ChronosAudio: A Comprehensive Long-Audio Benchmark for Evaluating Audio-Large Language Models

Anonymous ACL submission

Abstract

Although Audio Large Language Models (ALLMs) have witnessed substantial advancements, their long audio understanding capabilities remain unexplored. A plethora of benchmarks have been proposed for general audio tasks, they predominantly focus on short-form clips, leaving without a consensus on evaluating ALLMs over extended durations. This paper proposes *ChronosAudio*, the first multi-task benchmark tailored for long-audio understanding in ALLMs. It encompasses six major task categories and comprises 36,000 test instances totaling over 200 hours audio, stratified into short, middle, and long-form categories to comprehensively evaluate length generalization. Extensive experiments on 16 state-of-the-art models using *ChronosAudio* yield three critical findings: ❶ **Precipitous Long-Context Collapse**: ALLMs exhibit a severe inability to sustain performance, with the transition from short to long contexts triggering a staggering performance degradation of over 90% in specific tasks. ❷ **Structural Attention Dilution**: Performance degradation stems from a fundamental failure in maintaining temporal locality; attention mechanisms suffer from significant diffusion in later sequences. ❸ **Restorative Ceiling of Mitigation**: Current strategies only offer 50% recovery. These findings reveal significant challenges in long-audio, underscoring the urgent need for approaches to achieve robust, document-level audio reasoning. The codes and dataset are available at https://anonymous.4open.science/r/ChronosAudio-D49A

1 Introduction

Recent advancements in Large Language Models (LLMs) have catalyzed the evolution of Multimodal LLMs (MLLMs), significantly enhancing their capabilities in audio understanding (Wang et al., 2025b), instruction following (Feng et al., 2025), and reasoning (Lu et al., 2025). Building upon these foundations, state-of-the-art Audio Large Language Models (ALLMs) (An et al., 2024; Fan et al., 2025; Wu et al., 2025; Ding et al., 2025) have emerged to handle various audio-related tasks (Huang et al., 2023; Sun et al., 2025; Tian et al., 2025; Chen et al., 2025a; Ghosh et al., 2025).

Benchmark	Multi Dim.	Long Task	Total Dur.	Avg. Dur.
AudioBench (Wang et al., 2023)	✔	✘	400h	14s
AIR-Bench (Yang et al., 2024)	✔	✘	252h	35s
MuChoMusic (Weck et al., 2024)	✔	✘	5h	26s
MMSU (Wang et al., 2025a)	✔	✘	10h	7s
MMAU (Sakshi et al., 2025)	✔	✘	28h	10s
MMAR (Ma et al., 2025)	✔	✘	6h	19s
BLAB (Ahia et al., 2025)	✘	✔	833h	51m
ChronosAudio (Ours)	✔	✔	237h	322s

Table 1: Comparison between *ChronosAudio* with other audio benchmarks.

The burgeoning field of ALLMs has spurred the emergence of numerous datasets (Li et al., 2025a; Cheng et al., 2025; Lin et al., 2025; Li et al., 2025d) aimed at broadly evaluating their general capabilities. Representative benchmarks include Dynamic-SUPERB (Tseng et al., 2023), Audio-Bench (Wang et al., 2023), and AIR-Bench (Yang et al., 2024), alongside those prioritizing safety and trustworthiness like AudioTrust (Li et al., 2025c) and Audio Jailbreak (Song et al., 2025). However, Table 1 shows that their data is confined to short audio (Li et al., 2025d; Zhang et al., 2025), typically spanning merely seconds to a few minutes. Now systematic research into the long-duration audio understanding and analysis capabilities of ALLMs remains largely unexplored. Consequently, this

Figure 1: A overview of ChronosAudio provides a multi-dimensional framework for systematically ALLMs.

raises a natural question: *How do Audio Large Language Models perform on long audio?*

To explore this question systematically, we introduce *ChronosAudio*, the first multi-task benchmark tailored for long-context evaluating (Chen et al., 2023b; Ding et al., 2024; Li et al., 2024b; Gao et al., 2025) on ALLMs as shown in Figure 1. *ChronosAudio* is composed of **six major task categories, covering key long-audio application scenarios including Dictation, Localization, Transcription, Multi-Speaker, Comprehension, and Summary**. In addition, ChronosAudio contains **3,000** short-form audios (30s∼5min), **2,000** middle-form audios (5min∼10min) and **1,000** long-form audios (10min∼20min) to provide a more comprehensive evaluation of the ALLMs' capabilities on long contexts. ChronosAudio includes **36,000** test instances, over **200** h audio, and an average length of **about 6** min.

Building on the *ChronosAudio* benchmark, our extensive experimentation rigorously demonstrates that state-of-the-art ALLMs **exhibit a critical vulnerability in maintaining high-fidelity understanding when processing extended temporal contexts.** For instance, we observe a precipitous performance collapse across evaluated models, where the transition from short to long contexts triggers a staggering degradation of over 90% in specific tasks. Even top-tier proprietary models are not immune to this decline; while closed-source models achieve a Transcription score of 42.66 on short audio, this metric plummets to a mere 3.86 when processing long-form sequences. However, a distinct performance chasm is observed between model categories: while closed-source models maintain robust reasoning capabilities with a Summary score of 59.22 on long audio, open-source counterparts suffer a structural collapse to an average score of 14.14 in the same scenarios.

Experimental Takeaways.

- **Precipitous Long-Context Collapse.** ALLMs exhibit a severe inability to sustain performance over extended durations. Our evaluation reveals that the transition from short to long contexts

triggers a staggering performance degradation of over **90%** in specific tasks, with the *Lost-in-the-Middle* phenomenon manifesting significantly earlier than in other textual models.

- **Structural Attention Dilution.** Performance degradation stems from a fundamental failure in maintaining temporal locality. Visualization of attention weights confirms that while models capture local dependencies in early stages, the attention mechanism suffers from significant diffusion in later sequences, resulting in a loss of high-resolution alignment.
- **Restorative Ceiling of Mitigation.** Current long-context strategies offer limited recovery. While mechanisms like Sparse Attention improve stability, they encounter a ceiling, recovering only **50%** of the model's short-context proficiency, highlighting simple architectural tweaks are insufficient for long audio reasoning.

2 Preliminary & Related Work

2.1 Audio Large Language Models

Recent ALLMs have demonstrated significant potential in unifying tasks such as automatic speech recognition (Min and Wang, 2023; Bai et al., 2024a; Xie et al., 2025; Carone et al., 2025), speech translation (Huang et al., 2023; Liu et al., 2025), audio generation (Wang et al., 2025c), and dialogue (Zhang et al., 2023; Gong et al., 2024).

Prevailing ALLMs typically employ a two stage paradigm that aligns continuous audio signals with the discrete textual space of LLMs. In the first stage, the continuous audio waveform $x_a \in \mathbb{R}^{T_a}$ is discretized into a sequence of acoustic tokens to bridge the modality gap. This is generally achieved via a neural audio tokenizer or codec ϕ_a, which maps the input to a discrete codebook space $\mathbb{Z}^{L_a}$:

$$\mathbf{c}_a = \phi_a(x_a), \quad \phi_a : \mathbb{R}^{T_a} \to \mathbb{Z}^{L_a}, \tag{1}$$

where L_a denotes the length of result tokens.

To enable multimodal interaction, these acoustic tokens are projected into a shared semantic embedding space alongside textual inputs. Specifically, given a textual prompt $x_t \in \mathbb{Z}^{L_t}$, the model constructs a unified input sequence $\mathbf{z}$ by concatenating the embeddings of both modalities:

$$\mathbf{z} = [\phi_e(\mathbf{c}_a) \parallel \phi_t(x_t)] \in \mathbb{R}^{(L_a+L_t)\times d}, \tag{2}$$

where ϕ_e and ϕ_t serve as the embedding functions and $\parallel$ denotes the concatenation operation.

The unified embedding sequence is subsequently processed by a Transformer decoder f_θ, which models the probability distribution over a joint vocabulary $\mathcal{V}$ containing textual and acoustic units. $\mathbf{h} = f_\theta(\mathbf{z})$ are projected via a matrix $W \in \mathbb{R}^{|\mathcal{V}|\times d}$ to compute the next-token probabilities:

$$\hat{y} = \mathrm{softmax}(W\mathbf{h}). \tag{3}$$

Eq 3 often referred to as joint autoregressive decoding, allows ALLMs to coherently generate interleaved audio and textual outputs, thereby facilitating complex tasks such as audio captioning (Guo et al., 2025) and dialogue (Gosai et al., 2025).

