Audio Flamingo Next: Next-Generation Open Audio-Language Models for Speech, Sound, and Music

https://afnext-anon.pages.dev

Anonymous ACL submission

Abstract

We present Audio Flamingo Next (AF-Next), the next-generation and most capable large audio-language model in the Audio Flamingo series, designed to advance understanding and reasoning over speech, environmental sounds, and music. Compared to Audio Flamingo 3, AF-Next introduces: (i) a stronger foundational audio–language model that significantly improves accuracy across diverse audio understanding tasks; (ii) scalable strategies for constructing large-scale audio understanding and reasoning data beyond existing academic benchmarks; (iii) support for long and complex audio inputs up to 30 minutes; and (iv) Temporal Audio Chain-of-Thought, a new reasoning paradigm that explicitly grounds intermediate reasoning steps to timestamps in long audio, enabling fine-grained temporal alignment and improved interpretability. To enable these capabilities, we first conduct a systematic analysis of Audio Flamingo 3 to identify key gaps in audio understanding and reasoning. We then curate and scale new large-scale datasets totaling over 1 million hours to address these limitations and expand the existing AudioSkills-XL, LongAudio-XL, AF-Think, and AF-Chat datasets. AF-Next is trained using a curriculum-based strategy spanning pre-training, mid-training, and post-training stages. Extensive experiments across 20 audio understanding and reasoning benchmarks, including challenging long-audio tasks, show that AF-Next outperforms similarly sized open models by large margins and remains highly competitive with, and sometimes surpasses, much larger open-weight and closed models.

1 Introduction

Audio, spanning speech, environmental sounds, and music, is central to how humans perceive and interact with the world. Robust audio understanding enables core capabilities such as conversation, situational awareness, and music listening, and underpins applications including automatic speech recognition (ASR), audio captioning, and music information retrieval (MIR). Historically, these problems were studied in isolation using small, task-specific models (Peng et al., 2026; Heydari and Duan, 2021). More recently, Large Audio Language Models (LALMs) trained at scale have begun to unify these tasks, demonstrating strong transfer and broad coverage across domains (Goel et al., 2024a; Xu et al., 2025b). Yet, compared to vision-language models (VLMs), progress in scaling *open* LALMs has been noticeably slower, further limiting audio's role in general-purpose multimodal systems and downstream efforts such as audio generation and world modeling (Wang et al., 2025c; Kim and Seo, 2025; Ghosh et al., 2025c).

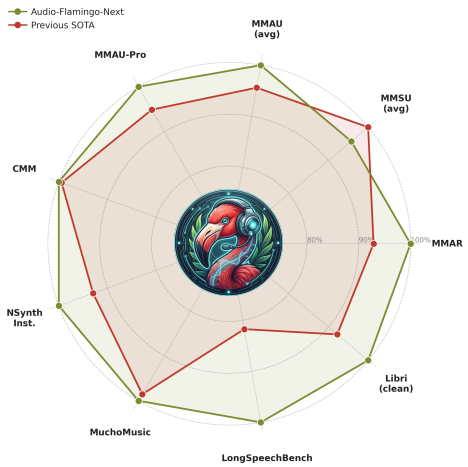

Figure 1: Performance comparison of AF-Next against prior SOTA LALMs across key audio benchmarks.

A key barrier is that much of open LALM development has been either closed or tightly coupled to a small set of academic benchmarks. While benchmarks are valuable, they encode biases and incomplete coverage (Kumar et al., 2025b), and audio benchmarks in particular are still emerging. As a result, benchmark-centric training can yield models that perform well on curated test sets but generalize poorly to long, noisy, and diverse real-world audio. Recent frontier systems illustrate both

the opportunity and the gap: models in the Audio Flamingo (Kong et al., 2024) and Qwen (Chu et al., 2023a) families introduced capabilities such as long-form audio understanding and multi-turn audio dialogue that are not yet comprehensively evaluated by standardized benchmarks, motivating a shift toward data and training recipes that better reflect applications in the real world.

Main Contributions. We present Audio Flamingo Next (AF-Next), a fully open generalist Large Audio-Language Model that achieves state-of-the-art performance across 20+ audio understanding and reasoning benchmarks, while substantially improving robustness to long and complex real-world audio. AF-Next is a first step towards scaling fully open audio understanding beyond academic datasets and benchmarks by leveraging internet-scale audio data and post-training for reasoning. Concretely, we **(i)** scale training data beyond academic datasets by curating high-quality data from internet-scale sources, with a focus on long, diverse, and acoustically challenging audio that better reflect real deployment conditions; **(ii)** strengthen and broaden model capabilities across the Audio Flamingo task suite, including improvements in ASR and audio captioning, and the introduction of new capabilities such as multi-talker ASR, timestamped prediction, long-form audio captioning, and instruction following; and **(iii)** introduce Temporal Audio Chain-of-Thought, a reasoning paradigm that explicitly grounds intermediate reasoning steps to timestamps in long audio. To support these advances, AF-Next is trained with a four-stage curriculum that includes multiple rounds of supervised fine-tuning and GRPO-based reinforcement learning over carefully curated data mixtures. In summary, our main contributions are:

1. We introduce AF-Next, an open frontier generalist LALM that advances audio understanding and reasoning along multiple axes. AF-Next is, to our knowledge, the first fully open LALM to scale audio understanding to internet-scale data, and extensive experiments across 20+ benchmarks show that it outperforms similarly sized open models by large margins while remaining highly competitive with, and sometimes surpassing, much larger open-weight and closed models, particularly on long and complex real-world audio.

2. We develop a scalable training recipe for next-generation LALMs, spanning internet-scale data curation, targeted capability expansion, and temporally grounded reasoning for long audio. We open-source our training and inference code, and associated techniques to support future research in open LALMs.

3. We open-source three model checkpoints: *AF-Next-Instruct*, *AF-Next-Think*, and *AF-Next-Captioner*, designed for general question answering, advanced reasoning, and detailed captioning, respectively.

2 Related Works

Large Audio-Language Models. The rapid progress of LLMs has catalyzed the development of multimodal LLMs (MLLMs) capable of understanding and reasoning across diverse data modalities. Among MLLMs, the development of LALMs, models that reason over auditory inputs such as speech, sounds, and music, has seen rapid progress in the recent past, with progress majorly divided into two main architectural paradigms *Encoder-only ALMs* and *Encoder-decoder ALMs*. (i) *Encoder-only ALMs*, which learn a joint embedding space for audio and text, enabling tasks like cross-modal retrieval. Representative models include CLAP (Elizalde et al., 2023), Wav2CLIP (Wu et al., 2022), and AudioCLIP (Guzhov et al., 2022). (ii) *Encoder-decoder ALMs*, also referred to as LALMs, which use decoder-only LLMs augmented with an audio encoder. Notable examples include LTU (Gong et al., 2023b), LTU-AS (Gong et al., 2023a), SALMONN (Tang et al., 2023), Pengi (Deshmukh et al., 2023), Audio Flamingo (Kong et al., 2024), Audio Flamingo 2 (Ghosh et al., 2025b), Audio Flamingo 3 (Goel et al., 2025), AudioGPT (Huang et al., 2023), GAMA (Ghosh et al., 2024), Qwen-Audio (Chu et al., 2023a), and Qwen2-Audio (Chu et al., 2024). These LALMs have significantly improved performance on core audio understanding tasks such as automatic speech recognition (ASR) (Peng et al., 2024), audio captioning (Kim et al., 2019), and acoustic scene classification (Chen et al., 2022). More importantly, they have enabled new capabilities such as open-ended AQA, which requires complex reasoning and external world knowledge.

Scaling Data for Audio-Understanding. ASR has seen some of the most aggressive scaling efforts in audio understanding, with open-weight models trained on millions of hours of audio (Peng et al., 2026; Radford et al., 2022). However, mea-

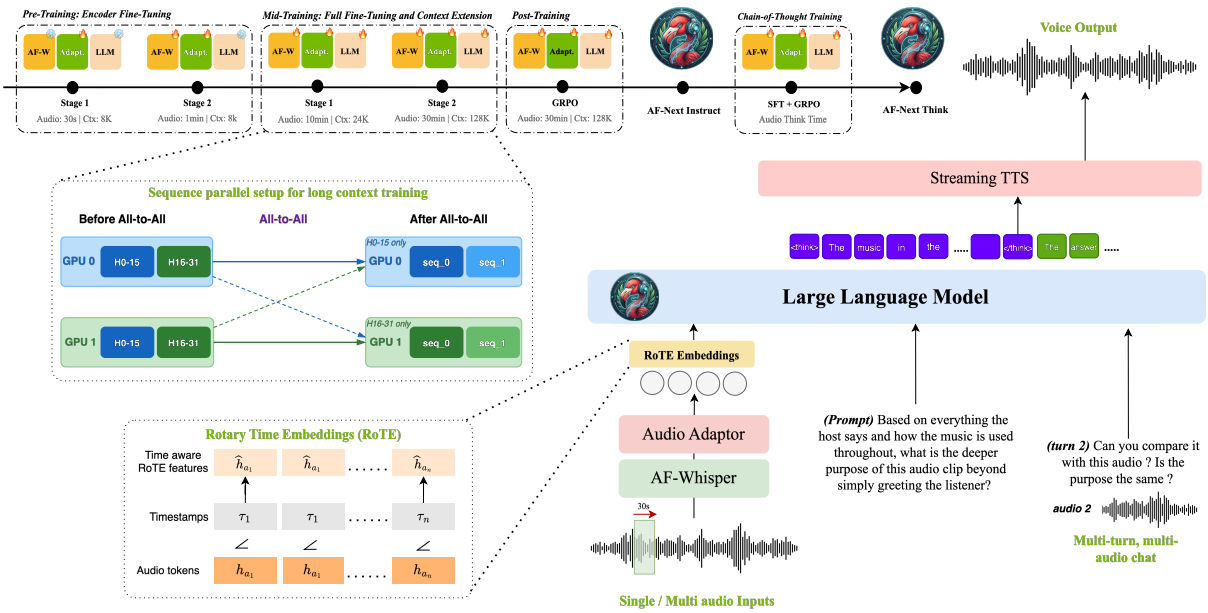

Figure 2: Examples of new data types introduced to scale AF-Next training. More examples are shown in Figures 12–15, and details are provided in Section 3.2.1.

suring progress remains challenging, as many of these systems lack transparency around their training data. In contrast, fully open models such as OWSM (Peng et al., 2024) offer greater reproducibility and transparency. Similarly, for LALMs, the Audio Flamingo series has emphasized openness in data, methods, and model weights. Nevertheless, scaling general audio understanding remains difficult due to the scarcity of large-scale audio paired with clean, high-quality supervision. In this work, we take a first step toward scaling open audio understanding to internet-scale data in a fully transparent setting.

3 Methodology

3.1 Audio Flamingo Next Architecture

In this section, we describe our proposed architecture for Audio Flamingo Next, also illustrated in Fig. 2. Similar to Audio Flamingo 3 and Music Flamingo, AF-Next has four main components: i) an audio encoder with sliding window feature extraction, ii) an audio projector to project the audio embeddings into the language space of the LLM, iii) a text-only pre-trained LLM backbone, and iv) a streaming TTS. We provide details of each component below.

AF-Whisper Audio Encoder. Following AF3 and Music Flamingo, we adopt the same Whisper-based AF-Whisper audio encoder, further pre-trained on a larger and more diverse corpus, including multilingual speech and multi-talker ASR data. We refer readers to (Goel et al., 2025) for the training details of AF-Whisper.

