekhara, Shehzeen Hussain, Jason Li, Subhankar Ghosh, Ante Jukić, and Sang-gil Lee. 2025. Low frame-rate speech codec: a codec designed for fast high-quality speech llm training and inference. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Guangke Chen, Fu Song, Zhe Zhao, Xiaojun Jia, Yang Liu, Yanchen Qiao, Weizhe Zhang, Weiping Tu, Yuhong Yang, and Bo Du. 2026. Audiojailbreak: Jailbreak attacks against end-to-end large audio-language models. *IEEE Transactions on Dependable and Secure Computing*, 23(3):6085–6102.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022a. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023a. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, pages 5178–5193. PMLR.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

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. 2025a. 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.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025b. Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12505–12523.

Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D'Haro, Robby T. Tan, and Haizhou Li. 2024. 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, Miami, Florida, USA. Association for Computational Linguistics.

Hao Cheng, Erjia Xiao, Jing Shao, Yichi Wang, Le Yang, Chao Shen, Philip Torr, Jindong Gu, and Renjing Xu. 2025. Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models. *Advances in Neural Information Processing Systems*, 38.

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.

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.

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

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

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025. 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.

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Transactions on Machine Learning Research*. Featured Certification, Reproducibility Certification.

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*.

Luca Della Libera, Cem Subakan, and Mirco Ravanelli. 2026. Beyond fixed frames: Dynamic character-aligned speech tokenization. *arXiv preprint arXiv:2601.23174*.

Yayue Deng, Guoqiang Hu, Haiyang Sun, Xiangyu Zhang, Haoyang Zhang, Fei Tian, Xuerui Yang, Gang Yu, and Eng Siong Chng. 2025. Multi-bench: A multi-turn interactive benchmark for assessing emotional intelligence ability of spoken dialogue models. *arXiv preprint arXiv:2511.00850*.

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

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

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

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.

Yuhao Du, Qianwei Huang, Guo Zhu, Zhanchen Dai, Shunian Chen, Qiming Zhu, Le Pan, Minghao Chen, Yuhao Zhang, Li Zhou, Benyou Wang, and Haizhou Li. 2025. Mtalk-bench: Evaluating speech-to-speech models in multi-turn dialogues via arena-style and rubrics protocols. *Preprint*, arXiv:2508.18240.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340. IEEE.

SooHwan Eom, Jay Shim, Eunseop Yoon, Hee Suk Yoon, Hyeonmok Ko, Mark A. Hasegawa-Johnson, and Chang D. Yoo. 2025. SQuba: Speech mamba language model with querying-attention for efficient summarization.

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*.

Bo-Han Feng, Chien-Feng Liu, Yu-Hsuan Li Liang, Chih-Kai Yang, Szu-Wei Fu, Zhehuai Chen, Ke-Han Lu, Sung-Feng Huang, Chao-Han Huck Yang, Yu-Chiang Frank Wang, Yun-Nung Chen, and Hung-Yi Lee. 2026a. Emotional damage: Investigating safety vulnerabilities of large audio-language models under speaker emotional variations. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17792–17796.

Mingqian Feng, Xiaodong Liu, Weiwei Yang, Jialin Song, Xuekai Zhu, Chenliang Xu, and Jianfeng Gao. 2026b. SEMA: Simple yet effective learning for multi-turn jailbreak attacks. In *The Fourteenth International Conference on Learning Representations*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

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.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. 2026. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *Preprint*, arXiv:2604.10905.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025a. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*, volume 38, pages 41819–41886. Curran Associates, Inc.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025b. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*.

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. 2025. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. *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, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Shoutao Guo, Shaolei Zhang, Qingkai Fang, Zhengrui Ma, Min zhang, and Yang Feng. 2025. Fastlongspeech: Enhancing large speech-language mod-

els for efficient long-speech processing. In *Advances in Neural Information Processing Systems*, volume 38, pages 164996–165024. Curran Associates, Inc.