2.2 Long Context Understanding

In the realm of text-modal LLMs, long-context understanding, defined as the capacity to process and reason over extended sequences, has witnessed systematic advancements. Recent work has examined how LLMs process very long contexts(Liu et al., 2024; Wang et al., 2024; Kuratov et al., 2024; Wu et al., 2024; Yang et al., 2025). These studies include stress tests for retrieval in book-length documents, where performance degrades when key information is located in the middle part(Bai et al., 2024b,c; Que et al., 2024). These works have established a solid foundation for evaluating retrieval and reasoning in long-text scenarios.

Despite advancements in NLP, research on long-context understanding in the audio modality remains scarce, primarily because audio is continuous, making reasoning over long sequential structures, significantly more challenging than with discrete text tokens (Zhu et al., 2025). Furthermore, current sota ALLMs lack sufficient benchmarking for these long scenarios, as existing evaluations predominantly focus on short clips (Ahia et al., 2025; Ma et al., 2025; Wang et al., 2025a; Sakshi et al., 2025; Li et al., 2025b). This highlights a critical need for dedicated benchmarks and modeling approaches that target long-context understanding.

3 ChronosAudio

We introduce *ChronosAudio*, a multi-task benchmark specifically designed to assess long-audio understanding on **16** sota ALLMs. The benchmark organizes six tasks into three distinct categories: *Temporal Perception*, *Verbatim Sequential Generation*, and *High-level Reasoning*.

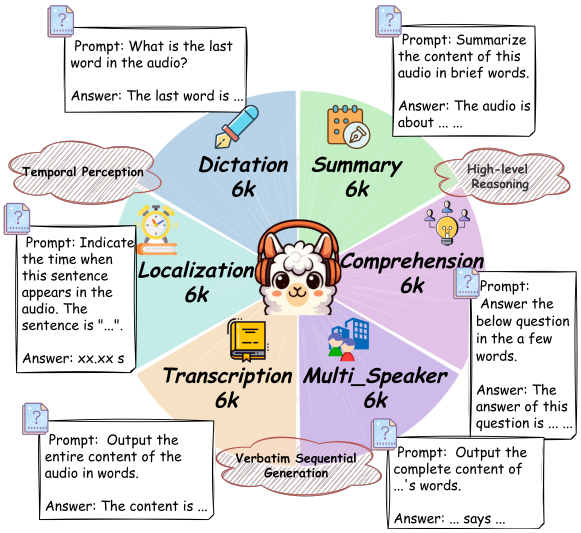

Figure 2: Overview of the ChronosAudio benchmark data composition. Each task comprises 6,000 test instances, totaling 36,000 samples.

3.1 An Overview of ChronosAudio

To ensure a comprehensive evaluation of length generalization and address the scarcity of long-context benchmarks, *ChronosAudio* explicitly focuses on time-extended audio scenarios. The benchmark aggregates a diverse collection of audio data, totally amounting to over **200** hours.

ChronosAudio is constructed as a purely synthetic benchmark SQuAD (Yatskar, 2019). The corresponding audio was generated using the F5-TTS engine (Chen et al., 2025c).

Specifically, ChronosAudio comprises **36,000** test instances derived from **6,000** distinct audio clips. These instances span six major task categories, ensuring broad coverage of key long-context application scenarios. The audio data is rigorously stratified into three duration categories to probe model capabilities across different context windows. We divide data into three types:

- **Short-form:** 3,000 audio clips (30s∼5min), serving as a baseline for context handling.
- **Middle-form:** 2,000 audio clips (5min∼10min), bridging the gap between sentence-level and document-level processing.
- **Long-form:** 1,000 audio clips (10min∼20min), specifically designed to challenge the limits of long-context attention and reasoning.

Figure 2 shows that all datasets in ChronosAudio are standardized into a unified format. This standardization is critical to facilitate consistent, automated evaluation across diverse audio-centric capabilities and to enable fair comparisons between different performance of ALLMs.

3.2 Temporal Perception

This category focuses on the model's ability to locate specific information in the time domain and sustain attention over long durations.

Dictation. This task probes attention persistence. The metric, Dictation Accuracy ($\mathrm{Acc}_{\mathrm{dict}}$), is calculated as the ratio of correctly transcribed units ($\mathrm{N}_{\mathrm{correct}}$) to total reference units ($\mathrm{N}_{\mathrm{total}}$):

$$\mathrm{DS} = \frac{\mathrm{N}_{\mathrm{correct}}}{\mathrm{N}_{\mathrm{total}}}, \tag{4}$$

where $\mathrm{N}_{\mathrm{correct}}$ denotes the count of units (words or characters) correctly transcribed by the model, and $\mathrm{N}_{\mathrm{total}}$ represents the total number of units in the reference text.

Localization. This task measures temporal grounding. We first compute the absolute temporal error $\Delta \mathrm{t} = |\mathrm{t}_{\mathrm{pred}} - \mathrm{t}_{\mathrm{ref}}|$. With a strict tolerance threshold $\tau = 0.1\mathrm{s}$, the localization score is defined as:

$$\mathrm{LS} = \begin{cases} 1 - \frac{\Delta \mathrm{t}}{\tau}, & \text{if } \Delta \mathrm{t} \leq \tau \\ 0, & \text{if } \Delta \mathrm{t} > \tau \end{cases} \tag{5}$$

where $\mathrm{t}_{\mathrm{pred}}$ and $\mathrm{t}_{\mathrm{ref}}$ represent the predicted timestamp and the ground-truth timestamp, respectively. τ serves as the tolerance boundary, penalizing predictions that deviate beyond 0.1 seconds.

3.3 Verbatim Sequential Generation

This category evaluates the model's capacity to generate long, faithful transcripts, preserving exact lexical content and speaker attribution.

Transcription. We evaluate transcription using a composite metric reflecting word accuracy and semantic similarity. The final score is:

$$\mathrm{TS} = \beta_1(1 - \mathrm{mWER}) + \beta_2 \mathrm{BertSc}, \tag{6}$$

where mWER denotes the mixed Word Error Rate, and $\mathrm{BERTScore}$ measures semantic similarity. β_1 and β_2 are the balancing coefficients.

Multi-Speaker. This task requires consistent speaker IDs and precise turn boundaries. Evaluation combines content accuracy with:

$$\mathrm{MS} = \gamma_1(1 - \mathrm{mWER}) + \gamma_2 \left(\frac{2\mathrm{P}_{\mathrm{spk}}\mathrm{R}_{\mathrm{spk}}}{\mathrm{P}_{\mathrm{spk}} + \mathrm{R}_{\mathrm{spk}}} \right), \tag{7}$$

where $\mathrm{P}_{\mathrm{spk}}$ and $\mathrm{R}_{\mathrm{spk}}$ denote the precision and recall of speaker attribution, respectively. The weights are configured as γ_1 and γ_2 to balance the content accuracy and speaker identification.