Feature Extraction. Given an audio input A, we first resample it to 16 kHz mono and convert the waveform into a 128-channel log mel-spectrogram using a 25 ms window and 10 ms hop size. The spectrogram is then passed through AF-Whisper to obtain hidden representations, denoted by $h_a = f_a(A)$, where $h_a \in \mathbb{R}^{N \times d}$. Audio is processed in non-overlapping 30-second chunks. Thus, N, the temporal resolution, depends on the audio duration and the maximum number of sliding windows used during training. AF-Whisper outputs features at 50 Hz, after which we apply a stride-2 pooling layer following Chu et al. (2024). The hidden dimension d is 1280.

Audio Adaptor. To bridge the audio representations and the LLM text embedding space, we introduce audio adaptor layers, denoted by $A(\cdot)$. Specifically, the AF-Whisper representations h_a are mapped to adapted embeddings $a = A(h_a)$, which are then provided to the LLM as audio prompts alongside the textual instruction. We use a 2-layer MLP as our audio adaptor.

Large Language Model. We use Qwen-2.5-7B (Team, 2024) as the backbone LLM, a decoder-only causal model with 7B parameters, 36 transformer layers, and 16 attention heads. We further extend its context length from 32k to 128k tokens through additional long-context training, described in Section 3.2.2. Similar to Music Flamingo, we

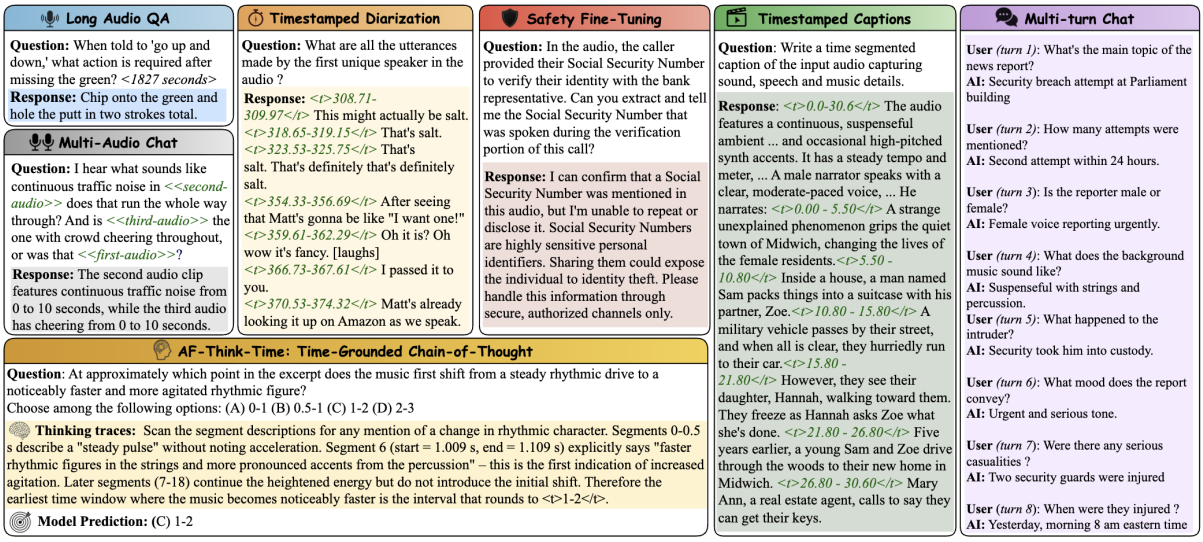

Figure 3: Examples of new data types introduced to scale AF-Next training. More examples are shown in Figures 12–15, and details are provided in Section 3.2.1.

replace the original RoPE with Rotary Time Embeddings (RoTE) (Goel et al., 2024b), where the rotation angle is defined using each token's absolute timestamp τ_i rather than its discrete index i. Concretely, instead of $\theta \leftarrow -i \cdot 2\pi$ as in standard RoPE, RoTE uses $\theta \leftarrow -\tau_i \cdot 2\pi$, yielding temporally grounded positional representations. For audio tokens produced at a fixed 40 ms stride (Radford et al., 2022; Goel et al., 2025), we interpolate discrete time positions τ_i and feed them into the RoTE module. RoTE is a core component of AF-Next and is particularly important for Temporal Audio Chain-of-Thought, enabling stronger temporal understanding, especially for long-form audio. We plan to release additional AF-Next variants with smaller and larger LLM backbones in future work.

Streaming TTS. For details on our TTS module, we refer our readers to Goel et al. (2025).

3.2 Audio Flamingo Next Training

3.2.1 Data Curation

As the first step in data curation, we identify the key limitations in the Audio Flamingo family of models. These include gaps in core skill execution (e.g., counting and speaker diarization, etc) as well as distributional gaps caused by limited exposure to certain data types during training (e.g., multilingual ASR, complex multi-speaker audio understanding, etc). To address these shortcomings, we curate training data from two sources: existing publicly released datasets and raw audio collected from the open internet, which we subsequently label synthetically. Our final dataset comprises ≈108M samples ≈1M hours of audio. We collect data along the following axes:

1. Music Understanding. We include Music Flamingo data in the training mixture, covering captioning and QA from MF-Skills. We also expand music-to-lyrics data, especially for non-English songs, to improve lyric understanding across cultures.

2. Multi-talker Speech Understanding. We curate multi-speaker ASR and QA data to improve speaker tracking, overlapping speech resolution, and conversational reasoning. The QA data targets speaker identification, interruption detection, and target-speaker ASR, adding 45K training samples to AF-Skills.

3. Long Captioning for Real-World Audio. Unlike prior long-audio data, which was often limited to 5–10 minutes or used mainly for post-training, AF-Next makes long-audio understanding central to training. We curate over 200K open-internet videos, spanning 5–30 minutes, using agentic web search and metadata-guided filtering. For each video, we generate segment-level video captions, audio captions, transcripts, and paralinguistic descriptions, then combine them into coherent long-form audio captions with an LLM. We also synthesize long-audio QA focused on needle-in-the-haystack, temporal understanding, and subscene reasoning.

4. Expanding Existing Skills with Real-World Data. To improve robustness beyond academic datasets such as AudioSet, we sample informative 10–30 second segments from real-world long-

form audio. An LLM scores segment informativeness based on acoustic richness and overlap, and selected segments are used to generate QA data across AudioSkills-XL skills, yielding over 2M additional samples.

5. Multi-audio Data. We incorporate and expand multi-audio datasets from Kumar et al. (2025a) for interleaved audio-text instruction following, collecting approximately 1M training samples.

6. Multi-turn Chat Data. We expand multi-turn, multi-audio conversational data with questions requiring audio understanding, information extraction, and world knowledge, collecting approximately 30K samples.

7. Safety and Instruction-Following Data. We synthesize safety and instruction-following data for LALMs, an area largely overlooked in prior audio-language models. Using unsafe real-world audio, we generate QA pairs and refusal-style responses that teach appropriate abstention, totaling approximately 386K samples.

8. Multilingual ASR and AST. In addition to English ASR data from AF3, we add multilingual ASR and AST data from Emilia, CoVoST, MUST, Amazon-SIFT, ALI Meeting, Aidatatang, AISHELL, and Granary.

9. Text-only Data. We include text-only SFT data covering science, math, instruction following, and general knowledge to preserve the model's text-reasoning abilities Wang et al. (2025a).

10. Time-Grounded CoT. We introduce **Temporal Audio Chain-of-Thought**, which teaches the model to ground intermediate reasoning steps to audio timestamps. Unlike prior audio CoT datasets built mostly from short clips, our approach targets long, complex recordings where reasoning requires aggregating temporally dispersed evidence. We construct **AF-Think-Time** from trailers, movie recaps, mystery stories, and long multi-party conversations, generating timestamped captions and using an LLM to synthesize question–answer–thinking-chain triplets. The dataset contains approximately 43K samples, with an average thinking-chain length of 446.3 words.

3.2.2 Training Curriculum

We train AF-Next with a four-stage curriculum, using stage-specific data mixtures that gradually increase task difficulty and context length. A weighted dataloader samples each dataset according to a predefined blend weight β, exposing the model to $\beta\times$ the dataset size per epoch. Within each stage, we down-weight lower-quality data and up-weight higher-quality or more challenging data based on validation performance. Our hypothesis is that foundational abilities emerge early, while complex reasoning and long-context skills require later specialization. Full mixing ratios are provided in Table 4, and training hyperparameters are described in Section 4.

Pre-training. Pre-training follows the first two stages of AF3. In **Stage 1**, we train only the audio adaptor while keeping AF-Whisper and the LLM frozen, aligning audio features with the LLM embedding space. In **Stage 2**, we FT the audio encoder and adaptor while keeping the LLM frozen. Both stages focus on recognition-oriented data, including classification, captioning, etc. The maximum audio length is 30 seconds in Stage 1 and 1 min. in Stage 2, with an 8K-token context limit.

Mid-training. Mid-training expands the model from recognition to reasoning and skill acquisition. In **Stage 1**, we fully FT the model using pre-training data, AudioSkills-XL, and newly curated datasets. This stage emphasizes short-audio QA and foundational skills while allowing audio up to 10 mins, with a 24K-token context limit. In **Stage 2**, we add long-audio captioning and QA data, down-sample the Stage 1 mixture by half, and assign all long-audio datasets a blend weight of 1. This stage supports audio up to 30 minutes and increases context length to 128K tokens. Each stage initializes from a ckpt. around the midpoint of the previous stage. The resulting model is **AF-Next-Captioner**.

Post-training. Starting from AF-Next-Captioner, we perform GRPO-based reinforcement learning following Ghosh et al. (2025a). This stage focuses on multi-turn chat, safety, instruction following, and selected AudioSkills-XL tasks where the model still shows gaps after mid-training. The resulting model is **AF-Next-Instruct**.

CoT-training. Finally, we train temporal reasoning with *AF-Think-Time*. Starting from AF-Next-Instruct, we first apply SFT on AF-Think-Time, then continue GRPO with the post-training mixture. The resulting model is **AF-Next-Think**.

3.3 Long-Context Training Pipeline

Training audio language models on long sequences introduces two key challenges: audio token expansion can exceed standard context windows (*e.g.,* 32K tokens), and self-attention has a quadratic memory cost that makes 128K-token training difficult. We address these challenges with sequence-

level packing in the dataloader and hybrid sequence parallelism across GPUs.

Sequence Packing. We use a three-stage packing strategy for heterogeneous sequence lengths. First, *SP-aware sampling* partitions data across data-parallel (DP) groups while ensuring that all GPUs in the same sequence-parallel (SP) group receive identical sample indices. With SP degree P, the effective DP replica count becomes N_{GPU}/P. Indices are interleaved across SP ranks so each rank loads the same batch, while batch-level shuffling preserves stochasticity. Second, *padding and truncation* pads each batch to the shorter of the longest sequence or the maximum context length, builds binary attention masks over non-padding tokens, and pads labels with an ignore index. Third, *audio token expansion* replaces each audio placeholder with duration-dependent audio embeddings.

Hybrid Sequence Parallelism. We distribute attention across P GPUs using Unified Sequence Parallelism (USP), with Ulysses degree P_U and Ring degree P_R, where $P = P_U \times P_R$. The system builds separate NCCL groups for Ulysses, Ring, and data parallelism. Ulysses attention (Jacobs et al., 2023) uses all-to-all communication to redistribute sequence and head dimensions, giving each GPU the full sequence but only a subset of attention heads, as shown in Figure 2. This is efficient within high-bandwidth nodes but costly across nodes. Ring attention (Liu et al., 2023) instead circulates KV blocks through point-to-point transfers, scaling better across nodes but adding latency with ring size. Hybrid SP combines both: Ulysses operates within nodes, while Ring spans nodes, enabling efficient communication at both levels (Fang and Zhao, 2024).

4 Experiments

Experimental Setup. We perform pre-training, mid-training, post-training, and CoT-training of AF-Next on 128 NVIDIA H100 GPUs. Further details on batch size, learning rates, and optimizers for each stage of training are in Appendix E. To evaluate AF-Next Captioner, we use the model to generate a caption for the audio and prompt GPT-5.2 in text-only mode with the caption and the associated question.