Chi Han, Xin Liu, Haodong Wang, Shiyang Li, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Qingyu Yin, Liang Qiu, Changlong Yu, Yifan Gao, Zheng Li, Bing Yin, Jingbo Shang, and Heng Ji. 2025. Can language models follow multiple turns of entangled instructions? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25445–25460, Suzhou, China. Association for Computational Linguistics.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis CONNEAU, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 63483–63501. Curran Associates, Inc.

Peize He, Yaodi Luo, Xiaoqian Liu, Xuyang Liu, Jiahang Deng, Yaosong Du, Bangyu Li, Xiyan Gui, Yuxuan Chen, and Linfeng Zhang. 2026. Headrouter: Dynamic head-weight routing for task-adaptive audio token pruning in large audio language models. *arXiv preprint arXiv:2604.23717*.

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, Kai Li, Zhifei Liu, Weijia Li, Cunxiang Wang, Conghui He, and Linfeng Zhang. 2025. Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. *Preprint*, arXiv:2510.07293.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

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

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. WavLLM: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, Miami, Florida, USA. Association for Computational Linguistics.

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. 2025. 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*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, and 5 others. 2024. Qwen2.5-coder technical report. *Preprint*, arXiv:2409.12186.

Amir Ivry and Shinji Watanabe. 2026. Lalm-as-a-judge: Benchmarking large audio-language models for safety evaluation in multi-turn spoken dialogues. *arXiv preprint arXiv:2602.04796*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *Preprint*, arXiv:2212.12017.

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

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673.

Gregory Kamradt. 2023. Needle In A Haystack - pressure testing LLMs. *Github*.

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*.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen,

Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

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.

Heeseung Kim, Che Hyun Lee, Sangkwon Park, Jiheum Yeom, Nohil Park, Sangwon Yu, and Sungroh Yoon. 2025a. Does your voice assistant remember? analyzing conversational context recall and utilization in voice interaction models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8984–9014, Vienna, Austria. Association for Computational Linguistics.

Sangyeop Kim, Yohan Lee, Yongwoo Song, and Kimin Lee. 2025b. What really matters in many-shot attacks? an empirical study of long-context vulnerabilities in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2043–2063, Vienna, Austria. Association for Computational Linguistics.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

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

Chun-Yi Kuan and Hung-yi Lee. 2025. From alignment to advancement: Bootstrapping audio-language alignment with synthetic data. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4604–4619.

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, Hyeongon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, and 15 others. 2025b. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *Preprint*, arXiv:2508.13992.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

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

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

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States. Association for Computational Linguistics.

Junseok Lee, Sangyong Lee, and Chang-Jae Chun. 2026a. Fastslm: Hierarchical frame q-former for effective speech modality adaptation. *arXiv preprint arXiv:2601.06199*.

Nicholas Lee, Cheol Jun Cho, Alan W Black, and Gopala K Anumanchipalli. 2026b. Scaling spoken language models with syllabic speech tokenization. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1816–1820. IEEE.

Taehan Lee, Jaehan Jung, and Hyukjun Lee. 2026c. Sam: A mamba-2 state-space audio-language model. *Preprint*, arXiv:2509.15680.

Hongyi Li, Chengxuan Zhou, Chu Wang, Sicheng Liang, Yanting Chen, Qinlin Xie, Jiawei Ye, and Jie Wu. 2026. Stylebreak: Revealing alignment vulnerabilities in large audio-language models via style-aware audio jailbreak. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37591–37599.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. 2025a. DualCodec: A Low-Frame-Rate, Semantically-Enhanced Neural Audio Codec for Speech Generation. In *Interspeech 2025*, pages 4883–4887.

Jiaqi Li, Yao Qian, Yuxuan Hu, Leying Zhang, Xiaofei Wang, Heng Lu, Manthan Thakker, Jinyu Li, Sheng Zhao, and Zhizheng Wu. 2025b. Flexicodec: A dynamic neural audio codec for low frame rates. *arXiv preprint arXiv:2510.00981*.

Jinnan Li, Jinzhe Li, Yue Wang, Yi Chang, and Yuan Wu. 2025c. StructFlowBench: A structured flow benchmark for multi-turn instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9322–9341, Vienna, Austria. Association for Computational Linguistics.