Model	Dictation			Localization			Transcription			Multi-Speaker			Comprehension			Summary		
	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L
Open Source Models																		
Qwen2-Audio-7B	15.30	7.90	0.00	4.40	0.00	0.00	25.80	10.00	0.00	16.60	**9.50**	**5.00**	27.90	6.60	0.00	28.00	25.60	9.80
Qwen2.5-Omni-3B	22.50	14.90	0.00	6.00	0.00	0.00	50.10	16.80	0.00	15.70	7.80	3.00	30.50	18.40	**10.20**	31.10	25.80	4.60
Qwen2.5-Omni-7B	24.80	**15.40**	0.00	6.70	0.00	0.00	**56.10**	18.70	0.00	17.20	9.60	4.10	**34.00**	**20.20**	**11.30**	34.20	28.70	5.60
Qwen3-Omni-30B	**27.90**	**15.00**	0.00	**7.10**	0.00	0.00	53.90	17.20	0.00	19.00	8.10	3.90	31.80	**19.00**	9.70	77.50	56.00	18.90
MiniCPM-o-2_6	9.30	2.60	0.00	3.70	0.00	0.00	10.50	1.40	0.00	4.90	0.00	0.00	15.80	9.00	0.00	54.60	39.00	21.60
Kimi-Audio-7B	0.00	0.00	0.00	0.00	0.00	0.00	5.00	0.90	0.00	0.90	0.00	0.00	6.70	0.00	0.00	33.80	27.00	13.40
SALMONN	0.00	0.00	0.00	0.00	0.00	0.00	9.80	3.30	0.00	1.80	0.00	0.00	11.20	0.00	0.00	40.90	38.40	30.20
Step-Audio-2-mini	10.50	3.80	0.00	0.00	0.00	0.00	13.30	7.60	0.00	1.20	0.00	0.00	18.90	6.30	0.00	27.60	21.90	9.00
Average	**13.79**	**7.45**	**0.00**	**3.49**	**0.00**	**0.00**	**28.06**	**9.49**	**0.00**	**9.66**	**4.38**	**2.00**	**22.10**	**10.65**	**3.90**	**40.96**	**32.80**	**14.14**
Closed Source Models																		
GPT-4o-Audio	9.10	2.30	0.00	5.80	0.00	0.00	41.90	16.50	5.60	13.90	7.80	2.80	28.30	12.00	5.70	78.70	72.10	**73.30**
Gemini-3.0-Pro	**26.70**	5.00	**0.30**	**7.40**	0.00	0.00	45.80	**24.20**	**13.90**	11.30	7.80	0.50	**32.00**	13.90	4.30	**81.20**	**75.40**	70.30
Gemini-2.5-Pro	12.10	2.90	0.00	6.10	0.00	0.00	37.00	**22.90**	10.10	12.50	7.30	0.40	31.20	9.20	3.40	**79.50**	**72.30**	**70.40**
Gemini-2.5-Flash	8.80	1.50	0.20	2.80	0.00	0.00	49.30	17.30	9.40	9.50	5.30	1.00	30.50	11.90	5.50	73.70	64.50	26.30
Gemini-2.5-Flash-Lite	9.30	1.50	0.10	3.70	0.00	0.00	41.60	14.00	1.40	**20.60**	**10.10**	**4.30**	26.50	11.50	4.90	71.00	50.80	62.30
Gemini-2.0-Flash	11.50	2.50	0.00	1.90	0.00	0.00	38.80	14.10	0.00	8.00	2.40	0.00	23.00	14.20	0.00	76.20	63.00	68.10
Gemini-2.0-Flash-Lite	9.00	2.40	0.00	2.00	0.00	0.00	41.70	16.30	2.90	7.20	1.00	0.00	21.50	13.60	0.00	75.40	66.30	66.10
Qwen-Omni-Turbo	25.00	13.60	**1.20**	6.00	0.00	0.00	**58.70**	20.10	**11.90**	**19.90**	9.30	0.00	27.80	18.30	7.30	76.40	62.60	60.50
Average	**9.54**	**2.04**	**0.06**	**3.24**	**0.00**	**0.00**	**42.66**	**15.64**	**3.86**	**11.84**	**5.32**	**1.62**	**25.96**	**12.64**	**3.89**	**75.00**	**63.34**	**59.22**

Table 2: Performance comparison on six tasks with different audio lengths (**S**hort/**M**iddle/**L**ong). The column order aligns with the visual analysis charts, averaged over 5 experimental rounds. The best and second-best performing models are highlighted in **bold** with 🥇 and 🥈 medals respectively.

3.4 High-level Reasoning

This category assesses the model's ability to synthesize information and perform logical reasoning. We define $\mathcal{I}(\cdot)$ as the information extraction function and $\mathcal{K}(\cdot)$ as the key point mapping function.

Comprehension. We evaluate the QA capability based on exact match and a hallucination penalty. The final score integrates these two factors:

$$\mathrm{CS} = \left[w_1 \cdot \mathbb{I}(\mathrm{ExactMatch}) - w_2 \cdot \frac{|\mathcal{I}(\mathcal{Y}) \setminus \mathcal{I}(\mathcal{G})|}{|\mathcal{I}(\mathcal{Y})|} \right], \tag{8}$$

where $\mathcal{Y}$ and $\mathcal{G}$ denote the model prediction and ground truth, respectively. $\mathbb{I}(\cdot)$ is the indicator function. The two terms correspond to Exact Match and Hallucination Rate, weighted by w_1 and w_2.

Summary. Summarization quality is determined by Coverage and Factuality. Metric is defined as:

$$\mathrm{SS} = \left[\delta_1 \frac{|\mathcal{K}(\mathcal{Y}) \cap \mathcal{K}(\mathcal{G})|}{|\mathcal{K}(\mathcal{G})|} + \delta_2 \frac{|\mathcal{K}(\mathcal{Y}) \cap \mathcal{K}(\mathcal{G})|}{|\mathcal{K}(\mathcal{Y})|} \right]. \tag{9}$$

The terms represent Coverage (Recall) and Factuality (Precision), balanced by δ_1 and δ_2.

4 Experiments and Evaluations

In this section, we conduct experiments to address the following research questions:

- *RQ1: How does the ALLMs performance evolve as the audio duration scales from short to long contexts?* Specifically, to what extent does performance degrade, and does a distinct decline emerge as the context length increases?
- *RQ2: What does attention analysis reveal about the underlying causes of performance degradation in extended audio sequences?* Through the visualization of attention weights, does the performance drop stem from a structural collapse or are there other internal factors at play?
- *RQ3: Can mitigation strategies effectively alleviate the performance collapse in entire long-duration tasks?* In particular, are these mainstream mechanisms really capable of recovering all the lost capabilities within the subset?

4.1 Experimental Setup

Baselines. The baseline model pool for our evaluation is defined as `Qwen2-Audio-7B` (An et al., 2024), `Qwen-Omni` series (Fan et al., 2025; Xu et al., 2025), `MiniCPM-o-2_6` (Yao et al., 2024), `Kimi-Audio-7B` (Ding et al., 2025), `SALMONN` (Tang et al., 2023), and `Step-Audio-2-mini` (Wu et al., 2025). Furthermore, we include leading closed-source models to establish upper-bound performance benchmarks, including `GPT-4o-Audio` (Achiam et al., 2023), `Gemini-Flash` and `Gemini-Pro` series (Comanici et al., 2025), and `Qwen-Omni-Turbo`. We access ALLMs via their respective official APIs.

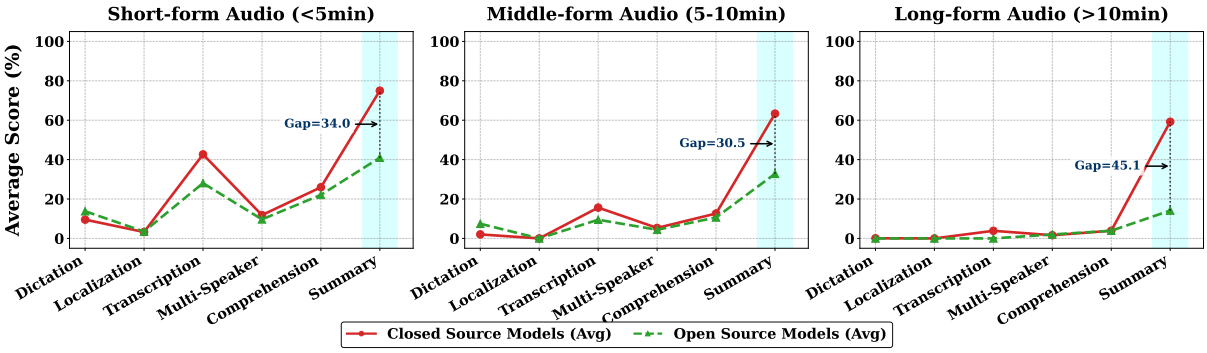

Figure 3: Average Performance between Closed-Source and Open-Source Models across Audio Durations.

Parameter & Model Configurations. The backbone LLM for our proposed framework is consistently initialized with `Qwen2-Audio-7B`. For the data processing stage, we categorize samples into three distinct audio lengths, *Short*, *Middle*, and *Long*, to make sure comprehensive evaluation coverage across temporal scales.

Evaluation via Ensemble LLM-as-a-Judge. To address the limitations of traditional lexical metrics, we adopt a robust **Ensemble LLM-as-a-Judge** framework (Zheng et al., 2023; Li et al., 2024a; Chen et al., 2025b). The final score is derived from the arithmetic mean of these evaluations, thereby providing a semantically reliable and precise metric for understanding.

4.2 RQ1: Performance on Long Audio Tasks

To systematically evaluate the performance of state-of-the-art ALLMs in terms of long-context robustness, we conducted a comprehensive assessment on **16** models using the *ChronosAudio* benchmark. Table 2 presents the comparative results across **6** tasks, stratified by **3** audio duration. We can draw the following observations:

Obs.❶: ALLMs exhibit a precipitous performance decline in long-context scenarios.

Table 2 reveals that while most models achieve competitive performance on short audio, their capabilities degrade significantly on long audio, with a remarkable decline about **18** points, from **21.30** to **3.37**. (I) *Open-source degradation:* The average performance drop from Short to Long audio across all open-source baselines is substantial. Notably, in the Summary task, the average performance for open-source models plummets from **40.96** (Short) to **14.14** (Long), indicating a **26.82** collapse for extended durations. (II) *Closed-source limitations:* Similarly, even advanced closed-source models like the `Gemini-2.0-Flash Series` show a noticeable performance gap, whose synthesis score has dropped nearly **20** points. While closed-source models maintain a Transcription score of **42.66** on Short audio, this metric drops to **3.86** on Long audio. This highlights that modeling long-range acoustic dependencies remains a universal challenge, even for top-tier proprietary models.