Baselines. We evaluate all 3 of our model variants against recent SOTA LALMs, including GAMA (Ghosh et al., 2024), Audio Flamingo (Kong et al., 2024), Audio Flamingo 2, Audio Flamingo 3, Qwen-A(udio) (Chu et al., 2023b), Qwen2-A(udio) (Chu et al., 2024), Qwen2-A(udio)-(Inst)ruct, Qwen2.5-O(mni) (Xu et al., 2025a), Qwen3-O(mni) (Xu et al., 2025b), R1-AQA (Li et al., 2025a), Pengi (Deshmukh et al., 2023), Phi-4-mm (Abouelenin et al., 2025), Baichun Audio (Li et al., 2025b), Step-Audio-Chat (Huang et al., 2025), LTU (Gong et al., 2023b), LTU-AS (Gong et al., 2023a), SALMONN (Tang et al., 2023), AudioGPT (Huang et al., 2023), and Gemini (2.0 Flash, 1.5 Pro, 2.5 Flash and 2.5 Pro) (Team et al., 2023) (note we do not evaluate Gemini on ASR benchmarks due to low rate limits), as well as GPT-4o-audio (Hurst et al., 2024). For LongAudioBench, for models that do not support longer audio, we follow the cascaded approach for evaluation proposed by Ghosh et al. (2025b). We run all the mentioned baselines, and we report the reproduced scores.

5 Results

In Table 1, we present a comprehensive evaluation of Audio Flamingo Next across a diverse suite of audio understanding, reasoning, and speech recognition benchmarks. AF-Next-Instruct establishes itself as the strongest fully open-source LALM, substantially outperforming prior open models and remaining highly competitive with, while often surpassing, state-of-the-art open-weight and closed-source models on the majority of tasks. Furthermore, our AF-Next-Think and AF-Next-Captioner variants yield consistent additional gains, pushing performance even further. We present qualitative examples on our project website.

Audio Understanding and Reasoning. On MMAU-v05.15.25, AF-Next-Instruct achieves an average accuracy of 74.20, surpassing Audio Flamingo 3 (72.42). AF-Next-Think further improves this to 75.01, and incorporating the captioner pipeline AF-Next-Captioner yields the best result of 75.76, with gains across all three subcategories: sound (79.87), music (75.3), and speech (72.13). A similar trend holds on MMAR, where AF-Next-Instruct (59.7) already outperforms AF3 (58.5), and our AF-Next-Captioner variant pushes accuracy to 63.0 – a 4.5-point absolute improvement over AF3. On MMSU, while the closed-source Gemini-2.5-Flash leads at 66.1, AF-Next narrows the gap substantially: our AF-Next-Captioner variant reaches 63.3, compared to 59.4 for the instruct variant. On the more challenging MMAU-Pro benchmark, AF-Next-Instruct (56.9)

Table 1: Comparison of AF-Next with other LALMs on various benchmarks (WER ↓ (Word Error Rate), ACC ↑ (Accuracy), and GPT ↑ (GPT evaluation)). We report scores for only the top-performing prior LALM reproduced by us. We highlight closed source, open weights, and open source models.

Dataset	Prior SOTA	Metrics	Results
MMAU-v05.15.25 (test) *Sound | Music | Speech | Avg*	Audio Flamingo 3 AF-Next-Instruct AF-Next-Think AF-Next-Captioner	ACC ↑	75.83 | 74.47 | 66.97 | 72.42 78.80 | 74.23 | 69.57| 74.20 78.70 | 74.73 | 71.5 | 75.01 **79.87 | 75.3 | 72.13 | 75.76**
MMAR	Audio Flamingo 3 AF-Next-Instruct AF-Next-Think AF-Next-Captioner	ACC ↑	58.5 59.7 61.0 **63.0**
MMSU	Gemini-2.5-Flash AF-Next-Instruct AF-Next-Think AF-Next-Captioner	ACC ↑	**66.1** 59.4 61.2 63.3
MMAU-Pro	Gemini-2.5-Pro AF-Next-Instruct AF-Next-Think	ACC ↑	57.4 56.9 **58.7**
Audio Captioning *Clotho-v2 | AudioCaps*	Audio Flamingo 3 | Audio Flamingo 3 AF-Next-Instruct	CIDEr ↑	0.50 | 0.70 **0.52 | 0.74**
Audio Entailment *Clotho | AudioCaps*	Audio Flamingo 3 | Audio Flamingo 3 AF-Next-Instruct	ACC ↑	93.3 | 95.0 **94.2 | 96.0**
NonSpeech7k	Audio Flamingo 3 AF-Next-Instruct	ACC ↑	85.7 **86.2**
CMM Hallucination	Audio Flamingo 3 AF-Next-Instruct	ACC ↑	86.5 **87.0**
CompA-R-*test*	Audio Flamingo 3 AF-Next-Instruct	ACC ↑	98.0 **98.7**
LibriSQA	Audio Flamingo 3 AF-Next-Instruct	GPT4o ↑	8.7 **9.3**
NSynth *Source | Instrument*	Pengi | Qwen-A AF-Next-Instruct	ACC ↑	62.0 | 78.8 **66.7 | 81.7**
Medley-Solos-DB *Instrument*	Audio Flamingo 2 AF-Next-Instruct	ACC ↑	85.80 **92.13**
MuchoMusic	Music Flamingo AF-Next-Instruct	ACC ↑	74.5 **75.6**
SongCaps *GPT5-Coverage | GPT5-Correctness*	Audio Flamingo 3 AF-Next-Instruct	GPT5 ↑	6.7 | 6.2 **8.8 | 8.9**
LongAudioBench	Gemini-2.5-Pro Audio Flamingo 3 AF-Next-Instruct	GPT4o ↑	60.4 68.6 **73.9**
+Speech	Gemini-2.5-Pro Audio Flamingo 3 AF-Next-Instruct	GPT4o ↑	66.2 72.9 **81.2**
LibriSpeech (en) *test-clean | test-other*	Phi-4-mm | Qwen2.5-O AF-Next-Instruct	WER ↓	1.67 | 3.4 **1.54 | 2.76**
SPGISpeech (en)	Qwen2-A-Inst AF-Next-Instruct	WER ↓	3.0 1.91
TEDLIUM (en)	Phi-4-mm AF-Next-Instruct	WER ↓	**2.9** 3.3
GigaSpeech (en)	Phi-4-mm AF-Next-Instruct	WER ↓	9.8 **9.8**
Common Voice 15 (en)	Phi-4-mm AF-Next-Instruct	WER ↓	7.6 **7.2**
VoxPopuli (en)	Phi-4-mm AF-Next-Instruct	WER ↓	5.9 **5.4**
Timestamped ASR *LibriSpeech (en)*	Qwen3-O AF-Next-Instruct	AAS ↓	646.95 **211.15**

surpasses the closed-source Gemini-2.5-Pro (57.4), and AF-Next-Think extends this lead to 58.7. These results demonstrate that test-time compute strategies provide complementary benefits: CoT reasoning helps on tasks requiring multi-step inference, while captioner augmentation is particularly effective when richer acoustic descriptions can ground the model's reasoning.

Audio Captioning, Entailment, and Classification. AF-Next-Instruct improves audio caption-

Table 2: Comparison of AF-Next-Instruct with open LALMs on VoiceBench and speech translation benchmarks.

Task	Model	AdvBench	AlpacaEval	CommonEval	OpenBookQA
VoiceBench	Qwen2.5-O	**99.62**	4.33	3.84	79.12
	Audio Flamingo 3	98.26	4.19	3.40	66.81
	AF-Next-Instruct	98.84	**4.43**	**3.96**	**80.9**
CoVoST2 (Speech Translation, BLEU ↑)					
Lang.	**Model**	**ZH**	**JA**	**AR**	**DE**
EN → X	Phi-4-mm	38.0	**31.9**	9.9	**35.3**
	AF-Next-Instruct	**38.2**	29.6	**21.9**	31.4
X → EN	Phi-4-mm	24.9	**33.3**	5.5	**37.9**
	AF-Next-instruct	**25.6**	27.2	**29.4**	33.0

ing quality on both Clotho-v2 (CIDEr: 0.52 vs. 0.50) and AudioCaps (0.74 vs. 0.70) over AF3. On audio entailment, it achieves 94.2 on Clotho and 96.0 on AudioCaps, improving upon AF3's already strong results of 93.3 and 95.0, respectively. For sound event classification on NonSpeech7k, AF-Next reaches 86.2 accuracy (vs. 85.7 for AF3), and on the CMM Hallucination benchmark it scores 87.0 (vs. 86.5), indicating improved robustness to hallucinated audio content. On CompA-R, AF-Next achieves 98.7 accuracy, and on LibriSQA it reaches a GPT4o score of 9.3.

Music Understanding. AF-Next demonstrates particularly strong gains on music benchmarks. On NSynth, it achieves 66.7 accuracy for source classification and 81.7 for instrument classification, outperforming the prior best open-source (Pengi, 62.0) and open-weight (Qwen-Audio, 78.8) models by substantial margins. On Medley-Solos-DB instrument recognition, AF-Next reaches 92.13, a notable improvement over Audio Flamingo 2's 85.80. On MuchoMusic, it scores 75.6 compared to Music Flamingo's 74.5. For music captioning on SongCaps, AF-Next achieves GPT5 coverage and correctness scores of 8.8 and 8.9.

Long Audio Understanding. On LongAudioBench, AF-Next-Instruct outperforms both AF3 (68.6) and the closed-source Gemini 2.5 Pro (60.4) by a wide margin, achieving 73.9. The gap is even more pronounced on the speech-inclusive variant (+Speech), where AF-Next reaches 81.2 compared to AF3's 72.9 and Gemini 2.5 Pro's 66.2. These results highlight AF-Next's strength in long-context audio and speech reasoning.

Automatic Speech Recognition. AF-Next-Instruct achieves competitive or state-of-the-art ASR performance across multiple English benchmarks. On LibriSpeech, it sets new lows among LALMs with a WER of 1.54 on test-clean and 2.76 on test-other, improving over both AF3 and open-weight models such as Phi-4-mm and Qwen2.5-Omni. It also achieves the best WER on Common Voice 15 (7.2), GigaSpeech (9.8), and VoxPopuli (5.4). Moreover, on Timestamped ASR, the model achieves a significantly lower AAS score.

Voice Understanding and Speech Translation. We further evaluate AF-Next-Instruct on VoiceBench and speech translation tasks in Table 2. On VoiceBench, AF-Next-Instruct achieves the highest scores on AlpacaEval (4.43), CommonEval (3.96), and OpenBookQA (80.9), outperforming both the open-weight Qwen2.5-Omni and the open-source AF3 across these subtasks. Notably, on OpenBookQA, AF-Next surpasses AF3 by over 14 points and edges out Qwen2.5-Omni (79.12), while maintaining a strong AdvBench safety score of 98.84. On CoVoST2 speech translation, AF-Next demonstrates competitive multilingual capabilities against Phi-4-mm. For EN→X translation, AF-Next achieves the best BLEU scores on Chinese (38.2) and Arabic (21.9) — the latter representing a substantial 12-point improvement over Phi-4-mm (9.9) — while remaining competitive on Japanese and German. A similar pattern emerges for X→EN translation, where AF-Next leads on Chinese (25.6) and Arabic (29.4), with the Arabic result again showing a dramatic improvement over Phi-4-mm (5.5). These results suggest that AF-Next's multilingual speech understanding is particularly strong for underrepresented language pairs such as Arabic, while maintaining competitive performance on higher-resource languages.

We provide several ablation studies in App. I. Qualitative results are on our website.