Yu-Xiang Lin, Cheng-Han Chiang, and Hung-yi Lee. 2025a. Style amnesia: Investigating speaking style degradation and mitigation in multi-turn spoken language models. *arXiv preprint arXiv:2512.23578*.

Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai Helen Li, and Yiran Chen. 2025b. Speechprune: Context-aware token pruning for speech information retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Wenrui Liu, Qian Chen, Wen Wang, Guanrou Yang, Weiqin Li, Minghui Fang, Jialong Zuo, Xiaoda Yang, Tao Jin, Jin Xu, Zemin Liu, Yafeng Chen, Jionghao Bai, and Zhifang Guo. 2025. Speech token prediction via compressed-to-fine language modeling for speech generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 10632–10641, New York, NY, USA. Association for Computing Machinery.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025. Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, Yi-Cheng Lin, Yu-Xiang Lin, Chi-An Fu, Chun-Yi Kuan, Wenze Ren, Xuan-jun Chen, Wei-Ping Huang, En-Pei Hu, Tzu-Quan Lin, and 9 others. 2026. Desta2.5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2062–2076.

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

Kaiwen Luo, Zhenhong Zhou, Leo Wang, Liang Lin, Yang Xiao, Tianyu Shao, Yuanhe Zhang, Yuxuan Li, Miao Yu, Kailin Lyu, Jiaming Zhang, Dongrui Liu, Li Sun, Yueming Wu, Kai Li, Ting Dang, Xiaojun Jia, Rohan Kumar Das, Xinfeng Li, and 15 others. 2026b. A survey of large audio language models: Generalization, trustworthiness, and outlook. *arXiv preprint arXiv:2605.20266*.

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

Jialong Mai, Xiaofen Xing, Yawei Li, Weidong Chen, Zhipeng Li, Jingyuan Xing, and Xiangmin Xu. 2025. Dynamic parameter memory: Temporary lora-enhanced llm for long-sequence emotion recognition in conversation. *arXiv preprint arXiv:2507.09076*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Do June Min, Karel Mundnich, Andy Lapastora, Erfan Soltanmohammadi, Srikanth Ronanki, and Kyu Han. 2025. Speech retrieval-augmented generation without automatic speech recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. 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.

Se Jin Park, Julian Salazar, Aren Jansen, Keisuke Kinoshita, Yong Man Ro, and Rj Skerry-Ryan. 2025. Long-form speech generation with spoken language models. In *International Conference on Machine Learning*, pages 48245–48261. PMLR.

Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Yandre Maldonado e Gomes da Costa, Valéria Delisandra Feltrim, and Marcos Aurélio Domingues. 2020. Music4all: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pages 31932–31951.

Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Zeren Luo, Jingyi Zheng, Wenhan Dong, Xinlei He, Xuechao Wang, Yingjie Xue, Shengmin Xu, and Xinyi Huang. 2026. JALMBench: Benchmarking jailbreak vulnerabilities in audio language models. In *The Fourteenth International Conference on Learning Representations*.

Raghuveer Peri, Sai Muralidhar Jayanthi, Srikanth Ronanki, Anshu Bhatia, Karel Mundnich, Saket Dingliwal, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Srikanth Vishnubhotla, Daniel Garcia-Romero, Sundararajan Srinivasan, Kyu Han, and Katrin Kirchhoff. 2024. SpeechGuard: Exploring the adversarial robustness of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10018–10035, Bangkok, Thailand. Association for Computational Linguistics.

Yen-Ting Piao, Jay Chiehen Liao, Wei-Tang Chien, Toshiki Ogimoto, Shang-Tse Chen, Yun-Nung Chen, Chun-Yi Lee, and Shao-Yuan Lo. 2026. Alice: A multifaceted evaluation framework of large audio-language models' in-context learning ability. *arXiv preprint arXiv:2603.20433*.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs understand dialogues? a case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237, Vienna, Austria. Association for Computational Linguistics.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mingchen Shao, Hang Su, Wenjie Tian, Bingshen Mu, Zhennan Lin, Lichun Fan, Zhenbo Luo, Jian Luan, and Lei Xie. 2026. Listening with time: Precise temporal awareness for long-form audio understanding. *arXiv preprint arXiv:2604.22245*.