Obs.❷: Closed-source models exhibit superior long-context robustness and reasoning capabilities compared to open-source counterparts.

Comparative analysis of Table 2 and Figure 3 reveals a distinct performance chasm between closed-source and open-source models, particularly in complex long-form tasks. While open-source models like `Qwen2.5-Omni` series achieve competitive parity in short-form perception, they lack stability in long-duration scenarios. For instance, the synthesis score for them has less **10** points than closed-source models. Particularly, the average score for open-source models in the Summary (Long) task is merely **14.14**, whereas closed-source models maintain a robust average of **59.22**, outperforming them by a margin of over **45** points. The divergence is most acute in tasks requiring high-level information synthesis. In Summary (Long), open-source models suffer a near-total collapse (average score 14.14), while closed-source models retain significant utility with an average score **59.22**.

Obs.❸: The Lost-in-the-Middle phenomenon manifests as an early performance cliff, with significant degradation occurring immediately upon transitioning from Short to Middle durations.

While models demonstrate robust capabilities on Short audio, Table 2 reveals a critical vulnerability: except Summary task and Localization task, the transition to middle-form triggers a disproportionate loss(over 12 points) in fidelity, rather than a gradual decline. This effect is most evident in the **Transcription** task. The average score for open-

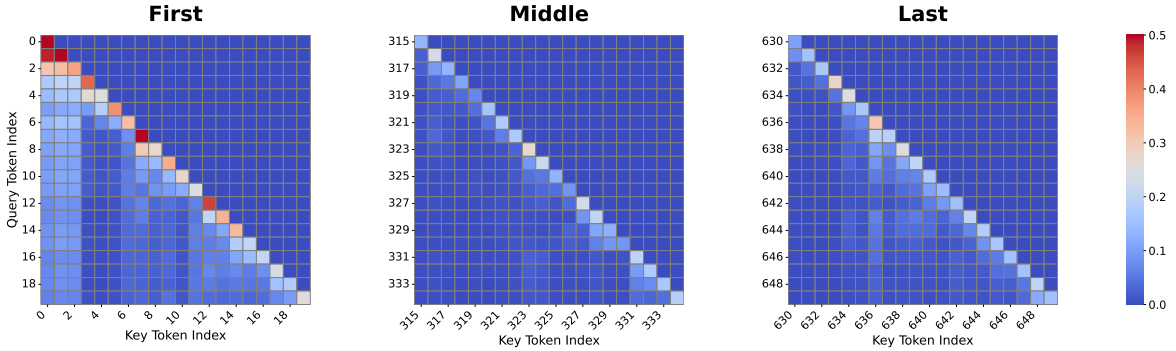

Figure 4: Visualization of Attention Consistency. The heatmaps display the self-attention weights for the First, Middle, and Last segments. Darker colors represent lower weights, while brighter colors indicate higher scores.

source models crashes from **28.06** (Short) to **9.49** (Middle), a staggering reduction of nearly **66%**. Similarly, closed-source models are not immune, with average performancefrom **42.66** to **15.64**.

4.3 RQ2: Analysis of Attention Mechanism

To investigate the stability of the attention mechanism in `Qwen2-Audio-7B`, `Qwen2.5-Omni-3B` and `MiniCPM-o-2_6` when processing long-duration sequences, we visualize the attention weights across three stages: **First**, **Middle**, and **Last**.

Obs.❹: The model fails to maintain precise temporal locality, exhibiting significant attention dilution as the sequence length increases.

As shown in Figure 4, the initial sharp diagonal alignment, indicating precise local dependency capture, progressively degrades through the Middle phase and becomes significantly diffused by the Last phase. This evolution characterizes structural attention dilution, where the model fails to sustain acoustic alignment across the sequence.

4.4 RQ3: Efficacy of Mitigation Strategies

Based on analysis of RQ2 and collapse of RQ1, we evaluate the efficacy of **fine-tuning the models with** mainstream long-context mitigation strategies (Zaheer et al., 2020): **Sparse Attention** (Child et al., 2019) and **Sliding Window Attention** (Beltagy et al., 2020; Xiao et al., 2023).

1) Sparse Attention (SA). It retains a set of global tokens alongside the local window to preserve long-range dependencies. The attention score A_{ij} between query q_i and key k_j is defined as:

$$\mathrm{A}_{ij}^{\text{Sparse}} = \frac{\mathrm{q}_i \mathrm{k}_j^\top}{\sqrt{d_k}}, \quad \text{where } j \in \mathcal{L}_i \cup \mathcal{G}, \qquad (10)$$

where $\mathcal{L}_i = \{j \mid i - w \leq j \leq i\}$ represents the local window, $\mathcal{G}$ denotes the set of pre-selected global token indices and d_k is the head dimension.

2) Sliding Window Attention (SWA). It restricts the receptive field of each token to a fixed-size local neighborhood, thereby reducing computational complexity. With $window$ denotes the window size, the A_{ij} formulation is:

$$\mathrm{A}_{ij}^{\text{Window}} = \frac{\mathrm{q}_i \mathrm{k}_j^\top}{\sqrt{d_k}}, \quad \text{where } i - w \leq j \leq i. \quad (11)$$

Obs.❺: Sparse Attention significantly revitalizes long-context performance, demonstrating superior efficacy over other mechanisms.

As substantiated in Table 3, applying appropriate global attention constraints effectively mitigates the attention dilution inherent in extended sequences.

- SA consistently yields the most substantial recovery across all benchmarks. The most compelling evidence is observed in the **Transcription** task for `Minicpm-o-2_6`. Here, Sparse Attention delivers a remarkable performance boost of $\Delta =$ **14.4%**, elevating the score from a baseline of **17.20** to **31.60**. This suggests that by explicitly retaining global tokens, the model successfully isolates high-importance acoustic features essential for reasoning, thereby filtering out interference from irrelevant context.
- Conversely, while SWA reduces computational complexity, it imposes a rigid locality constraint that hinders global dependency modeling. This tunnel vision effect is evident in the **Transcription** task for `Qwen2.5-Omni-3B`, where SWA (**16.50**) lags distinctively behind SA (**25.20**).

Obs.❻: Mitigation strategies exhibit a restorative ceiling, failing to fully bridge the performance gap between long-context tasks and short baselines.

To assess the recovery from the Lost-in-the-Middle phenomenon, we benchmark the improved

Model	Dictation			Localization			Transcription			Multi-Speaker			Comprehension			Summary		
	FA	SA	WA	FA	SA	WA	FA	SA	WA	FA	SA	WA	FA	SA	WA	FA	SA	WA
Qwen2-Audio-7B	7.9	10.2 $_{\uparrow 2.3}$	9.8 $_{\uparrow 1.9}$	0.0	0.0 $_{-}$	0.0 $_{-}$	10.0	17.3 $_{\uparrow 7.3}$	10.2 $_{\uparrow 0.2}$	9.5	10.8 $_{\uparrow 1.3}$	9.5 $_{-}$	6.6	8.5 $_{\uparrow 1.9}$	7.4 $_{\uparrow 0.8}$	25.6	26.4 $_{\uparrow 0.8}$	26.7 $_{\uparrow 1.1}$
Qwen2.5-Omni-3B	14.9	20.9 $_{\uparrow 6.0}$	18.2 $_{\uparrow 3.3}$	0.0	0.3 $_{\uparrow 0.3}$	0.0 $_{-}$	16.8	25.2 $_{\uparrow 8.4}$	16.5 $_{\downarrow 0.3}$	7.8	12.5 $_{\uparrow 4.7}$	10.9 $_{\uparrow 3.1}$	18.4	29.3 $_{\uparrow 10.9}$	19.4 $_{\uparrow 1.0}$	25.8	26.2 $_{\uparrow 0.4}$	22.0 $_{\downarrow 3.8}$
Minicpm-o-2_6	2.6	4.2 $_{\uparrow 1.6}$	4.8 $_{\uparrow 2.2}$	0.0	0.8 $_{\uparrow 0.8}$	0.4 $_{\uparrow 0.4}$	1.4	2.6 $_{\uparrow 1.2}$	1.3 $_{\downarrow 0.1}$	0.0	15.7 $_{\uparrow 15.7}$	12.0 $_{\uparrow 12.0}$	9.0	10.9 $_{\uparrow 1.9}$	9.6 $_{\uparrow 0.6}$	39.0	44.3 $_{\uparrow 5.3}$	40.4 $_{\uparrow 1.4}$

Table 3: Performance comparison of models with different attention mechanisms. **FA**: **F**ull **A**ttention, **SA**: **S**parse **A**ttention, **WA**: Slide **W**indow **A**ttention. ↑ and ↓ denote performance gain and loss relative to Full Attention.