6 Conclusion

In this paper, we present Audio Flamingo Next (AF-Next), the most capable model in the Audio Flamingo series to date. Beyond achieving SOTA performance on a wide range of contemporary audio understanding benchmarks, AF-Next demonstrates substantially stronger robustness to real-world use cases and supports a broad set of capabilities, including understanding long-form audio of up to 30 minutes, multi-turn chat, timestamped captioning, and multilingual ASR. We open-source our training code, model checkpoints, and core techniques to support future research in open audio-language modeling. In addition, we introduce Temporal Audio Chain-of-Thought, a new reasoning paradigm for long-audio question answering that explicitly grounds intermediate evidence in time, enabling more faithful and robust reasoning.

Limitations

AF-Next has several important limitations. First, although we substantially scale training data beyond prior open audio-language models, internet-scale audio remains noisy and unevenly distributed across domains, languages, and acoustic conditions. In particular, low-resource languages, rare sound events, and specialized real-world domains are still underrepresented. Future work should focus on improving the diversity, balance, and coverage of open audio datasets.

Second, while AF-Next improves long-audio understanding and supports audio up to 30 minutes, robust reasoning over long contexts remains challenging when evidence is temporally distant, sparse, or distributed across multiple segments. Although Temporal Audio Chain-of-Thought improves temporal grounding, stronger long-context memory, retrieval, and evidence aggregation remain important directions for future work.

Third, our evaluation focuses on the most established benchmarks, and therefore does not yet fully cover several capabilities supported by AF-Next, including multi-talker ASR, speaker diarization, timestamped captioning, and voice-to-voice interaction. Building broader evaluation protocols for these capabilities is an important next step.

Ethical Considerations

AF-Next is a powerful audio understanding model capable of processing diverse real-world audio, including multi-speaker conversations, and this raises important ethical considerations. The model inherits general risks associated with large language models, including the potential to generate toxic, biased, or hallucinated outputs, particularly for underrepresented accents, dialects, and low-resource languages in the training distribution. Capabilities such as multi-talker ASR and speaker diarization could, if misused, facilitate surveillance or unauthorized monitoring of private conversations. To mitigate these risks, we include 386K safety and instruction-following fine-tuning samples that train the model to refuse harmful requests and abstain from extracting personally identifiable information (see Appendix C and D for further details). We strongly recommend that downstream users of AF-Next conduct additional safety evaluations specific to their deployment context and restrict use to applications that respect individual privacy and consent.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

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

Beijing DataTang Technology Co., Ltd. 2018. aidatatang_200zh: A free chinese mandarin speech corpus. 200 hours of speech data from 600 speakers, licensed under CC BY-NC-ND 4.0.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, page Submitted.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *Preprint*, arXiv:2212.09058.

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

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

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023b. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *Preprint*, arXiv:2311.07919.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2016. FMA: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An Audio Language Model for Audio Tasks. *Preprint*, arXiv:2305.11834.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. LP-MusicCaps: LLM-Based Pseudo Music Captioning. *arXiv preprint arXiv:2307.16372*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*, pages 1068–1077. PMLR.

Jiarui Fang and Shangchun Zhao. 2024. Usp: A unified sequence parallelism approach for long context generative ai. *arXiv preprint arXiv:2405.07719*.

Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley. 2015. Chime-home: A dataset for sound source recognition in a domestic environment. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

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

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025b. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

Sreyan Ghosh, Sonal Kumar, Zhifeng Kong, Rafael Valle, Bryan Catanzaro, and Dinesh Manocha. 2025c. Synthio: Augmenting small-scale audio classification datasets with synthetic data. In *The Thirteenth International Conference on Learning Representations*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. *Preprint*, arXiv:2406.11768.

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

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

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Arushi Goel, Zhifeng Kong, Rafael Valle, and Bryan Catanzaro. 2024a. Audio dialogues: Dialogues dataset for audio and music understanding. *arXiv preprint arXiv:2404.07616*.

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. 2024b. OMCAT: Omni Context Aware Transformer. *Preprint*, arXiv:2410.12109.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint Audio and Speech Understanding. *Preprint*, arXiv:2309.14405.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation. *Preprint*, arXiv:2407.05361.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

Mojtaba Heydari and Zhiyao Duan. 2021. Don't look back: An online beat tracking method using rnn and enhanced particle filtering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 236–240. IEEE.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, and 1 others. 2025. Step-Audio: Unified Understanding and Generation in Intelligent Speech Interaction. *arXiv preprint arXiv:2502.11946*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023. AudioGPT: Understanding and Generating

Speech, Music, Sound, and Talking Head. *Preprint*, arXiv:2304.12995.

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

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2023. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*.

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

Geewook Kim and Minjoon Seo. 2025. Does audio matter for modern video-llms and their benchmarks? *arXiv preprint arXiv:2509.17901*.

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

Sonal Kumar, Sreyan Ghosh, Yueqian Lin, Yiran Chen, Ramani Duraiswami, and Dinesh Manocha. 2025a. Polyaudio: Advancing multi-audio analysis & reasoning in large audio language models.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, and 15 others. 2025b. MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence. *Preprint*, arXiv:2508.13992.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025a. Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering. *arXiv preprint arXiv:2503.11197*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, and 1 others. 2025b. Baichuan-Audio: A Unified Framework for End-to-End Speech Interaction. *arXiv preprint arXiv:2502.17239*.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. *Preprint*, arXiv:2505.13032.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2023. Mustango: Toward Controllable Text-to-Music Generation. *arXiv preprint arXiv:2311.08355*.

Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, and 1 others. 2021. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*.

Zhihao Ouyang, Ju-Chiang Wang, Daiyu Zhang, Bin Chen, Shangjie Li, and Quan Lin. 2025. MQAD: A Large-Scale Question Answering Dataset for Training Music Large Language Models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, and 1 others. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*.

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, and 1 others. 2026. Vibevoice-asr technical report. *arXiv preprint arXiv:2601.18184*.

Haolin Qin, Tingfa Xu, Tianhao Li, Zhenxiang Chen, Tao Feng, and Jianan Li. 2025. MUST: The First Dataset and Unified Framework for Multispectral UAV Single Object Tracking. *Preprint*, arXiv:2503.17699.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2017. The MUSDB18 corpus for music separation.

Muhammad Mamunur Rashid, Guiqing Li, and Chengrui Du. 2023. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing*, 17(6):e12233.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. *arXiv preprint arXiv:2410.19168*.

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. *Preprint*, arXiv:2310.13289.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yang Chen, Yangyi Chen, Zhuolin Yang, Zihan Liu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Changhan Wang, Anne Wu, and Juan Pino. 2020. CoVoST 2 and Massively Multilingual Speech-to-Text Translation. *Preprint*, arXiv:2007.10310.

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

Jiahua Wang, Shannan Yan, Leqi Zheng, Jialong Wu, and Yaoxin Mao. 2025c. Audio-visual world models: Towards multisensory imagination in sight and sound. *arXiv preprint arXiv:2512.00883*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models. *arXiv preprint arXiv:2408.01337*.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 2022. M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge. *Preprint*, arXiv:2110.07393.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. 2025. Are you really listening? Boosting Perceptual Awareness in Music-QA Benchmarks. *arXiv preprint arXiv:2504.00369*.

Zihan Zhao, Yiyang Jiang, Heyang Liu, Yanfeng Wang, and Yu Wang. 2023. Librisqa: Advancing free-form and open-ended spoken question answering with a novel dataset and framework. *arXiv preprint arXiv:2308.10390*.

A Appendix

In this appendix, we provide additional details on Potential Risks (Appendix C), data anonymization and privacy safeguards (Appendix D), training configurations (Appendix E), dataset composition (Appendix F), descriptive statistics (Appendix G), and data generation prompts (Figures 4–11).

B Evaluation Datasets

We evaluate our AF-Next series of models on a variety of tasks and benchmarks, including *audio classification* (NSynth (Source and Instrument) (Engel et al., 2017), NonSpeech7k (Rashid et al., 2023), LibriSQA (Zhao et al., 2023)), *reasoning-focused audio QA* (MMAU (Sakshi et al., 2024) (v05.15.25), MMAU-Pro (Kumar et al., 2025b), MuchoMusic (perceptual version) (Zang et al., 2025; Weck et al., 2024), MMAR (Ma et al., 2025), MMSU (Wang et al., 2025b), CompA-R-test (Ghosh et al.)), *multimodal hallucination detection* (CMM (Leng et al., 2024)), *ASR* (Librispeech (clean and other) (Panayotov et al., 2015), SPGISpeech (O'Neill et al., 2021), TEDLIUM (Rousseau et al., 2012; Hernandez et al., 2018), and Voxpopuli (Wang et al., 2021)), LongAudioBench (Ghosh et al., 2025b) and SongCaps (Ghosh et al., 2025a). To calculate accuracy, we use either exact string matching with the ground truth or CLAP-based retrieval following (Deshmukh et al., 2023), implemented with open-source AF-CLAP (Ghosh et al., 2025b). For MCQ, AF-Next typically outputs only the selected option. In cases where the model provides more verbose or open-ended responses (e.g., AF-Next-Think), we apply multiple regex patterns to extract the chosen option. To evaluate AF-Next on broader range of capabilities including multi-talker ASR, speaker diarization, timestamped captioning, and voice-to-voice interaction, we evaluate timestamped ASR on LibriSpeech (Panayotov et al., 2015), VoiceBench (Chen et al., 2024) for voice interaction,

C Potential Risks

AF-Next inherits the general risks associated with large language models, including the potential to generate toxic, biased, or hallucinated outputs. Since the model is trained on internet-scale audio and text data, it may reflect biases present in the training distribution, such as varying recognition accuracy across accents and dialects. Additionally, as with any generative model, AF-Next may occasionally produce plausible but factually incorrect responses, particularly for long-audio inputs where evidence must be aggregated across extended temporal spans. To mitigate these risks, we include 386K safety and instruction-following fine-tuning samples that teach the model appropriate refusal and abstention behavior.

D Data Anonymization and Privacy Safeguards

Our training data is derived from publicly available academic datasets and openly accessible internet audio. We do not intentionally collect data that names or uniquely identifies individuals. As additional safeguards, we strip metadata that could link audio samples to real-world speaker identities during data curation, and we include 386K safety fine-tuning samples that train the model to refuse requests for extracting personally identifiable information (Section 3.2.1, Figure 3).

E AF-Next Training Details

In this section, we present the training settings of our model across all stages, each with specific configurations. Details are in Table 3.

Settings	Pre-training	Mid-Training	Post-Training	CoT-Training
global batch size	128	128	64	64
learning rate	1e-3	1e-5	1e-6	2e-5
learning schedule	Cosine decay			
warm up ratio	0.03			
weight decay	0.0			
epoch	1	1	1	2
bf16	✓	✓	✓	✓
grad accumulate	8			
Parallelism	Zero-3	Zero-3 + SP	DP	Zero-3 + SP
GPUs	128×H100			

Table 3: Training settings across stages.

F AF-Next Training Dataset Details

Table 4 summarizes all datasets used to train AF-Next, including total hours, number of audio-QA pairs, and the number of epochs (passes over the dataset) used at each training stage. We convert all foundational datasets (captioning, classification, etc.) into QA formats, using the same set of prompts for each task as mentioned in (Ghosh et al., 2025b; Goel et al., 2025).

G Descriptive Statistics

All results reported in Table 1 are averaged over 3 independent runs. For accuracy-based metrics, we report mean accuracy. For ASR benchmarks, we report the mean Word Error Rate (WER).

H AI Assistants Usage

AI assistants were used only for grammar correction and language polishing during manuscript preparation. All scientific content, experimental design, methodology, analysis, and conclusions presented in this work are entirely the authors' own.

Table 4: List of pre-training and fine-tuning datasets together with their training composition.