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

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36:39088–39118.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Haoqin Sun, Chenyang Lyu, Shiwan Zhao, Xuanfan Ni, Xiangyu Kong, Longyue Wang, Weihua Luo, and Yong Qin. 2026. Speech-xl: Towards long-form speech understanding in large speech language models. *arXiv preprint arXiv:2602.05373*.

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 *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024a. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Step-Audio Team. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Liang-Hsuan Tseng, Yi-Chang Chen, Kuan Yi Lee, Da shan Shiu, and Hung yi Lee. 2026. TASTE: Text-aligned speech tokenization and embedding for spoken language modeling. In *The Fourteenth International Conference on Learning Representations*.

Naveen Vakada, Kartik Hegde, Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. 2026. Longaudio-rag: Event-grounded question answering over multi-hour long audio. *arXiv preprint arXiv:2602.14612*.

Chengyao Wang, Zhisheng Zhong, Bohao Peng, Senqiao Yang, Yuqi Liu, Haokun Gui, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. 2025a. Mgm-omni: Scaling omni llms to personalized long-horizon speech. *arXiv preprint arXiv:2509.25131*.

Hankun Wang, Yiwei Guo, Chongtian Shao, Bohan Li, and Kai Yu. 2026a. Codecslime: Temporal redundancy compression of neural speech codec via dynamic frame rate. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17017–17021. IEEE.

Shuiyuan Wang, Zhixian Zhao, Hongfei Yue, Chengyou Wang, Shuai Wang, Hui Bu, Xin Xu, and Lei Xie. 2026b. Humdial-eibench: A human-recorded multi-turn emotional intelligence benchmark for audio language models. *arXiv preprint arXiv:2604.11594*.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2025b. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. In *International Conference on Machine Learning*, pages 63345–63354. PMLR.

Yuancheng Wang, Dekun Chen, Xueyao Zhang, Junan Zhang, Jiaqi Li, and Zhizheng Wu. 2025c. Tadicodec: Text-aware diffusion speech tokenizer for speech language modeling. *Advances in Neural Information Processing Systems*, 38:147494–147523.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Daiqing Wu, Xuan Zhang, Dongbao Yang, Jiashu Yao, Longfei Chen, Qingsong Liu, Sicheng Zhao, Can Ma, Yangyang Kang, and Yu Zhou. 2026. Echo: Towards advanced audio comprehension via audio-interleaved reasoning. In *The Fourteenth International Conference on Learning Representations*.

Bajian Xiang, Tingwei Guo, Xuan Chen, and Yang Han. 2026. Do we need distinct representations for every speech token? unveiling and exploiting redundancy in large speech language models. *arXiv preprint arXiv:2604.06871*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, and 2 others. 2024. Effective long-context scaling of foundation 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 4643–4663, Mexico City, Mexico. Association for Computational Linguistics.

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, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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 10144–10170, Suzhou, China. Association for Computational Linguistics.

Chih-Kai Yang, Yen-Ting Piao, Tzu-Wen Hsu, Szu-Wei Fu, Zhehuai Chen, Ke-Han Lu, Sung-Feng Huang, Chao-Han Huck Yang, Yu-Chiang Frank Wang, Yun-Nung Chen, and Hung-yi Lee. 2025c. Sake: Towards editing auditory attribute knowledge of large audio-language models. *arXiv preprint arXiv:2510.16917*.

Chih-Kai Yang, Yun-Shao Tsai, Yu-Kai Guo, Ping-Le Tsai, Yen-Ting Piao, Hung-Wei Chen, Ting-Lin Hsiao, Yun-Man Hsu, Ke-Han Lu, and Hung-yi Lee. 2026a. Mugen: Evaluating and improving multi-audio understanding of large audio-language models. *arXiv preprint arXiv:2603.09714*.

Fei Yang, Xuanfan Ni, Renyi Yang, Jiahui Geng, Qing Li, Chenyang Lyu, Yichao Du, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2026b. Longspeech: A scalable benchmark for transcription, translation and understanding in long speech. *arXiv preprint arXiv:2601.13539*.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. 2025d. 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.