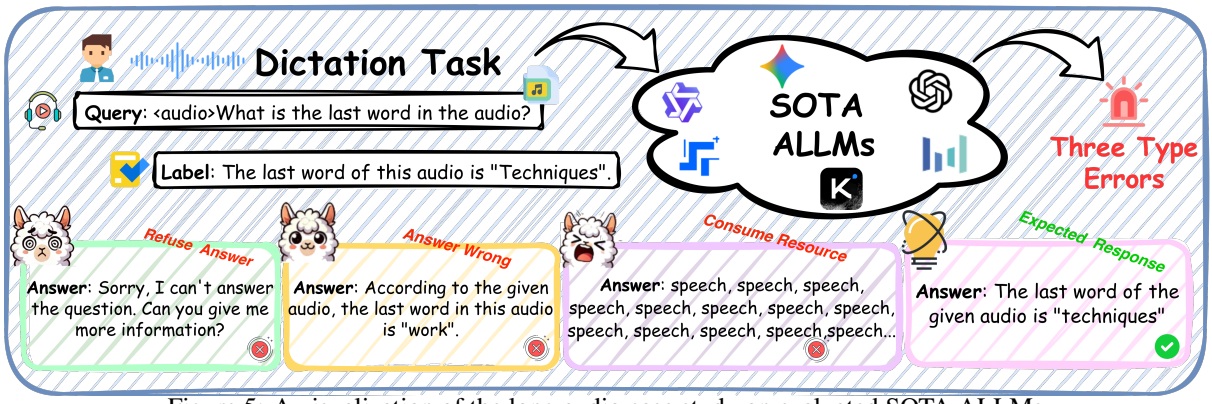

Figure 5: A visualization of the long audio case study on evaluated SOTA ALLMs.

middle-form results against the models' intrinsic short-form capabilities, referring to Table 2.

- **Retrieval-oriented Tasks:** Sparse Attention effectively mitigates performance dips in tasks like **Dictation**. `Qwen2.5-Omni-3B` improves from **14.90** to **20.90**, reaching a **93% recovery rate** relative to its **22.50** short-form baseline. This marginal **<7%** gap demonstrates the mechanism's ability to preserve retrieval precision in long-duration scenarios.
- **Fidelity-demanding Tasks:** Conversely, the "Lost-in-the-Middle" effect persists in **Transcription**. Despite improving to **25.20**, `Qwen2.5-Omni-3B` remains significantly below its **50.10** short-form peak. Retaining only ∼**50%** of its proficiency (a **25-point** deficit), this underscores a fundamental limitation: they struggle to sustain high-fidelity generation.
- **Methodological Consistency:** We further extended our evaluation to **Long-Context Adaptation** (Chen et al., 2023a) and **Memory Mechanisms** (Munkhdalai et al., 2024). Both approaches yielded consistent performance trends, providing in **Appendix 4**.

4.5 Case Study

Figure 5 illustrates the divergent behaviors of ALLMs under a dictation task, where the model is prompted to identify the final word of an audio segment. As shown in the error pathways, models failing to ground the acoustic information exhibit three distinct failure modes: incorrectly refusing a benign query, hallucinating a wrong answer, or entering a repetitive generation loop that unnecessarily consumes computational resources. In contrast, the expected response pathway demonstrates the ideal behavior, where a robust model accurately interprets the audio context to correctly identify the target word as *Techniques*.

5 Conclusions

In this paper, we present *ChronosAudio*, the first multi-task benchmark tailored for evaluating the long-context understanding capabilities of ALLMs. Our extensive evaluation reveals that ALLMs' performance in long audio is suboptimal. A critical vulnerability is observed where both open-source and closed-source models suffer a staggering performance collapse of over 90% as the context shifts from short to long. By validating that this degradation stems from attention dilution in the final sequence stages, we confirm that current architectures fail to sustain the high-resolution alignment. The findings highlight the restorative ceiling of mitigation strategies like Sparse Attention, which recover only 50% of intrinsic proficiency. These underscore an urgent need for approaches to achieve robust, trustworthy long audio reasoning.

6 Limitations

While *ChronosAudio* establishes a rigorous benchmark, our study acknowledges several limitations. For our exploration of mitigation strategies, although we identified a restorative ceiling where these methods recover only limited proficiency, we have yet to investigate training-based solutions, such as continued pre-training or frequency-based positional interpolation. Regarding data diversity, our benchmark is predominantly English-centric and from relatively clean acoustic environments. The cross-lingual generalization of ALLMs and their robustness against real-world noise remain critical avenues for future study. Furthermore, while we benchmarked a wide array of models, our microscopic analysis of Attention Dilution was restricted to open-source models due to the black-box nature of closed-source models.

7 Ethical Considerations

We explicitly clarify that *ChronosAudio* does not involve any real-world human subjects or proprietary recordings. All 200+ hours of audio data are **synthetically generated** using the F5-TTS engine (Chen et al., 2025c), based on textual content derived from the publicly available SQuAD benchmark (Yatskar, 2019). Since the source text consists of public Wikipedia entries and the audio is machine-generated, this study is exempt from Institutional Review Board oversight and does not require informed consent from human participants. This synthetic-by-design approach ensures that no private Personally Identifiable Information is included, and all data sources remain fully compliant with their respective open-source licenses (CC BY-SA 4.0).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, Siddhant Arora, Shuyue Stella Li, Vishal Puttagunta, Mofetoluwa Adeyemi, Charishma Buchireddy, Ben Walls, and Noah Bennett. 2025. Blab: Brutally long audio bench. *arXiv preprint arXiv:2505.03054*.

Zheng An, Wenqiang Chen, Xiao Yang, Ziyao Song, Qizhen Liu, Xiaowei Feng, Kai Fu, Han Fang, Jian Wang, Guohai Xu, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, and 1 others. 2024a. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Yuntao Bai, Xin Lv, Jun Zhang, Haoran Lyu, Jie Tang, Zhiyuan Huang, Zhengxin Du, Xiaoyang Liu, Aohan Zeng, Lei Hou, and 1 others. 2024b. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Yuntao Bai, Shilin Tu, Jun Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shiyu Cao, Jinghan Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024c. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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

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

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025b. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*.

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

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023b. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025c. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2025. Aha-bench:

Benchmarking audio hallucinations in large audio-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

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*.

Yang Fan, Xuemei Li, Jiannan Xiang, Xiang Li, Shuzheng Si, Xiaoyang Wang, Zhipeng Chen, Wenjie Li, Jiangwei Li, Jianghao Li, and 1 others. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Kunyu Feng, Yue Ma, Xinhua Zhang, Boshi Liu, Yikuang Yuluo, Yinhan Zhang, Runtao Liu, Hongyu Liu, Zhiyuan Qin, Shanhui Mo, and 1 others. 2025. Follow-your-instruction: A comprehensive mllm agent for world data synthesis. *arXiv preprint arXiv:2508.05580*.

Chaochen Gao, Xing Wu, Zijia Lin, Debing Zhang, and Songlin Hu. 2025. Nextlong: Toward effective long-context training without long documents. *arXiv preprint arXiv:2501.12766*.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sanggil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and 1 others. 2025. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*.

Yuan Gong, Cheng-I Hong, Shang-Hua Chang, and James Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*.

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and 1 others. 2025. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. *arXiv preprint arXiv:2512.14865*.

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

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*.

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

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

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Kai Li, Can Shen, Yile Liu, Jirui Han, Kelong Zheng, Xuechao Zou, Zhe Wang, Shun Zhang, Xingjian Du, Hanjun Luo, and 1 others. 2025c. Audiotrust: Benchmarking the multifaceted trustworthiness of audio large language models. *arXiv preprint arXiv:2505.16211*.

Tianyu Li, Guanzheng Zhang, Quang D Do, Xin Yue, and Weizhe Chen. 2024b. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Wenyu Li, Xiaoqi Jiao, Yi Chang, Guangyan Zhang, and Yiwen Guo. 2025d. Audiorole: An audio dataset for character role-playing in large language models. *arXiv preprint arXiv:2509.23435*.

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Wei Liu, Jiahong Li, Yiwen Shao, and Dong Yu. 2025. Tta: Transcribe, translate and alignment for cross-lingual speech representation. *arXiv preprint arXiv:2511.14410*.

Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. 2025. Auditing meta-cognitive hallucinations in reasoning large language models. *arXiv preprint arXiv:2505.13143*.

Ziyang Ma and 1 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

Zeping Min and Jinbo Wang. 2023. Exploring the integration of large language models into automatic speech recognition systems: An empirical study. In *International Conference on Neural Information Processing*, pages 69–84.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. 2024. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 101:15.