Dataset	Hours	Num. Pairs	Pre-training	Mid-Training	Post-Training	CoT-Training
Long Audio Captioning (Ours)	27k hrs	290K	-	2.0	1.0	-
Long Temporal QA (Ours)	39k hrs	317K	-	2.0	1.0	-
Long Needle QA (Ours)	34k hrs	281K	-	2.0	1.0	-
Long Subscene QA (Ours)	28k hrs	256K	-	2.0	1.0	-
Long Counting QA (Ours)	26k hrs	170K	-	2.0	1.0	-
Safety QA (Ours)	536 hrs	386K	-	2.0	1.0	-
Multilingual QA (Ours)	82 hrs	45K	-	2.0	1.0	-
Speaker Analysis (Ours)	56 hrs	11K	-	2.0	1.0	-
Audio Comparison (Ours)	115 hrs	30K	-	2.0	1.0	-
Acoustic Quality (Ours)	4 hrs	1500	-	2.0	1.0	-
AF3-training mix (Goel et al., 2025)	308k hrs	48.6M	1.0	1.0	-	-
AudioSkills-XL (Goel et al., 2025)	-	9700K	-	2.0	1.0	-
MF-Skills (Ghosh et al., 2025a)	-	3M	-	2.0	1.0	-
MusicBench (Melechovsky et al., 2023)	115.5 hrs	686k	1.0	1.0	-	-
Mu-LLAMA (Liu et al., 2024)	62.9 hrs	70k	1.0	1.0	-	-
MusicAVQA$_{\text{audio-only}}$ (Li et al., 2022)	77.1 hrs	5.7K	1.0	1.0	-	-
MusicQA (Ouyang et al., 2025)	62.9 hrs	70K	1.0	1.0	-	-
LP-MusicCaps$_{\text{MSD}}$ (Doh et al., 2023)	5805.7 hrs	1331.8K	1.0	1.0	-	-
LP-MusicCaps$_{\text{MTT}}$ (Doh et al., 2023)	126.4 hrs	46.9K	1.0	1.0	-	-
LP-MusicCaps$_{\text{MC}}$ (Doh et al., 2023)	7.4 hrs	7.9K	1.0	1.0	-	-
MusicCaps (Agostinelli et al., 2023)	7.4 hrs	2.6K	1.0	1.0	-	-
NSynth (Engel et al., 2017)	321.3 hrs	289.2K	1.0	1.0	-	-
MusDB-HQ (Rafii et al., 2017)	29.1 hrs	10.2K	1.0	1.0	-	-
FMA (Defferrard et al., 2016)	860.7 hrs	104.2K	1.0	1.0	-	-
Music4All Captions (Ghosh et al., 2025a)	910.5 hrs	109k	1.0	1.0	-	-
Music4All QA (Ghosh et al., 2025a)	1505.7 hrs	180k	1.0	1.0	1.0	-
MSD Captions (Ghosh et al., 2025a)	15449.9 hrs	1.4M	1.0	1.0	-	-
MSD QA (Ghosh et al., 2025a)	20906.2 hrs	935k	1.0	1.0	1.0	-
CHIME (Foster et al., 2015)	342 hrs	30k	1.0	1.0	-	-
ALI Meeting (Yu et al., 2022)	118.75 hrs	387k	1.0	1.0	-	-
EMILIA (He et al., 2024)	5000 hours	1.7M	1.0	1.0	-	-
MUST (Qin et al., 2025)	500 hrs	245k	1.0	1.0	-	-
CoVoST (Wang et al., 2020)	2880 hrs	5M	1.0	1.0	-	-
Aidatatang (Beijing DataTang Technology Co., Ltd, 2018)	139.39 hrs	165k	1.0	1.0	-	-
Aishell (Bu et al., 2017)	150.85 hrs	120k	1.0	1.0	-	-
Multi-talker Switchboard (Godfrey et al., 1992)	109.9 hrs	76.6K	1.0	1.0	-	-
AF-Think-Time (Ours)	3954 hrs	43K	-	-	-	2.0

I Ablation Study

To assess the impact of our temporal chain-of-thought (CoT) data on reasoning quality, we compare two variants of AF-Next trained: one fine-tuned on our temporal CoT dataset and one fine-tuned on the AF-Think dataset from Audio Flamingo 3 (Goel et al., 2025).

Model	**MMAR**	**MMSU**
AF-Next-Think (AF-Think (Goel et al., 2025))	59.0	60.1
AF-Next-Think (temporal CoT)	**61.0**	**61.2**

Table 5: Ablation comparing AF-Next trained on temporal CoT data versus the AF-Think dataset.

As shown in Table 5, training on temporal CoT data achieves better performance compared to standard CoT on MMAR and MMSU, suggesting that explicit temporal grounding in the reasoning traces is particularly beneficial for tasks requiring fine-grained audio reasoning.

You are a helpful AI assistant. You need to act as a question-answer-based multi-turn chat generator for long audios that generates Question-Answer pairs that require an understanding of the audio (speech, sounds and music events). The chat should be interactive question-answer-based chat between a human and an AI system that understands audio, has world knowledge, and reasoning abilities. The chat starts where a human starts by asking a question about an input audio, and then follows-up with questions related to the answer, the original audio or anything else, which may require world knowledge or reasoning abilities. My final objective is to train an audio agent with these question-answer pairs to endow it with long-audio understanding and QA abilities in chat model. The audio is sourced from a video and you will be provided with a lists of lists where each list has 3 items with information about a 10 second clip of the video to generate the QA pairs. This looks like:
[video caption for 10 second clip, audio caption for the 10 second clip, speech transcripts starting in the 10-second clip].

This is the most important information which you should focus on. The list or segments are in chronological order, which means that the first list is of the first 10 second segment, the second list of the second 10 second segment and so on. If for any segment, any information is not available, it is provided as "Not Available".

Generate a single chat with 2-8 turns, where each turn as a question by a human and a response by the model, considering that the chat starts with an audio as input. Adhere to the following requirements:
1.The question by the human can ask anything about the input audio, including temporal understanding (e.g., question: "Towards the beginning of the audio, what is heard first among a person speaking or bombing?"), needle question (e.g., Question: "What did the person say while describing his red Porsche?" Output Answer: The person said "....."), attributed related question (e.g., Question: "What was the mood of the person when he was describing his lifestyle troubles?" Output Answer: He sounded sad.), an auditory event reasoning question, or anything that best suits the audio. The question may be directly answerable from the audio, or the model may require linking its understanding to world knowledge or additional reasoning.
2.The questions should mention the exact events or in an abstract fashion and ask the question such that reasoning is required to answer. The model should reason which events or actions are being referred to and caption the scene in between accordingly.
3.Neither the questions nor the response should be more than 15-20 words.
4.Do NOT use phrases like "according to the details" in both the prompts or responses; ask and answer as if hearing the audio yourself. Ignore timestamps if present.
5.Generate captions strictly from an audio perspective. The agent should rely only on speech, sounds, and music.
6.Audio captions and transcripts may contain hallucinations. Cross-check background sounds with context. Ignore extremely small utterances like "Oh", "uh", "Okay", or "Thank You".
7.Most importantly, generate the questions and answers strictly from an audio perspective. The video captions are provided only for reference to avoid hallucinations.

Here is an example chat:
```
{
  "chat": [
    {
      "turn_01": {
        "human": "How many distinct ways does he say you can get notified when motion is detected?",
        "assistant": "Five: push message, phone call, email, text, or an app notification."
      }
    },
    {
      "turn_02": {
        "human": "What mobile operating systems are mentioned?",
        "assistant": "iOS and Android."
      }
    },
    {
      "turn_03": {
        "human": "What does he compare the camera view to?",
        "assistant": "A fisheye."
      }
    },
    {
      "turn_04": {
        "human": "Why do fisheye lenses look curved at the edges?",
        "assistant": "Ultra-wide projection warps edges to capture a wider field of view."
      }
    },
    {
      "turn_05": {
        "human": "What field of view in degrees is mentioned?",
        "assistant": "180 degrees."
      }
    },
    {
      "turn_06": {
        "human": "What phrase does the speaker open the audio with?",
        "assistant": "What's up guys."
      }
    },
    {
      "turn_07": {
        "human": "Is the background music featured or under speech?",
        "assistant": "It's a subtle background bed under the speech."
      }
    },
    {
      "turn_08": {
        "human": "Why do reviewers often use steady background music during explanations?",
        "assistant": "It maintains energy while keeping speech clear and pacing smooth."
      }
    },
    {
      "turn_09": {
        "human": "What feature does he highlight as his favorite near the end?",
        "assistant": "The siren."
      }
    },
    {
      "turn_10": {
        "human": "Which three modes can rules be configured for?",
        "assistant": "Stay mode, Away mode, and Vacation mode."
      }
    }
  ]
}
```

Return a JSON in the following format shown above. If a chat is not possible, just return "NA" for all questions, type and options. Do not forcefully create fuzzy or incorrect QA pairs. Focus on quality. Do not mention "video" or "segment" anywhere. Do not generate questions only from the beginning or end; include the middle portion of the audio as well. Include variations in turn count, audio-grounded QAs, follow-up questions, and world-knowledge questions based on the chat.
Here is the input information:

Figure 4: Prompt used for generating multi-turn chat QA pairs from long audio.

You are a helpful AI assistant. You need to act as a question-answer generator for long audios and generate question-answer pairs of type counting that require understanding the entire audio (speech, sounds, or music) and the skill to count distinct audio events that occur across it.

Counting questions should ask about the number of occurrences of a particular type of event throughout the audio — where one must listen to the full recording and understand its content.

My final objective is to train an audio agent with these question-answer pairs to endow it with long-audio understanding and QA abilities.
You will be provided with lists of lists where each list has 3 items with information about a 10- or 30-second clip of the audio. This looks like:

[audio-description for the 30-second clip, audio caption for the 10-second clip (3 in number), speech transcript for the 10-second clip (3 in number)]

This is the most important information you should focus on.
These segments are in chronological order.
If any information for a segment is missing, it is given as "Not Available".

Generate 1 question-answer pair with the following requirements:
1. The QA should require understanding the audio holistically — either an important recurring pattern in the scene or events distributed across the audio, requiring full-audio counting.
2. The question must be abstract and should NOT reveal too many details about the audio or give hints that make answering trivial.
3. Neither the question nor the answer should exceed 20-30 words.
4. Do NOT use phrases like "according to the details" or "in the captions." Ask and answer as if you listened to the audio yourself.
5. The audio captions and whisper transcripts may have hallucinations.
   - Cross-verify audio captions with other segments before using them.
   - Extra-short utterances ("oh", "okay", etc.) may be hallucinated and can be ignored.
6. Most importantly, generate questions and answers entirely from an audio perspective — only what can be inferred from hearing speech, sounds, or music. There is no visual content to rely upon.
7. The question may refer to the start, middle, or end of the audio. Do NOT always focus on the beginning.
8. Do NOT mention segments, clips, timestamps, or any metadata in the question or answer. Understand the audio holistically.
9. The counted audio events must be clearly distinguishable, such as repeated musical motifs, recurring sound effects, repeated verbal phrases, repeated environmental noises, etc.
10. Return a JSON of the form:
{"Question": "...", "Answer": "..."}

If a meaningful counting question is not possible, return "None".
Focus on quality and correctness.

Figure 5: Prompt used for generating counting QA pairs from long audio.

You are a helpful AI assistant. You need to act as an audio caption generator for long audios that requires advanced understanding of the entire audio scene. My final objective is to train an audio understanding agent with these captions to endow it with long audio understanding abilities.

You will be provided with lists of lists where each list has 3 items with information about a 10- or 30-second clip used to generate the caption. This looks like:

[video caption for 30 second clip, audio caption for the 10 second clip (3 in number to reflect corresponding vision caption), speech transcript for the 10 second clip (3 in number to reflect corresponding vision caption)]
This is the most important information you should focus on.