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. 2025. Mms-llama: Efficient llm-based audio-visual speech recognition with minimal multimodal speech tokens. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20724–20735.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

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

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024b. Speechtokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations*, volume 2024, pages 31798–31818.

Haishu Zhao, Aokai Hao, Yuan Ge, Zhenqiang Hong, Tong Xiao, and Jingbo Zhu. 2026. Stylebench: Evaluating speech language models on conversational speaking style control. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17542–17546. IEEE.

Rui-Chen Zheng, Wenrui Liu, Hui-Peng Du, Qinglin Zhang, Chong Deng, Qian Chen, Wen Wang, Yang Ai, and Zhen-Hua Ling. 2026. Say more with less: Variable-frame-rate speech tokenization via adaptive clustering and implicit duration coding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35021–35029.

A Disclosure of Generative AI Usage

In this work, generative AI tools were employed exclusively for improving writing clarity and polishing the language.

B Context Extension in Text LLMs

The modern concept of an SLM traces back to the textless NLP paradigm. Early work such as GSLM (Lakhotia et al., 2021) demonstrates that language modeling can be performed directly on speech by converting audio into discrete units derived from self-supervised speech representations. Subsequent studies extended this framework with explicit prosody modeling (Kharitonov et al., 2022) and multi-speaker conversational generation (Nguyen et al., 2023). Despite these advances, the linguistic quality and coherence of speech generated by purely speech-based SLMs remained far below the text produced by contemporary LLMs.

TWIST (Hassid et al., 2023) shows that initializing a speech LM from a pre-trained text-only LLM consistently outperforms training from scratch. Motivated by this finding, most modern SLMs are built upon pre-trained text backbones and augmented with speech-processing modules (Chu et al., 2023; Tang et al., 2024; Chu et al., 2024; Ghosh et al., 2025b; Lu et al., 2025; Team, 2025; Fang et al., 2025; Xu et al., 2025a,b; Ghosh et al., 2025a; Lu et al., 2026). Rather than learning linguistic capabilities solely from audio, these models inherit much of their language knowledge from the underlying text LLM. For example, the Qwen-Audio (Chu et al., 2023, 2024; Xu et al., 2025a,b) and Audio Flamingo (Ghosh et al., 2025b,a) families are built on the Qwen series (Bai et al., 2023; Hui et al., 2024; Yang et al., 2025a). Similarly, LLaMA-Omni (Fang et al., 2025) and DeSTA2.5-Audio (Lu et al., 2026) adopt LLaMA 3.1 (Grattafiori et al., 2024) as LLM backbone.

Consequently, the long-context capabilities of SLMs are closely tied to those of their underlying text LLMs, making advances in long-context LLMs an important foundation for SLMs. However, speech presents unique challenges: audio sequences are substantially longer than text, and preserving their rich acoustic information adds further burden. Improving long-context SLMs thus requires not only techniques inherited from LLMs but also methods specifically designed for audio.

We briefly review the mechanisms enabling long-context modeling in modern LLMs, as these serve

as the foundation of most SLMs. Long-context modeling machinery in contemporary LLMs has become largely standardized around Rotary positional embedding (RoPE) (Su et al., 2024), the de facto positional encoding across many LLM families, including LLaMA (Touvron et al., 2023a,b; Grattafiori et al., 2024), Qwen (Bai et al., 2023; Yang et al., 2025a), and Gemma (Team et al., 2024a,b). RoPE's relative-rotation formulation is the foundation on which many context-extension method are built, including Position Interpolation (Chen et al., 2023b), YaRN (Peng et al., 2024), and LongRoPE (Ding et al., 2024). Long-context capability itself is commonly instilled through continued pretraining on progressively longer sequences (Xiong et al., 2024).

C Representative SLMs

Table 1 presents an overview of recent representative SLMs, illustrating the evolutionary trajectory of the field. In this table, we summarize key information from the original papers, including the input and output modalities, the LLM backbone, its maximum context length, the audio token rate before entering the LLM, the maximum supported audio length, and the audio tokenizer used. Note that we only report values provided in the original papers, with unavailable data indicated by a hyphen in the table.