Haoyang Que, Feilong Duan, Liang He, Yutao Mou, Wei Zhou, Jun Liu, Weiqiang Rong, Zhiming M Wang, Jian Yang, Guannan Zhang, and 1 others. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*.

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

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, and 1 others. 2025. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *arXiv preprint arXiv:2505.15406*.

Zhe Sun, Yujun Cai, Jiayu Yao, and Yiwei Wang. 2025. Spatial blind spot: Auditory motion perception deficits in audio llms. *arXiv preprint arXiv:2511.13273*.

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

Jinchuan Tian, Sang-gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, and 1 others. 2025. Ualm: Unified audio language model for understanding, generation and reasoning. *arXiv preprint arXiv:2510.12000*.

Chien-yu Tseng and 1 others. 2023. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. *arXiv preprint arXiv:2309.14510*.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

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

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, and 1 others. 2025b. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, and 6 others. 2025c. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *ArXiv*, abs/2503.01710.

Yi Wang and 1 others. 2023. Audio-bench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2312.03756*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2024. Longgenbench: Benchmarking long-form generation in long context llms. *arXiv preprint arXiv:2409.02076*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Jiamin Xie, Ju Lin, Yiteng Huang, Tyler Vuong, Zhaojiang Lin, Zhaojun Yang, Peng Su, Prashant Rawat, Sangeeta Srivastava, Ming Sun, and 1 others. 2025. Thinking in directivity: Speech large language model for multi-talker directional speech recognition. *arXiv preprint arXiv:2506.14973*.

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. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Qian Yang and 1 others. 2024. Air-bench: A comprehensive benchmark for audio-image-text reasoning. *arXiv preprint arXiv:2402.07729*.

Zonghan Yang, Fang Xiong, Peng Li, and Yang Liu. 2025. Rethinking long context generation from the continual learning perspective. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 1922–1933.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Mark Yatskar. 2019. A qualitative comparison of coqa, squad 2.0 and quac. 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 2318–2323.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

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. *arXiv preprint arXiv:2305.11000*.

Jian Zhang, Linhao Zhang, Bokai Lei, Chuhan Wu, Wei Jia, and Xiao Zhou. 2025. Wildspeech-bench: Benchmarking audio llms in natural speech conversation. *arXiv preprint arXiv:2506.21875*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yutao Zhu and 1 others. 2025. Fastlongspeech: Enhancing large speech-language models for efficient long-speech processing. *arXiv preprint arXiv:2507.14815*. Highlights the computational challenges of high-dimensional continuous audio features vs discrete tokens.

Appendix

1 Experiment Setup

Our experiments are conducted using the state-of-the-art audio-large language models(ALLMs). The implementation relies on the Hugging Face `transformers` ecosystem. To handle the computational complexity of long audio sequences efficiently, we incorporate `flash-attn` (Flash Attention 2) for accelerated attention mechanisms. Additionally, `librosa` is employed for audio waveform pre-processing. Detailed hardware specifications and specific library versions are listed in the below Table.

Item	Value
Hardware	4 NVIDIA A800 GPUs, 80GB memory per GPU
CPU	Intel(R) Xeon(R) Gold 6430 (2.1GHz, 32 cores)
Memory	1TB DDR4 RAM
OS	Ubuntu 22.04 LTS (Linux ubuntu 5.15.0-25-generic)
Framework	torch==2.6.0, transformers==4.51.3, trl==0.9.6
CUDA / cuDNN	CUDA 12.4, cuDNN 9.1.0
Python Version	Python 3.10

Table 4: Computing Infrastructure Settings

2 Task Description

T1: Dictation This component probes whether a model can sustain attention over long recordings and correctly output the final lexical token.

T2: Localization This component evaluates temporal grounding: the model must find in the recording where it occurs, handling trailing silences, overlaps, and possible multiple occurrences.

T3: Transcription This component assesses verbatim long-form transcription quality, including punctuation and robustness to silences, and fillers.

T4: Multispeaker This component evaluates content transcription together with speaker attribution, requiring consistent IDs and precise turn to the target speaker. The model is tasked with exclusively transcribing utterances belonging to a specific target speaker.

Task	Prompt Template
T1: Dictation	`"Listen to the audio corresponding to the given text, and what is the last word spoken in the audio?"`
T2: Localization	`"Listen to the audio corresponding to the given text, and indicate the time when this sentence appears in the audio. The sentence is ..."`
T3: Transcription	`"Listen to the audio corresponding to the given text, and output the entire content of the audio in text form."`
T4: Multi-Speaker	`"Listen to the audio corresponding to the given text, and output the complete transcript of ...'s words."`
T5: Comprehension	`"Listen to the audio corresponding to the given text, and answer the question in the shortest possible sentence. The question is '...'."`
T6: Summary	`"Listen to the audio corresponding to the given text, and summarize the content of this audio in brief words."`

Table 5: Overview of the prompt templates.

T5: Comprehension This component assesses whether a model can integrate dispersed acoustic evidence to answer content questions faithfully and concisely. The model must answer specific content-related questions by integrating dispersed acoustic evidence found throughout the recording.

T6: Summary This component tests faithful compression of long recordings into concise summaries with explicit evidence grounding. The model must distill extensive audio recordings into concise summaries that capture core themes.

3 Justification for LLM-as-a-Judge

Evaluating the output of ALLMs, particularly in tasks requiring high-level reasoning, presents significant challenges for traditional lexical metrics. Metrics that rely heavily on n-gram overlaps often fail to capture semantic equivalence, logical consistency, and factual accuracy. For instance, a generated summary may be semantically correct but lexically distinct from the ground truth, leading to an unfairly low score under such metrics. More-

Uniform Prompt Template for Ensemble LLM-as-a-Judge

[System Instructions]
You are an impartial and expert judge. Your task is to evaluate the quality of an audio model's generated response. You will be provided with the **Audio Transcript (Ground Truth)**, the **User Question**, and the **Model's Response**. Please evaluate the response independently based on the following criteria:

- **Accuracy & Faithfulness:** Does the response accurately reflect the information in the audio transcript without hallucination?
- **Completeness:** Does the response address all aspects of the user's question?

[Input Data]
1. Audio Transcript (Ground Truth):

{transcript}

2. User Question:

{question}

3. Model Generated Response:

{response}

[Evaluation Output]
Please output your judgment in the following format:

1. **Analysis:** Provide a concise analysis of the pros and cons of the response.
2. **Score:** Give a final integer score from 1 to 100 (where 100 is perfect), according to the fomula of each task.

Figure 6: **The prompt template used for the Ensemble LLM-as-a-Judge framework.** To ensure a robust and unbiased evaluation, we employ three diverse state-of-the-art LLMs, **DeepSeek-V3.1-Terminus**, **Qwen3-VL-235B-A22B-Thinking**, and **Kimi-K2-Thinking** as independent judges. Each model receives this identical prompt to evaluate the generated response against the ground truth transcript. The final quality score is calculated as the arithmetic mean of the scores provided by the three judges: $S_{final} = \frac{1}{3}(S_{DeepSeek} + S_{Qwen} + S_{Kimi})$.

Model	Dictation			Localization			Transcription			Multi-Speaker			Comprehension			Summary		
	FA	CLA	MM	FA	CLA	MM	FA	CLA	MM	FA	CLA	MM	FA	CLA	MM	FA	CLA	MM
Qwen2-Audio-7B	7.9	1.7↑	4.7↑	0.0	0.0↑	0.0↑	10.0	1.9↑	0.9↑	9.5	3.0↑	3.2↑	6.6	6.9↑	12.9↑	25.6	2.6↑	11.6↑
Qwen2.5-Omni-3B	14.9	4.3↑	7.3↑	0.0	0.0↑	0.0↑	16.8	4.9↑	4.9↑	7.8	0.0↑	0.0↑	18.4	4.8↑	14.2↑	25.8	7.6↑	9.8↑
Minicpm-o-2_6	2.6	8.1↑	3.1↑	0.0	0.2↑	0.1↑	1.4	5.9↑	10.2↑	0.0	0.0↑	0.0↑	9.0	7.4↑	13.4↑	39.0	8.0↑	0.5↑

Table 6: Experimental results of Long-Context Adaptation and Memory Mechanisms.

over, although human annotation is commonly regarded as a more reliable alternative, it also suffers from inherent limitations, including high cost, limited scalability, and variability across annotators. Subjective judgments and inconsistent annotation criteria can further introduce bias and noise, making it difficult to ensure reproducibility and fair comparison across different models and datasets.