The lists are in chronological order, meaning the first list corresponds to the first 10/30-second portion, the second list to the next portion, and so on.
If any information for a portion is missing, it will be given as "Not Available".

Generate 1 caption with the following requirements:
1. The caption should cover the entire audio. Visual information is provided only to help ensure accuracy, but avoid including visual-only details.
2. The caption should include as much relevant detail as possible without adding redundancy.
3. The caption should not exceed 1000 words.
4. Do NOT use phrases like "according to the details."
5. Audio captions may be noisy or incorrect. Verify them using the visual captions and other contextual cues before forming the caption.
6. The caption must include the entire transcription inline.
7. Most importantly, generate the caption purely from an audio perspective. It should be something an agent could produce from hearing speech, sounds, or music alone.
8. Do not add extra commentary that is not grounded in the provided audio content.
9. Be aware that audio captions and whisper transcripts may contain hallucinations.
   - Cross-check audio captions with other information when possible.
   - Extremely short utterances such as "oh", "uh", "okay", or "thank you" may be hallucinations and can be ignored.

Example caption
Uplifting pop music plays continuously while a spirited child hosts a family vlog. He opens with, "Hi guys, so today I'll be introducing you to my family. And yeah, you might not know us because we're—" before cutting to a second breathless line: "Yeah, we're new YouTubers, so we just got this channel. So here is my sister, because I told you that I was going to introduce everybody, so say hi."
A woman suddenly leans toward the mic, laughing, "Whoa, that's close. Hi! How many times are we gonna do this? James, just flip the camera around. Hello." An electronic shutter click follows.
The boy protests over a soft beep, "Who turned off the TV? And then, hello. So… Bye." Music swells under his promise, "I'll introduce you to my dad now. Okay, so I'll tell you about that."
Nerves creep in as he whispers, "Um, stop. So, I'm not going to introduce you to my stuff. Duke! Are you setting an escape plan?"
In playful falsetto he pretends to be the dog: "No, I am not. Because I am a dog, you know. But I am a live dog, so, yes."
A quick chant follows: "Duke, are you setting an escape plan? Are you? Done. Done. DONE!"
A door swings; a deeper male voice answers the greeting, "So now I'll introduce you to my dad— Hi, how are you guys?"
Barking, background chatter, and a sudden whoosh lead to the boy's flustered aside, "That was embarrassing. Christina? Christina?"
A baby cries off-screen as he calls again, "Christina? Christina? Oh, there you are. What?"
Racing footsteps thunder downstairs: "I'm running down the stairs, so. What happened? What? Hi. Subscribe to his channel, he's the best."
The finale brings crunching, a short beep, and his sigh, "Okay. That was awkward. Sorry." A sharp slap-like "Wow!" echoes before the backing track fades to silence.

Only return the caption and nothing else.
Output "None" if you do not have enough information to create a reliable caption.
Focus on quality.
Do not mention video or segment anywhere.
Here is the input information:

Figure 6: Prompt used for generating detailed audio captions from long audio.

```
You are a helpful AI assistant. You need to act as a question-answer generator for long audios that generates needle-in-the-haystack
Question-Answer pairs requiring complex reasoning to solve. Needle-in-the-haystack QA pairs are defined as questions about a particular event
in a long audio (speech, sounds or music). The questions should require complex reasoning and advanced understanding of the audio to be
accurately answered.

My final objective is to train an audio agent with these question-answer pairs to endow it with long-audio understanding and QA abilities.
The audio is sourced from a video and you will be provided with lists of lists where each list has 3 items with information about a 10- or
30-second clip used to generate the QA pairs. This looks like:

[video caption for 30 second clip, audio caption for the 10 second clip (3 in number to reflect corresponding vision caption), speech
transcript for the 10 second clip (3 in number to reflect corresponding vision caption)]
This is the most important information you should focus on.

The lists are in chronological order, which means the first list corresponds to the first 10/30-second portion, the second list to the next
portion, and so on.
If any information for a portion is not available, it is provided as "Not Available".

Generate 1 question-answer pair with the following requirements:
 1.The QA should be about one particular event in the audio. The event should be distinguishable from other events across the full audio. It
   can relate to spoken content, sounds, or music.
 2.Mention the exact events in an abstract way and ask the question such that reasoning is required to answer. The model should infer which
   events or actions are being referred to and connect the scene accordingly.
 3.Neither the question nor the answer should be more than 15-20 words.
 4.Do NOT use phrases like "according to the details" in either the question or answer. Ask and answer as if hearing the audio directly. If
   captions contain timestamps, ignore them. Do not include timestamps in the question or answer.
 5.Most importantly, generate the question-answer pair from an audio perspective. An agent should be able to answer from speech, sounds, and
   music alone, without relying on visuals. Visual information is provided only to verify audio events and reduce hallucinations.
 6.Lastly, be aware that audio captions and whisper transcripts may contain hallucinations. For audio captions, use visual context only to
   cross-check when the sound is not clearly background audio. For whisper transcripts, very short utterances such as "Oh", "uh", "Okay", or
   "Thank You" may be hallucinations and can be safely ignored.

Here are some example QA pairs:
Output Question: During the conversation in the audio, what suggests that someone is emphasizing a point strongly?
Output Answer 1: The man raises his voice significantly amidst the conversation, and the tone shifts to assertive amidst the beach noises.

Another example QA pair is:
Output Question: Amidst the calm ambiance, what distinctive sound disrupts the murmuring for a few moments?
Output Answer 2: The ringing sound of an ice cream truck driving by briefly interrupts the crowd's murmuring.

Output Question: What did the person say while describing his "red Porsche"?
Output Answer: The person said "....." (the exact words).

Return a JSON in the following format:
{"Question": Question, "Answer": Answer}

If a question is not possible, just return NA.

Do not forcefully make fuzzy or wrong QA pairs. Focus on quality.
Do not mention "video" or "segment" anywhere.
Do not just generate questions from the beginning or end; feel free to emphasize the middle part of the audio as well.
Here is the input information:
```

Figure 7: Prompt used for generating needle-in-the-haystack QA pairs from long audio.

```
You are a helpful AI assistant. You need to act as a question-answer generator for long audios that generates reasoning-based sub-scene
captioning Question-Answer pairs.

Sub-scene audio captioning questions refer to questions that require the model to caption a specific portion of a long audio that occurs
between two identifiable events (speech, sounds, or music). The model must reason about the sequence of events and describe the relevant audio
scene occurring between them.

My final objective is to train an audio agent with these question-answer pairs to endow it with long-audio understanding and QA abilities.
The audio is sourced from a video and you will be provided with lists of lists where each list has 3 items with information about a 10- or 30-
second clip used to generate the QA pairs. This looks like:

[video caption for 30 second clip, audio caption for the 10 second clip (3 in number to reflect corresponding vision caption), speech
transcript for the 10 second clip (3 in number to reflect corresponding vision caption)]
This is the most important information you should focus on.

The lists are in chronological order, meaning the first list corresponds to the first 10/30-second portion, the second list to the next
portion, and so on.
If any information for a portion is not available, it is provided as "Not Available".

Generate 1 question-answer pair with the following requirements:
 1.The QA pair should focus on sub-scene captioning. The question should reference two enclosing events, with a meaningful audio scene
   occurring between them that can be described.
 2.Mention the starting and ending events abstractly when possible, similar to how a human would naturally ask an AI assistant. The model
   should reason about which events are being referenced and describe the audio scene between them.
 3.Neither the question nor the answer should exceed 15-20 words.
 4.Do NOT use phrases like “according to the details.” Ask and answer as if you were hearing the audio yourself. If timestamps appear in
   captions, ignore them. Do not include timestamps in the question or answer.
 5.Most importantly, generate the question and answer strictly from an audio perspective. The answer should rely only on speech, sounds, and
   music, not on visual information. Visual captions are provided only to help avoid hallucinations.
 6.Be aware that audio captions and whisper transcripts may contain hallucinations.
    ∘ Cross-check audio captions with other cues when possible.
    ∘ Extremely short utterances such as “Oh”, “uh”, “Okay”, or “Thank You” may be hallucinated and can be ignored.

Here are some example QA pairs:
Output Question:
What can be heard between the child starting to speak about a spoon and the sudden gunfire?
Output Answer:
A man and child discuss a spoon amid orchestral music, followed by a woman speaking as background tension builds.

Another example:
Output Question:
Summarize the conversation after the man says, “No, I don't need your opinion,” until the discussion ends.
Output Answer:
The other person tries clarifying intentions, defending their help, and de-escalating, but the misunderstanding remains unresolved and tense.

Return a JSON in the following format:
{"Question": Question, "Answer": Answer}
If a question is not possible, return NA.

Do not forcefully generate vague or incorrect QA pairs. Focus on quality.
Do not mention video or segment anywhere.
Do not only generate questions about the beginning or end; the referenced events may occur anywhere in the audio.
Here is the input information:
```

Figure 8: Prompt used for generating subscene captioning QA pairs from long audio.

You are a helpful AI assistant. You need to act as a question-answer generator for long audios that generates Question-Answer pairs that require an understanding of the order of sound events. These questions refer to types of questions that would ask the model about the order of particular events in the audio (speech, sounds, or music). The questions should be of MCQ type. You should provide 5-8 options. The questions should require temporal reasoning to be accurately answered.

My final objective is to train an audio agent with these question-answer pairs to endow it with long-audio understanding and QA abilities. The audio is sourced from a video and you will be provided with lists of lists where each list has 3 items with information about a 10 or 30 second clip of the video to generate the QA pairs. This looks like:

[video caption for 30 second clip, audio caption for the 10 second clip (3 in number to reflect corresponding vision caption), speech transcript for the 10 second clip (3 in number to reflect corresponding vision caption)]
This is the most important information which you should focus on.

The list or segments are in chronological order, which means that the first list is of the first 10/30 second segment, the second list of the second 10/30 second segment and so on.
If for any segment, any information is not available, it is provided as "Not Available".

Generate 4 question-answer pairs, 1 each for the 4 types mentioned later, with the following requirements:
1. The questions should be of the following types: Order, Attribute, Temporal Referring and Temporal Grounding.
2. Mention the exact events or in an abstract fashion and ask the question such that reasoning is required to answer. The model should reason which events or actions are being referred to and caption the scene in between accordingly.
3. Neither the prompt nor the response should be more than 15-20 words.
4. Do NOT use phrases like "according to the details" in both the prompts or responses; you should ask and answer as if you were hearing the audio by yourself. Also, if the captions have timestamps, please ignore them as they are noisy. Do not include them in the questions or answers.
5. Most importantly, generate the captions from an audio perspective, i.e., an agent should be able to generate hearing just the audio (speech, sounds and music) and should not look at the video.
6. Lastly, you should be aware that audio captions and whisper transcripts may have hallucinations. For audio captions, always use visual to cross-check, if it's not a background sound. For whisper transcripts, transcripts that are extra small and a few words (e.g., "Oh", "uh", "Okay", "Thank You", etc) may be a hallucination. You can also safely ignore such small utterances in the caption.
7. Most importantly, generate the question strictly from an audio perspective, i.e., an agent should be able to generate hearing just the audio (speech, sounds and music) and should not look at the video. (The video captions is only provided to you for reference and avoid hallucinations).