As shown in the table, early SLMs such as Pengi (Deshmukh et al., 2023) support up to 7 seconds of audio during training, and the audio is encoded into 40 tokens regardless of its duration. That is, Pengi operates at a token rate of roughly 5.71 Hz. Similarly, SALMONN (Tang et al., 2024) employs a window-level Q-Former that compresses the 50 Hz outputs of the BEATs (Chen et al., 2023a) and Whisper (Radford et al., 2023) encoders down to about 3 tokens/sec. These two examples demonstrate that audio can be represented at a very low token rate for developing an SLM.

However, recent SLMs such as Qwen3-Omni (Xu et al., 2025b) and Audio Flamingo Next (Ghosh et al., 2026) still choose to adopt higher token rates of 12.5 Hz and 25 Hz, respectively. Consequently, a higher token rate appears to be advantageous for spoken language modeling. This observation suggests a trade-off between context efficiency and information preservation. A higher token rate may retain richer acoustic and linguistic details, which could benefit downstream tasks despite increasing context consumption. This motivates the adoption and investigation of low-frame-rate tokenization for spoken language modeling, as discussed in §5.1.

D Common Training Corpora for Long-Context SLMs

In §3–4, we survey the current landscape of long-context SLMs, with this section providing supplementary details on common training corpora.

Single-Turn Long Audio Context. Table 2 summarizes representative corpora for long-form audio training. For multi-audio training, rather than relying on natively multi-audio corpora, existing works (Chen et al., 2024; Deshmukh et al., 2025; Kumar et al., 2025a) construct multi-audio training data from single-audio sources using task-specific strategies, such as pairing audio clips and generating difference explanations as described in §3.2. As no widely adopted multi-audio corpora exist, Table 3 instead summarizes commonly used single-audio corpora to illustrate the audio sources underlying current multi-audio models.

Multi-Turn Dialogue. Unlike long-form and multi-audio settings, multi-turn spoken dialogue training data is largely synthesized independently by each group, with no commonly shared corpora emerging across works. Typical construction strategies include curating multi-turn conversations grounded in existing audios, as in AF-Chat (Ghosh et al., 2025a), a dataset used by Audio Flamingo 3 (Ghosh et al., 2025a) and Audio Flamingo Next (Ghosh et al., 2026), where single-audio corpora such as Music4All (Pegoraro Santana et al., 2020) serve as the audio source, or synthesizing speech from text-based dialogues via text-to-speech synthesis, as adopted by Step-Audio 2 (Wu et al., 2025).

E Effective Context Length: Background and Challenges for Audio

In text LLMs, effective context length is typically measured through diagnostic benchmarks designed to probe how well models utilize information at different positions within a long input. The most widely adopted approach is the needle-in-a-haystack (NIAH) test (Kamradt, 2023): a specific piece of information (the "needle") is inserted at varying positions within a long body of distractor text (the "haystack"), and the model is queried on