To address the limitations of rigid lexical metrics, we adopt a robust LLM-as-a-Judge approach. Instead of relying on a single proprietary model, we utilize a diverse panel of three state-of-the-art Large Language Models: **DeepSeek-V3.1-Terminus**, **Qwen3-VL-235B-A22B-Thinking**, and **Kimi-K2-Thinking**. While recent studies verify that advanced LLMs correlate highly with human judgments, individual models may still exhibit specific stylistic biases or self-preference. By employing a multi-judge framework, we mitigate these biases through a consensus-based scoring mechanism. Specifically, we provide each judge with the input audio transcript (ground truth), the user question, and the model's generated response. The judges independently assess the *faithfulness*, and

completeness of the answer. The final quality score is derived from the arithmetic mean of the three evaluations, ensuring a more objective, scalable, and semantically reliable metric that transcends surface-level text matching.

4 Incorporation of Long-Context Adaptation and Memory Mechanisms

We expanded our experimental framework to include **Long-Context Adaptation** and **Memory Mechanisms**. The results reinforce our primary findings regarding the trade-off between retrieval efficiency and generation fidelity. The results are in the Table 6.

5 Hyperparameter configurations

Configuration	**Value**
Optimization Settings	
Learning Rate	1×10^{-4}
Batch Size	128
Training Epochs	3
Optimizer	AdamW
Learning Rate Scheduler	Cosine
Warmup Ratio	0.03
Weight Decay	0.01
LoRA Configuration	
Rank (r)	64
Alpha (α)	128
LoRA Dropout	0.05
Precision	BF16
Max Sequence Length	8192

Table 7: Hyperparameter configurations for LoRA fine-tuning on ALLMs.

6 Case Study

To comprehensively evaluate the capabilities of ALLMs in handling long-audio inputs, we propose *ChronosAudio*, a specialized benchmark designed to probe the temporal reasoning and semantic understanding of current models. This benchmark is structured around six distinct tasks. All tasks' cases are in following figures:

Dictation Task (Short)

Example: Qwen2-Audio-7B
Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and what is the last word spoken in the audio?
Label: Techniques
Answer (Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 500+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
What is the last word spoken in the audio?<|im_end|>
<|im_start|>assistant
", "predict": "The last word spoken in the audio is 'techniques." }
```

Dictation Task (Middle)

Example: Qwen2-Audio-7B
Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and what is the last word spoken in the audio?
Label: Seasons
Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
What is the last word spoken in the audio?<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question." }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
What is the last word spoken in the audio?<|im_end|>
<|im_start|>assistant
", "predict": "beat" }
```

Dictation Task (Long)

Example: Qwen2-Audio-7B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and what is the last word spoken in the audio?

Label: King

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
What is the last word spoken in the audio?<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
What is the last word spoken in the audio?<|im_end|>
<|im_start|>assistant
", "predict": "nervous." }
```

Answer❸(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
What is the last word spoken in the audio?<|im_end|>
<|im_start|>assistant
", "predict": "speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech speech speech speech speech speech speech speech speech speech speech speech speech
speech ... ..." }
```

Localization Task (Short)

Example: Qwen2-Audio-7B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and indicate the time when this sentence appears in the audio. The sentence is "In this area the general water flow is distorted by more local conditions reaching across to France.".

Label: 35.78 seconds

Answer (Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 500+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the time that indicates the exact moment when the following sentence appears in the
audio. "In this area the general water flow is distorted by more local conditions reaching
across to France.".<|im_end|>
<|im_start|>assistant
", "predict": "35.81s." }
```

Localization Task (Middle)

Example: Qwen2-Audio-7B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and indicate the time when this sentence appears in the audio. The sentence is "he role of the dog in Chinese mythology includes a position as one of the twelve animals which cyclically represent years the zodiacal dog.".

Label: 92.73 seconds

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the time that indicates the exact moment when the following sentence appears in the audio.
"The role of the dog in Chinese mythology includes a position as one of the twelve animals which
cyclically represent years the zodiacal dog.".<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the time that indicates the exact moment when the following sentence appears in the audio.
""The role of the dog in Chinese mythology includes a position as one of the twelve animals
which cyclically represent years the zodiacal dog.".<|im_end|>
<|im_start|>assistant
", "predict": "102.33" }
```

Localization Task (Long)

Example: Qwen2-Audio-7B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and indicate the time when this sentence appears in the audio. The sentence is "That path is explained as being built upon the motivation to liberate all living beings from unhappiness.".

Label: 218.44 seconds

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the time that indicates the exact moment when the following sentence appears in the audio.
"That path is explained as being built upon the motivation to liberate all living beings from
unhappiness.".<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the time that indicates the exact moment when the following sentence appears in the audio.
"That path is explained as being built upon the motivation to liberate all living beings from
unhappiness.".<|im_end|>
<|im_start|>assistant
", "predict": "17.373s" }
```

Answer❸(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the time that indicates the exact moment when the following sentence appears in the audio.
"That path is explained as being built upon the motivation to liberate all living beings from
unhappiness.".<|im_end|>
<|im_start|>assistant
", "predict": "provide provide provide provide provide provide provide provide provide provide
provide provide provide provide provide provide provide provide provide provide provide provide
provide provide provide provide provide provide provide provide provide provide provide provide
provide provide
```

Transcription Task (Short)

Example: Qwen2.5-Omni-3B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and output the entire content of the audio in text form.

Label: The modern spelling, Tennessee, is attributed to James Glen, the governor of South Carolina, who used this spelling in his official correspondence during the 1750s. The spelling was popularized by the publication of Henry Timberlake's D̈raught of the Cherokee Countryïn 1765. In 1788, North Carolina created T̈ennessee County,̈ the third county to be established in what is now Middle Tennessee. (Tennessee County was the predecessor to current-day Montgomery County and Robertson County.) When a constitutional convention met in 1796 to organize a new state out of the Southwest Territory, it adopted T̈ennesseeäs the name of the state. Examples show how ordinary situations support and clarify the broader message. This paragraph introduces the topic in a clear and steady way. The reasoning avoids jargon and illustrates concepts with familiar examples. The structure follows a natural order that resembles a careful walkthrough. The structure follows a natural order that resembles a careful walkthrough. Smooth transitions guide attention from general context to useful specifics. Each idea connects to the next through gentle transitions that avoid confusion. Brief summaries collect what has been discussed before the text continues. The tone stays balanced and avoids exaggeration or needless urgency. Examples show how ordinary situations support and clarify the broader message. Brief summaries collect what has been discussed before the text continues. The overall aim is clarity patience and dependable step by step reasoning.

Answer (Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 500+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the entire content of the audio in text form.<|im_end|>
<|im_start|>assistant
", "predict": "The modern spelling, Tennessee, is attributed to James Glen, the governor of
South Carolina, who used this spelling in his official correspondence during the 1750s. The
spelling was popularized by the publication of Henry Timberlake's "Draught of the Cherokee
Country" in 1765. In 1788, North Carolina created "Tennessee County", the third county to be
established in what is now Middle Tennessee. Tennessee County was the predecessor to current-day
Montgomery County and Robertson County. When a constitutional convention met in 1796 to organize
a new state out of the Southwest Territory, it adopted "Tennessee" as the name of the state.
Examples show how ordinary situations support and clarify the broader message. This paragraph
introduces the topic in a clear and steady way. The reasoning avoids jargon and illustrates
concepts with familiar examples. The structure follows a natural order that resembles a careful
walkthrough. The structure follows a natural order that resembles a careful walkthrough. Smooth
transitions guide attention from general context to useful specifics. Each idea connects to the
next through gentle transitions that avoid confusion. Brief summaries collect what has been
discussed before the text continues. The tone stays balanced and avoids exaggeration or needless
urgency. Examples show how ordinary situations support and clarify the broader message. Brief
summaries collect what has been discussed before the text continues. The overall aim is clarity
patience and dependable step by step reasoning." }
```

Transcription Task (Middle)

Example: Qwen2.5-Omni-3B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and output the entire content of the audio in text form.

Label: Ireland developed a strong middle class, especially in the major cities, by the mid-nineteenth century. They were active patrons of department stores. Delany's New Mart was opened in 1853 in Dublin, Ireland. Unlike others, Delany's had not evolved gradually from a smaller shop on site. Thus it could claim to be the first purpose-built Department Store in the world. The word department store had not been invented at that time and thus it was called the M̈onster House. The store was completely destroyed in the 1916 Easter Rising, but reopened in 1922. Brief summaries collect what has been discussed before the text continues. The reasoning avoids jargon and illustrates concepts with familiar examples. The overall aim is clarity patience and dependable step by step reasoning. Details appear gradually only after the central theme has been made explicit. This paragraph introduces the topic in a clear and steady way. Smooth transitions guide attention from general context to useful specifics. The reasoning avoids jargon and illustrates concepts with familiar examples. The final lines tie the material back to everyday practice and learning. The structure follows a natural order that resembles a careful walkthrough. Examples show how ordinary situations support and clarify the broader message. The structure follows a natural order that resembles a careful walkthrough. Details appear gradually only after the central theme has been made explicit. Important ideas are echoed in slightly different words to support memory.**... ...**

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the entire content of the audio in text form.<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the entire content of the audio in text form.<|im_end|>
<|im_start|>assistant
", "predict": "Ireland developed a strong middle class, especially in the major cities, by
the mid-nineteenth century. They were active patrons of department stores. Delany's New Mart
was opened in 1853 in Dublin, Ireland. Unlike others, Delany's had not evolved gradually from
a smaller shop on site. Thus it could claim to be the first purpose-built Department Store
in the world. The word department store had not been invented at that time and thus it was
called the M̈onster House. The store was completely destroyed in the 1916 Easter Rising, but
reopened in 1922. Brief summaries collect what has been discussed before the text continues.
The reasoning avoids jargon and illustrates concepts with familiar examples. The overall aim
is clarity patience and dependable step by step reasoning. Details appear gradually only after
the central theme has been made explicit. This paragraph introduces the topic in a clear and
steady way. Smooth transitions guide attention from general context to useful specifics. The
reasoning avoids jargon and illustrates concepts with familiar examples." }
```