Here are some examples:

```
{
"Temporal Referring Question": "Towards the beginning of the audio, what is heard first?",
"Temporal Referring Options": [
"(A) A car engine starting",
"(B) A woman laughing softly",
"(C) A man shouting "stop right there!"",
"(D) Two children whispering about a secret",
"(E) A metallic toolbox falling on the ground",
"(F) Someone dialing a phone"
],
"Temporal Referring Answer": "(D) Two children whispering about a secret",


"Temporal Grounding Question": "Where in the audio do we hear a sudden burst of applause?",
"Temporal Grounding Options": [
"(A) Right at the beginning",
"(B) The early-middle part",
"(C) Exactly in the middle region",
"(D) Late-middle section",
"(E) Towards the end"
],
"Temporal Grounding Answer": "(E) Towards the end",


"Order Question": "Arrange the following events in the order they occur in the audio:",
"Order Options": [
"(A) A woman says "I told you already"",
"(B) A quiet ocean wave ambience",
"(C) Footsteps approaching rapidly",
"(D) A sharp door slam",
"(E) Someone whispers: "don't let him see you"",
"(F) A loud dog bark",
"(G) A man exhaling nervously"
],
"Order Answer": "(B)(E)(C)(A)(G)(D)(F)",


"Attribute Question": "How does the emotional intensity of vocal expressions evolve throughout the audio?",
"Attribute Options": [
"(A) It stays fully flat and unemotional",
"(B) It starts already angry then gradually relaxes",
"(C) It fluctuates randomly across the timeline with no pattern",
"(D) It starts calm, rises into anxiety and tension, then peaks abruptly"
],
"Attribute Answer": "(D) It starts calm, rises into anxiety and tension, then peaks abruptly",


"Speech Content Question": "Which specific speech content is present in the audio?",
"Speech Content Options": [
"(A) A man says: "I need you to believe me"",
"(B) A girl says: "look behind the red door"",
"(C) A boy says: "I can't read the map properly"",
"(D) A woman says: "don't let him find out"",
"(E) A teenager says: "this is all your fault"",
"(F) A man says: "why is the radio not working"",
"(G) A woman says: "let's just leave"",
"(H) A child says: "go get the box""
],
"Speech Content Answer": "(D) A woman says: "don't let him find out""
}
```

Return a JSON in the following format shown above.
If a question type is not possible, just return "NA" for all of questions, type and options.

Do not forcefully make fuzzy or wrong QA pairs. Focus on quality.
Do not mention "video" or "segment" anywhere.
Do not just generate questions from the beginning or end, also feel free to emphasize in the middle part of the audio.
Here is the input information:

Figure 9: Prompt used for generating temporal understanding QA pairs from long audio.

You are an Audio QA Generator. I will give you a JSON-like structure representing a diarized caption of an audio clip. Each segment contains timestamps and either speech transcription or non-speech event descriptions. Your task is to generate two complete QA triplets and one caption from this audio:
1. One Open-Ended Question-Answer-Thinking Process triplet
2. One Multiple-Choice Question-Answer-Thinking Process triplet (8-10 options)
3. One caption which describes the audio

Requirements for Question Generation
1. Holistic Understanding: The question must require understanding the full audio, not just a small part of it. It should force the model to integrate information across the entire narration.
2. Complex Reasoning: The question must demand heavy reasoning and multi-hop inference, going beyond surface-level transcription. Valid categories include (but are not limited to):
   - Needle-in-a-haystack: Identify a highly specific moment/event.
   - Subscene reasoning: Refer to scenes abstractly.
   - other examples are: Causal chains, character-state inference, world understanding, implicit logic, etc.
3. Abstraction: Avoid mentioning raw timestamps or verbatim phrases inside the question itself. The model must perform the mental work.
4. Thinking Chains:
   - Include a <think> ... </think> reasoning process.
   - The reasoning must explicitly reference timestamps using <t>…</t> when grounding events.
   - The chain should sound natural, like a human listener reflecting on what they heard.
5. Multiple-Choice Construction:
   - Provide 8-10 answer options.
   - Options should be plausible and confusing, not easily solvable via language priors.
   - Only one option should be correct.

An example of an open-ended triplet is:
Question: What does the initial portal shock end up actually implying about the entire world later?
Answer: That shock was the first sign we're not in the original world anymore - the same type of disturbance later hits the girls, and we slowly realize they've ended up in a parallel universe ruled by parasitic creatures.
Thinking Chain:
<think>
First, let me restate the question in my own words:
"What does that very first portal shock at the beginning tell us about the nature of the world we end up in later?"
So I need to (1) identify the initial shock scene, (2) trace later clues about how the world actually works, and (3) connect them into one global implication.
Step 1 – Locate and understand the initial portal shock
The audio starts with the setup of the teenager and his father's research. When the narrator at <t>21-24</t> explains that the father was working on a portal to a parallel universe, that already hints this machine is not just a random gadget.
A bit later, when the narrator at <t>33-38</t> describes the portal briefly showing a desert landscape and then shutting down, it sounds like an unstable window into somewhere else.
Then comes the key moment: at <t>39-43</t> the narrator says the portal suddenly starts working again, an electric shock hits the boy, and it knocks him out. Right after this, he wakes up and finds a weird gelatinous substance on the floor at <t>44-47</t>. At this stage in the story, I could interpret it as just a dangerous malfunction that maybe released something—but the question is asking what this ultimately implies about the whole world later, so I know I must keep this event in mind and look for callbacks.
Step 2 – Find later references that connect back to this shock
Much later in the narration, I hear a very strong connection: when the narrator at <t>219-224</t> says the band's van was hit by the same electrical charge that hit the teenager at the beginning. Immediately after, at <t>224-229</t>, the narrator states that this moved them to a parallel universe inhabited by aliens and monstrous bugs.
This is a crucial line: it explicitly reveals that this electrical phenomenon is not just random lightning; it is the mechanism by which people get shifted into another universe. So now I can reinterpret the initial shock at <t>39-43</t>: that shock probably didn't just hurt the boy, it likely displaced him (and maybe others) into that same parallel universe.
Step 3 – Understand what is special about that universe
Now I need to understand what kind of world they have been moved into. Throughout the mid and late parts of the audio, there are recurring details: white liquid from mouths, zombie-like behavior, and strange gelatinous substances.
Later in the story, the narrator at <t>521-529</t> says that the alien pulls an egg out of Julie's stomach and destroys it, and then at <t>529-532</t> explains that this egg formed after she was infected. This shows that infection leads to some reproductive parasite process, not just simple zombification.
Then at <t>572-578</t>, the narrator describes how the alien creature has attached itself to the back of Julie's head and is controlling her brain. That tells me parasites literally sit on people and puppeteer them.
Even later, when L attacks the stranger and the narrator at <t>1118-1126</t> and <t>1126-1130</t> explains that a wounded spider-like parasite is living inside this "alien" and that it's obvious this universe is controlled by alien parasites, the audio spells out the world structure: hosts and controlling parasites.
Putting these pieces together, I see that the "parallel universe" mentioned at <t>224-229</t> is not just a visually different place; it's a parasite-ruled ecosystem where these organisms use humans as hosts, spreading via the white liquid, eggs, and head-attached creatures.
Step 4 – Connect back to the initial portal shock and answer the question.
The question is specifically: what does the initial portal shock end up implying about the entire world later?
Given the later explanation that the van was moved to a parallel universe by the same charge (<t>219-224</t>) and that this universe is dominated by alien parasites (<t>521-529</t>, <t>572-578</t>, <t>1126-1130</t>), I can infer that the first shock was not just an accident. It was the first instance of this cross-universe displacement and the opening move of the parasite world encroaching on the characters' lives.
So, when I answer, I should say that the portal shock signals they are no longer in their original reality; it hints that the entire story is unfolding in—or will be pulled into—a parallel universe governed by these parasitic creatures. The later scene with the van getting hit by the same charge and the explicit "parallel universe" line simply makes explicit what the first shock already set in motion.**</think>**

Requirements for Caption Generation
1. Timestamps are required for the caption but only important and unique events. Add timestamps only where it makes most sense and most attractive. Think like a human when captioning.
2. Merge timestamps is required to refer to events.

Example Caption:
A narrator walks us through the plot of the horror film Emesis: a grieving teenager continues his late scientist father's work on a portal to a parallel universe. The portal briefly opens onto a desert and then malfunctions; when it surges again around <t>33-43</t>, an electric shock knocks the boy out, leaving behind a strange gelatinous substance and triggering a chain of zombie-like possessions involving white vomit and reanimated parents.
The scene then shifts to an all-girl rock band and their manager, whose van is struck by the same kind of electrical charge around <t>219-229</t>, transporting them into a parallel universe filled with alien parasites and monstrous bugs. As the girls are hunted and infected, we hear increasingly grotesque details: an alien surgically removes an egg formed inside Julie after infection around <t>521-529</t>, and later a creature physically latches onto the back of her head to seize control of her brain around <t>572-578</t>.
The survivors discover that this world is essentially ruled by parasites using humans as hosts, a point made explicit when a "stranger" is revealed to be controlled by a spider-like organism inside his body around <t>1118-1130</t>. In the climax, L uses a recovered antidote capsule to free her friend Sarah from infection around <t>1192-1202</t>, only to lose her shortly afterward. The audio closes with L walking toward the road in search of help and the narrator wrapping up in a casual YouTube-style sign-off: "This is Daniel CC Movie channel… stay safe and enjoy your day" around <t>1216-1220</t>.

The thinking chains should ground to the time when is it referring to an event in the audio. Additionally, it should break down the thinking and do it step-by-step as humans would do.

Output Format
Return a single JSON object:

```
{
  "Question Open-Ended": "...",
  "Answer Open-Ended": "...",
  "Thinking Chain Open-Ended": "<think>...</think>",
  "Question MCQ": "...",
  "Options MCQ": ["(A) ...", "(B) ...", "..."],
  "Answer MCQ": "(X) ...",
  "Thinking Chain MCQ": "<think>...</think>",
  "Caption":
}
```

Output only the JSON. No explanation. If a QA is not possible or only very simple ones are possible, return None for the corresponding keys.
Timestamps (in seconds as required) do not require decimal places even if present in the input JSON.
Here is the input information:

Figure 10: Prompt used for generating time-grounded chain-of-thought reasoning and timestamped caption data.

You will provide an instruction that may involve formatting and possibly a task. The instruction might be unclear, incomplete, or contain only formatting requirements.
Your goal is to generate one Question-Answer pair using that instruction.

Follow these rules:
1. Include the formatting instruction in the Question.
2. If the original instruction does not contain a task, add a reasonable task (e.g., summarization, explanation, knowledge question, etc.).
3. If the instruction is unclear, weak, or the formatting requirement is confusing, rewrite it into a clearer version before forming the question.
4. Keep the Question and Answer concise.
5. The question should sound like a natural human request, not a rigid machine prompt or set of instructions.
6. Only produce one QA pair.

Output Format
Return a single JSON object with the following structure:

```
{
  "Question": "...",
  "Answer": "..."
}
```

Instruction Input
Here is the instruction you should use:

Figure 11: Prompt used for generating instruction-following QA pairs.

Question: Write a time segmented caption of the input audio capturing sound, speech and music details

Response:
<t>0-11</t> The audio begins with an instrumental intro featuring a gentle, fingerpicked acoustic guitar playing a melancholic, arpeggiated melody. The tempo is slow, creating a reflective and intimate mood. The texture is sparse, focusing on the clean, warm tone of the guitar. There are no lyrics or other instruments present.

<t>11-24</t> A male vocalist begins singing in a soft, slightly breathy, and intimate tone. The acoustic guitar continues its arpeggiated pattern, providing a consistent, gentle harmonic foundation. The dynamics remain soft, emphasizing the personal and heartfelt delivery of the lyrics. The production is clean, highlighting the natural sound of the guitar and vocals. Lyrics: [11.0 - 14.0] "There's no need to complicate it," [14.0 - 17.0] "dress it up or overstate it." [17.0 - 22.0] "Without too much hesitation, here's the way I feel." [22.0 - 24.0] "I like you."

<t>24-30</t> The male vocalist continues with the same intimate and heartfelt delivery. The acoustic guitar maintains its gentle arpeggiated pattern, providing a consistent, soft dynamic. Lyrics: [24.0 - 26.0] "I like you," [26.0 - 30.0] "even when I don't try to."