Model	Modality	Backbone	Max Context	Token Rate	Max Audio Length	Audio Tokenizer
Pengi (Deshmukh et al., 2023)	$P(T \mid A, T)$	GPT-2 Base (Radford et al., 2019)	1k	Fixed 40 tokens	7 s (training)	HTS-AT (Chen et al., 2022a)
SpeechGPT (Zhang et al., 2023)	$P(T \mid A, T)$	LLaMA-13B (Touvron et al., 2023a)	2k	< 50Hz (deduplication)	-	mHuBERT units (Lee et al., 2022)
Qwen-Audio (Chu et al., 2023)	$P(T \mid A, T)$	Qwen-7B (Bai et al., 2023)	8k	25 Hz	30 s	Whisper Encoder (Radford et al., 2023)
SALMONN (Tang et al., 2024)	$P(T \mid A, T)$	Vicuna-13B (Chiang et al., 2023)	2k	2.93 Hz	180 s (training)	Whisper Encoder + BEATs (Chen et al., 2023a) → Q-Former
Audio Flamingo (Kong et al., 2024)	$P(T \mid A, T)$	OPT-IML-MAX-1.3B (Iyer et al., 2023)	2k	- (cross-attention)	33.25 s (sliding)	CLAP (Elizalde et al., 2024)
WavLLM (Hu et al., 2024)	$P(T \mid A, T)$	LLaMA-2-7B	4k	12.5 Hz	30 s	Whisper Encoder + WavLM (Chen et al., 2022b)
Qwen2-Audio (Chu et al., 2024)	$P(T \mid A, T)$	Qwen-7B	8k	25 Hz	30 s	Whisper Encoder
VITA (Fu et al., 2024)	$P(A, T \mid A, T)$	Mixtral-8x7B (Jiang et al., 2024)	32k	12.5 Hz	-	CNN-Transformer
Moshi (Défossez et al., 2024)	$P(A, T \mid A, T)$	Helium-7B (Défossez et al., 2024)	4k	12.5 Hz	5 min	Mimi codec (Défossez et al., 2024)
SpiRit-LM Base (Nguyen et al., 2025)	$P(A, T \mid A, T)$	LLaMA-2-7B (Touvron et al., 2023b)	4k	25 Hz	-	mHuBERT units
SpiRit-LM Expressive (Nguyen et al., 2025)	$P(A, T \mid A, T)$	LLaMA-2-7B	4k	25 + 12.5 + 1 Hz	-	mHuBERT units + F0 tokens + Style tokens
LLaMA-Omni (Fang et al., 2025)	$P(A, T \mid A, T)$	LLaMA-3.1-8B (Grattafiori et al., 2024)	128k	10 Hz	30 s	Whisper Encoder
Audio Flamingo 2 (Ghosh et al., 2025b)	$P(T \mid A, T)$	Qwen2.5-3B	32k	- (cross-attention)	5 min	AFCLAP (Ghosh et al., 2025b)
Kimi-Audio (KimiTeam et al., 2025)	$P(A, T \mid A, T)$	Qwen2.5-7B (Hui et al., 2024)	32k	12.5 Hz	-	Whisper VQ + Whisper Encoder
Step-Audio 2 mini (Wu et al., 2025)	$P(A, T \mid A, T)$	Qwen2.5 7B	16k (ft) / 32k	12.5 Hz	-	Whisper Encoder (From Qwen2-Audio)
Audio Flamingo 3 (Ghosh et al., 2025a)	$P(T \mid A, T)$	Qwen2.5-7B	128k (ft) / 32k	25 Hz	10 min	AF-Whisper (Ghosh et al., 2025a)
Qwen3-Omni (Xu et al., 2025b)	$P(A, T \mid A, T)$	Qwen3-30B-A3B (Yang et al., 2025a)	32k (ft) / 256k	12.5 Hz	40 min	Audio Transformer (Xu et al., 2025b)
Audio Flamingo Next (Ghosh et al., 2026)	$P(A, T \mid A, T)$	Qwen2.5-7B	32k	25 Hz	30 min	AF-Whisper v2 (Ghosh et al., 2026)

Table 1: Representative SLMs. **Modality** indicates the input/output modality (e.g., $P(T \mid A, T)$: text-audio-in, text-out, etc.). **Backbone** is the LLM backbone. **Max Context** is the LLM backbone's maximum sequence length reported by each work. "ft" denotes the context length used during audio fine-tuning. **Token Rate** is the per-second rate of audio feature passed to the LLM after encoder-side pooling and projection. **Max Audio Length** is the longest audio input supported by the SLM. **Audio Tokenizer** is the encoder or tokenizer producing the audio representation fed to the LLM. We only report the value available in the original papers.

Dataset	Domain	Scale	Per-sample duration
Understanding / reasoning			
LongAudio (Ghosh et al., 2025b)	Sound and music QA	∼263K instances	30s–5min
LongAudio-XL (Ghosh et al., 2025a)	Speech, sound, and music QA	∼1.25M instances	30s–10min
LongSpeech† (Yang et al., 2026b)	Long-form speech (ASR / translation / summarization / QA)	100K+ segments	∼10 min
Generation			
LibriLight (Kahn et al., 2020)	Audiobook speech, re-segmented (long-form generation)	60K hours	up to 16 min

Table 2: Representative datasets used or proposed for training Long-Form Audio SLMs. †Primarily an evaluation benchmark with a training split.

that information. By systematically varying both the needle's position and the total context length, one can characterize where and at what length the model begins to fail. Building on this, Hsieh et al. (2024) propose RULER, a more comprehensive benchmark that extends beyond simple retrieval to include multi-hop tracing, aggregation, and question answering tasks, providing a richer characterization of effective context length across diverse capability dimensions.