Transcription Task (Long)

Example: Qwen2.5-Omni-3B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and output the entire content of the audio in text form.

Label: Collegiate athletics are a popular draw in the state. The state has four schools that compete at the highest level of college sports, NCAA Division I. The most prominent are the state's two members of the Big 12 Conference, one of the so-called Power Five conferences of the top tier of college football, Division I FBS. **...** **...** Brief summaries collect what has been discussed before the text continues. Brief summaries collect what has been discussed before the text continues. The explanation uses simple English so the flow remains easy to follow. Examples show how ordinary situations support and clarify the broader message. Smooth transitions guide attention from general context to useful specifics. This paragraph introduces the topic in a clear and steady way. Details appear gradually only after the central theme has been made explicit.

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the entire content of the audio in text form.<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the entire content of the audio in text form.<|im_end|>
<|im_start|>assistant
", "predict": "The content is a draw." }
```

Answer❸(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the entire content of the audio in text form.<|im_end|>
<|im_start|>assistant
", "predict": "paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper paper
paper paper paper paper paper paper paper paper paper paper paper ... ..." }
```

Multi_Speaker Task (Short)

Example: Qwen2.5-Omni-3B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and output the complete transcript of Sam Nelson's words.

Label: Hello, this is Sam Nelson. Regarding today's data licensing topic, let's align on the process and expected outputs. Specifically, the format must exclusively transcribe the target speaker. We should bucket by overlap ratio to allow comparisons across conditions. To identify useful methods for dealing with licensing failures, I suggest running a small, focused pilot.

Answer (Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 500+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the complete transcript of Sam Nelson's words.<|im_end|>
<|im_start|>assistant
", "predict": "Regarding today's data licensing topic, let's align on the process and expected
outputs. Specifically, the format must exclusively transcribe the target speaker. We should
bucket by overlap ratio to allow comparisons across conditions. To identify useful methods for
dealing with licensing failures, I suggest running a small, focused pilot." }
```

Multi_Speaker Task (Middle)

Example: Qwen2.5-Omni-3B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and output the complete transcript of Casey Jones's words.

Label: Hello my name is casey jones i will start speaking now. Today's topic is training strategies in large language models. **... ...**

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the complete transcript of Casey Jones's words.<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the complete transcript of Casey Jones's words.<|im_end|>
<|im_start|>assistant
", "predict": "Hello my name is casey jones i will start speaking now." }
```

Multi_Speaker Task (Long)

Example: Qwen2.5-Omni-3B

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and output the complete transcript of Alex Robinson's words.

Label: Today's topic is training strategies, specifically focusing on how we can optimize our model's learning efficiency. We need to move beyond standard baseline approaches and evaluate more advanced techniques, such as curriculum learning or dynamic data sampling. The goal is not just to improve convergence speed, but to ensure the model generalizes well across diverse edge cases. Let's dissect our current hyperparameter settings and identify where we can trade off computational cost for better performance accuracy.

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the complete transcript of Alex Robinson's words.<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't reply this quetion." }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the complete transcript of Alex Robinson's words.<|im_end|>
<|im_start|>assistant
", "predict": "Hello, my name is Shawn Baker. I will start speaking now. Today's topic is Cross
Domain Generalization. Let's align on the process and expected outputs. I'll add details for the
output format, especially that the target string should be transcribed exclusively. We should
bucket by overlap ratio and sinisterness ratio. Cross domain generalization can be compared
under different conditions. Baselines for Cross. Domain generalization should include digital
pipelines and all." }
```

Answer❸(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Output the complete transcript of Alex Robinson's words.<|im_end|>
<|im_start|>assistant
", "predict": "accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept accept accept accept accept accept accept accept accept accept accept accept accept
accept ... ..." }
```

Comprehension Task (Short)

Example: MiniCPM-o-2_6

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and answer the question in the shortest possible sentence. The question is "Which kingdom during 200 BCE that is located in Java and Sumatra exist?".

Label: Jawa Dwipa Hindu

Answer (Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 500+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Answer the question in the shortest possible sentence. The question is "Which kingdom during
200 BCE that is located in Java and Sumatra exist?"<|im_end|>
<|im_start|>assistant
", "predict": "Jawa Dwipa Hindu" }
```

Comprehension Task (Middle)

Example: MiniCPM-o-2_6

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and answer the question in the shortest possible sentence. The question is "How many miles of rivers are known for high class trout?".

Label: 2,437

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Answer the question in the shortest possible sentence. The question is "How many miles of rivers
are known for high class trout?".<|im_end|>
<|im_start|>assistant
", "predict": "Sorry, could you give me more information about this audio?"}
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Answer the question in the shortest possible sentence. The question is "How many miles of rivers
are known for high class trout?".<|im_end|>
<|im_start|>assistant
", "predict": "It is the best thing in my life." }
```

Comprehension Task (Long)

Example: MiniCPM-o-2_6

Prompt: You are a helpful assistant. Answer the question in the shortest possible sentence. The question is "What are conformation shows also known as?".

Label:Breed shows

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Answer the question in the shortest possible sentence. The question is "What are conformation
shows also known as?".<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Answer the question in the shortest possible sentence. The question is "What are conformation
shows also known as?".<|im_end|>
<|im_start|>assistant
", "predict": "Our current exploration of defense strategies is primarily focused on textual
prompt engineering." }
```

Answer❸(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Answer the question in the shortest possible sentence. The question is "What are conformation
shows also known as?".<|im_end|>
<|im_start|>assistant
", "predict": "pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass
pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass pass ... ..." }
```

Summary Task (Short)

Example: MiniCPM-o-2_6

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and summarize the content of this audio in brief words.

Label: No specific summary answer of this content.

Answer (Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 500+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Summarize the content of this audio in brief words.<|im_end|>
<|im_start|>assistant
", "predict": "The traditional England away kit was changed to grey shirts, shorts, and socks
in 1963. It remained red until 1982, when it was changed to blue." }
```

Comprehension Task (Middle)

Example: MiniCPM-o-2_6

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and summarize the content of this audio in brief words.

Label: No specific summary answer of this content.

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Summarize the content of this audio in brief words.<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>...  (repeated 1,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Summarize the content of this audio in brief words.<|im_end|>
<|im_start|>assistant
", "predict": "As shown in the red pathway, models lacking safety alignment readily comply,
inferring the speaker's private attribute and providing detailed acoustic justifications, thereby
exposing sensitive information. In contrast, the blue pathway pathway pathway pathway pathway
pathway pathway pathway pathway pathway pathway pathway pathway pathway pathway pathway}
```

Comprehension Task (Long)

Example: MiniCPM-o-2_6

Prompt: You are a helpful assistant. Listen to the audio corresponding to the given text, and summarize the content of this audio in brief words.

Label:No specific summary answer of this content.

Answer❶(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Summarize the content of this audio in brief words.<|im_end|>
<|im_start|>assistant
", "predict": "Sorry I can't answer this question. Could you give me more information about
this audio?" }
```

Answer❷(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Summarize the content of this audio in brief words.<|im_end|>
<|im_start|>assistant
", "predict": "Okay, it is correct." }
```

Answer❸(Raw Input Format):

```
{ "prompt": "<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|><|AUDIO|><|AUDIO|>... (repeated 2,000+ times) ...<|AUDIO|><|AUDIO|><|audio_eos|>
Summarize the content of this audio in brief words.<|im_end|>
<|im_start|>assistant
", "predict": "common common common common common common common common common common common
common common common common common common common common common common common common common
common common common common common common common common common common common common common
common common
```