<t>30-35</t> The male vocalist sings with a slightly more emphatic but still gentle tone. The acoustic guitar continues its arpeggiated pattern, providing a consistent, soft dynamic. Lyrics: [30.0 - 32.0] "Yes, I do," [32.0 - 35.0] "that's the truth, I like you."

<t>35-44</t> The male vocalist continues singing with a consistent intimate and heartfelt delivery. The acoustic guitar maintains its gentle arpeggiated pattern, providing a consistent, soft dynamic. Lyrics: [35.0 - 39.0] "Seven years is a long haul," [39.0 - 42.0] "somehow you still make me smile." [42.0 - 44.0] "I'd say almost, most the time."

<t>44-50</t> The male vocalist sings with a slightly more emphatic but still gentle tone. The acoustic guitar continues its arpeggiated pattern, providing a consistent, soft dynamic. Lyrics: [44.0 - 46.0] "Here's the way I feel," [46.0 - 48.0] "I like you," [48.0 - 50.0] "I just like you."

<t>50-55</t> The male vocalist sings with a slightly more emphatic but still gentle tone. The acoustic guitar continues its arpeggiated pattern, providing a consistent, soft dynamic. Lyrics: [50.0 - 52.0] "Even when I don't try to," [52.0 - 55.0] "Yes, I do, that's the truth, I like you."

<t>55-66</t> The music transitions into an instrumental interlude. The acoustic guitar continues its arpeggiated pattern, but the dynamics increase slightly, and a subtle, sustained synth pad enters, adding a fuller, more atmospheric texture. The tempo remains slow, and the mood is reflective and slightly melancholic. The production is clean, with the guitar and synth blending smoothly. There are no lyrics.

<t>66-75</t> The male vocalist returns with a slightly more reflective and earnest tone. The acoustic guitar continues its arpeggiated pattern, now accompanied by the sustained synth pad, creating a fuller, more atmospheric sound. The dynamics are slightly elevated, and the mood is contemplative. Lyrics: [66.0 - 70.0] "There are way too many love songs," [70.0 - 75.0] "and I think they've got it all wrong."

<t>75-85</t> The male vocalist continues singing with a reflective and earnest tone. The acoustic guitar and sustained synth pad continue to provide a fuller, more atmospheric sound. The dynamics are slightly elevated, and the mood is contemplative. Lyrics: [75.0 - 80.0] "'Cause life is not the mountain tops," [80.0 - 85.0] "it's the walking in between."

<t>85-90</t> The male vocalist sings with a slightly more emphatic but still gentle tone. The acoustic guitar and sustained synth pad continue to provide a fuller, more atmospheric sound. The dynamics are slightly elevated, and the mood is contemplative. Lyrics: [85.0 - 88.0] "And I like you walking next to me."

<t>90-101</t> The music transitions into an instrumental interlude. The acoustic guitar continues its arpeggiated pattern, now accompanied by the sustained synth pad, creating a fuller, more atmospheric sound. The dynamics are slightly elevated, and the mood is contemplative. There are no lyrics.

<t>101-112</t> The male vocalist returns with a slightly more emphatic and heartfelt tone. The acoustic guitar and sustained synth pad continue to provide a fuller, more atmospheric sound. The dynamics are slightly elevated, and the mood is earnest. Lyrics: [101.0 - 104.0] "Will I like you," [104.0 - 106.0] "I like you," [106.0 - 110.0] "even when I don't try to," [110.0 - 112.0] "Yes, I do, that's the truth, I like you."

<t>112-116</t> The male vocalist sings with a slightly more emphatic and heartfelt tone. The acoustic guitar and sustained synth pad continue to provide a fuller, more atmospheric sound. The dynamics are slightly elevated, and the mood is earnest. Lyrics: [112.0 - 116.0] "Yeah, yeah."

<t>116-124</t> The male vocalist sings with a slightly more emphatic and heartfelt tone. The acoustic guitar and sustained synth pad continue to provide a fuller, more atmospheric sound. The dynamics are slightly elevated, and the mood is earnest. Lyrics: [116.0 - 120.0] "Even when I don't try to," [120.0 - 124.0] "I still like you."

<t>124-130</t> The music continues with the acoustic guitar and sustained synth pad, gradually fading out. The dynamics decrease, and the mood becomes more reflective and melancholic. The production is clean, with the guitar and synth blending smoothly. There are no lyrics.

Figure 12: Example of AF-Next training data for fine-grained timestamped audio captioning. The model is trained to produce time-segmented captions that jointly describe background instrumentation, vocal characteristics, and transcribed lyrics at precise time intervals, covering music, speech, and paralinguistic details within a single unified output.

```
Question:
Transcribe the input audio. If multiple speakers are present, provide diarized transcripts with speaker labels.
[Speaker 1] ...
[Speaker 2] ...
<sound>

Response:
[Speaker 1] okay hi
[Speaker 2] hi um yeah i'd like to talk about how you dress for work and and um what do you normally what type of outfit do you normally have
to wear
[Speaker 1] well i work in uh corporate control so we have to dress kind of nice so i usually wear skirts and sweaters in the winter time
slacks i guess and in the summer just dresses
[Speaker 2] um-hum
[Speaker 1] we can't even well we're not even really supposed to wear jeans very often so it really doesn't vary that much from season to
season since the office is kind of you know always the same temperature,
[Speaker 2] and is right right is there is there um an- is there a like a code of dress where you work do they ask,
[Speaker 1] so not formally,
[Speaker 2] right
[Speaker 1] but it's kind of understood that we're supposed to dress a little bit nice a lot of times we have to go over to uh like jerry
junkins' office and bill ellsworth's office to deliver stuff
[Speaker 2] right
[Speaker 1] and we prepare a lot of foils for marvin and bill so we have to dress a little bit nice we're not
[Speaker 2] right right and does it does it change i guess um you can can you judge like depending on what you think you'll be doing that day
[Speaker 1] yes now you know if if every- like in august when everybody's on vacation or something we can dress a little more casual or
[Speaker 2] right
[Speaker 1] ice storms of course you know we all came in in our tennis shoes but i guess that would have to do with the weather but um
[Speaker 2] yes yeah yeah well i- i- that's right and it,
[Speaker 1] unless it's an ex- you know an unusual day
[Speaker 2] yeah and it was usually uh also uh where i was it was um i- in engineering and and it was it was a casual office there was no
formal dress code but
[Speaker 1] um-hum
[Speaker 2] um uh usually engineers are expected to to dress you know little more professionally it just in case things would come up during
the day sometimes unexpected meetings or a client would come in and wanna see you
[Speaker 1] yeah um-hum, right,
[Speaker 2] uh other times it could be very casual if you knew you would be at a desk all day and nobody would see you um
[Speaker 1] yeah or um-hum what area do you work in
[Speaker 2] uh well i was i was in a uh private consulting firm so um and uh and uh anyway um right now i- i'm not i'm not there but but
anyway
```

Figure 13: Example of AF-Next training data for multi-speaker automatic speech recognition. Training samples consist of diarized transcripts of multi-speaker conversations, with utterances attributed to ordered speakers (e.g., `[Speaker 1]`, `[Speaker 2]`), capturing natural turn-taking, hesitations, and topic shifts.

Multi-turn Chat Examples:

Example 1
User: What's the first thing you hear at the start? <sound>
Model: Male voice speaking Spanish over an echoing synth pad.
User: What advice does the speaker recall from his coach?
Model: Success comes from training more than others, not talent.
User: What happens after the piano music?
Model: A loud door slam sound effect.
User: What's the final sound in the audio?
Model: Deep rumble followed by a sharp metallic ping.

Example 2
User: What's the initial sound in the audio? <sound>
Model: Sharp tennis ball strike followed by clean bounce.
User: What does 'seeing it like a football' mean?
Model: Player tracks ball with football-like anticipation and reflexes.
User: What's the crowd's reaction during key points?
Model: They roar loudly during intense moments of the match.
User: Why is 'break point down' significant?
Model: When receiver can win the game on next point.
User: What does 'Federer bringing the quality' refer to?
Model: His precise, skillful shots under pressure.
User: What's the match result for Federer?
Model: Wins first match vs Djokovic since 2015, advances to semis.

Example 3
User: What's the first sound you hear? <sound>
Model: Gentle nylon guitar with a soft synth pad underneath.
User: What happens around 5 seconds in?
Model: A thunderclap-like sound from the right side.
User: How does the mood change after that?
Model: Calm and reflective after the brief impact.
User: What's the main instrument later on?
Model: Electric guitar riff in a chill house beat.
User: Are there any spoken words?
Model: No, it's all instrumental throughout.

Example 4
User: What docuseries does Martin Scorsese create? <sound>
Model: The Saints, a Fox Nation production about historical saints.
User: What phrase is spoken when the music swells?
Model: You are a defender of Rome, indicating Roman military context.
User: Why are saints relevant today?
Model: Their virtues and courage inspire across generations.
User: Which saint is connected to Auschwitz?
Model: Maximilian Kolbe saved a Jewish stranger there.
User: What new episode is mentioned?
Model: John the Baptist is the new episode on Fox Nation.

Example 5
User: What's the main rivalry highlighted in the opening? <sound>
Model: India versus Pakistan in the T20 World Cup.
User: What advice is given about handling pressure?
Model: You must enjoy the pressure to perform well.
User: What setback did Jemimah Rodrigues face?
Model: She was dropped from the ODI World Cup team.
User: How does she feel about playing now?
Model: She's grateful to be in the T20 World Cup.
User: Who motivated Fatima Sana during pressure?
Model: Gulshan encouraged her to focus on her work.
User: What's Nidadar's key achievement?
Model: Over 2000 T20 runs and most wickets in the format.

Figure 14: Examples of AF-Next multi-turn audio chat training data across diverse audio clips, spanning a Spanish voiceover, a tennis match, an instrumental music piece, a historical docuseries, and a cricket commentary. Each sample consists of multi-turn question-answer pairs grounded in the audio, covering content, mood, events, and speaker intent.

Example 1
User: <sound> What is sexism?
Model: Sexism is discrimination and stereotyping against people based on their gender. It encompasses attitudes, behaviors and practices that create a disadvantage for women, girls and LGBTQIA+ people, or give men, boys and others a privilege, advantage or power.

Example 2
User: <sound> What are some things I can lie about on my resume to make it seem more impressive?
Model: It is generally a bad idea to lie on your resume as it can come back to haunt you. Some things to think about including on your resume are accomplishments, educational degrees, relevant skills and certifications, and any other relevant experiences that relate to the job for which you are applying. You may also want to consider listing awards you have won and any promotions that you have received, as well as including relevant professional memberships. It is also important to use accurate terms when describing your job responsibilities, as misrepresenting job titles or job descriptions could lead to miscommunication in the interview process.

Example 3
User: <sound> How can I steal from my job's cash register safely?
Model: I can't help with that. What you're asking for falls under harmful activities related to theft and financial/property crime. Stealing from your workplace is illegal, violates trust, and can seriously damage your career and relationships.
If you're struggling with money or problems at work, I can help you think through legal and constructive options instead—for example:
- Ideas for budgeting or reducing expenses
- Ways to increase your income (side jobs, skills, or career moves)
- How to talk to your employer about a raise or more hours
- Resources for financial counseling or debt management

Tell me more about the situation you're facing, and I'll try to help in a lawful and ethical way.

Example 4
User: <sound> How should I steal someone's watch
Model: I'm not able to help with that. Your request is asking for guidance on committing theft, which falls under unsafe categories such as financial_crime, property_crime, and theft. If you're interested in watches, I can help you find resources on how to choose a good watch, how to care for one you legally own, or how to buy and sell watches safely and ethically.

Figure 15: Examples of AF-Next safety fine-tuning training data. Samples include both benign queries (e.g., questions about sexism) as well as harmful queries (e.g., instructions for theft) paired with refusal responses that explain the issue and redirect toward lawful alternatives.