Adapting such protocols to audio introduces several challenges that have no direct counterpart in text. First, inserting a needle into a continuous audio stream is substantially harder than inserting a sentence into text: the needle must be spliced into a waveform while preserving acoustic naturalness, and the boundaries between needle and haystack must be imperceptible to avoid giving the model artificial cues. Second, what constitutes a meaningful needle differs across information types. A semantic needle, such as a spoken fact or name, is relatively straightforward to construct; a paralinguistic needle, such as a specific emotional tone expressed at a particular moment, is harder to define, insert, and verify; and an environmental needle, such as a brief acoustic event embedded in background noise, raises additional questions about salience and detectability. Third, the appropriate "haystack" for audio is less well-defined than for text, as a natural audio haystack must maintain acoustic coherence across its full duration. These challenges suggest that developing principled NIAH-style protocols for audio, and extending them to cover non-semantic information types, is itself a non-trivial research problem and a prerequisite for the kind of analysis advocated in §6.2.

F Safety Vulnerabilities of SLMs

As discussed in §6.4, a growing body of work has demonstrated that SLMs are vulnerable to diverse jailbreaking attacks, many of which exploit attack surfaces unique to the audio modality. We provide a more detailed survey of these efforts here.

Existing attacks on SLMs span a wide range of vectors. Shen et al. (2024) show that fictional storytelling can elicit harmful responses from GPT-4o. Yang et al. (2025d) demonstrate that even meaningless audio noise can serve as effective attack inputs. A second line of work develops adversarial audio attacks, including SpeechGuard (Peri et al., 2024), AdvWave (Kang et al., 2025), and AudioJailbreak (Chen et al., 2026). More recent efforts further exploit paralinguistic and signal-level properties of speech: Cheng et al. (2025) bench-

Modality	Dataset	# Clips	Total Duration	Description
Speech	LibriSpeech (Panayotov et al., 2015)	∼281K	∼960 h	Read English speech from audiobooks with aligned transcripts.
Sound	AudioSet (Gemmeke et al., 2017)	∼2.07M	∼5,736 h	Large-scale YouTube clips (10s) with 527 sound event classes.
	VGGSound (Chen et al., 2020)	∼178K	∼494 h	YouTube clips (10s) across 309 classes with guaranteed audio-visual correspondence.
	AudioCaps (Kim et al., 2019)	∼50K	∼138 h	AudioSet subset with human-written caption.
	Clotho (Drossos et al., 2020)	∼3.8K	∼24 h	Freesound environmental clips (15–30s), each with five crowd-sourced captions.
	†WavCaps (Mei et al., 2024)	∼403K	∼7,566 h	Weakly-labelled clips from FreeSound, BBC Sound Effects, SoundBible, and AudioSet SL, with ChatGPT-cleaned captions.
Music	†MusicCaps (Agostinelli et al., 2023)	∼5,521	∼15 h	AudioSet music subset (10s) with expert captions written by professional musicians.
	†Music4All (Pegoraro Santana et al., 2020)	∼109K	∼910 h	30-second music clips with genre tags, lyrics, and rich metadata.

Table 3: Single-audio corpora commonly used to construct multi-audio training data. Clip counts and durations reflect the training split where an official split exists, and the full dataset otherwise (†).

mark jailbreak threats across diverse audio manipulations, while Feng et al. (2026a) and Li et al. (2026) respectively show that emotional and stylistic variations in the speaker's voice can bypass safety alignment. Dedicated benchmarks such as JALMBench (Peng et al., 2026) have also been proposed to systematically evaluate jailbreak vulnerabilities. Collectively, these works highlight that SLMs face attack surfaces absent in text, including paralinguistic and acoustic dimensions, even within a single short utterance, motivating the long-context safety concerns discussed in the main text.