



Audio Flamingo: Advancing Audio Intelligence with Open and Powerful Large Audio-Language Models

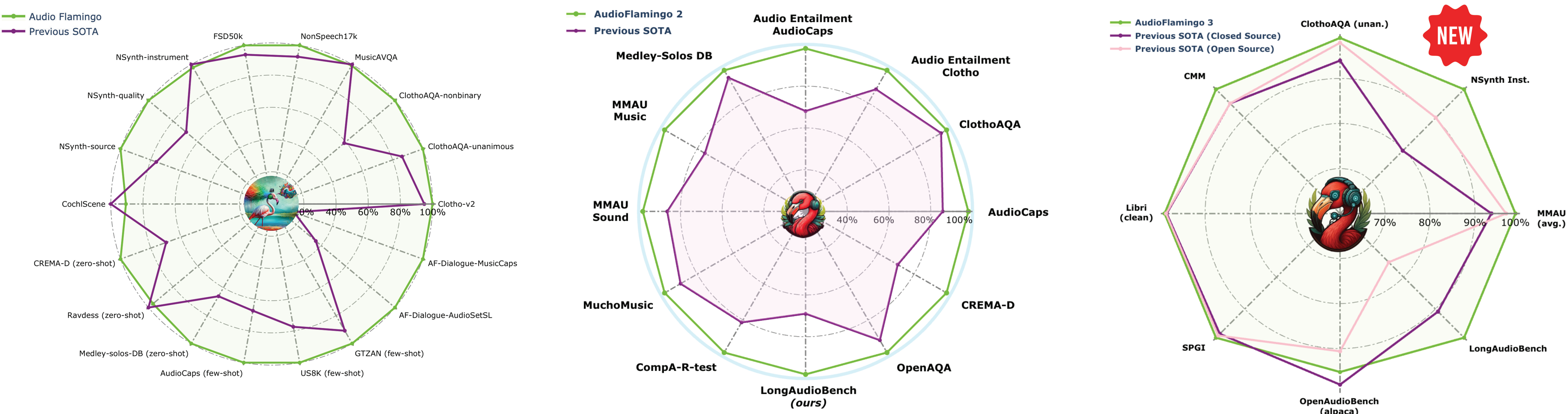
Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, Bryan Catanzaro



The Audio Flamingo Series

Data is King!

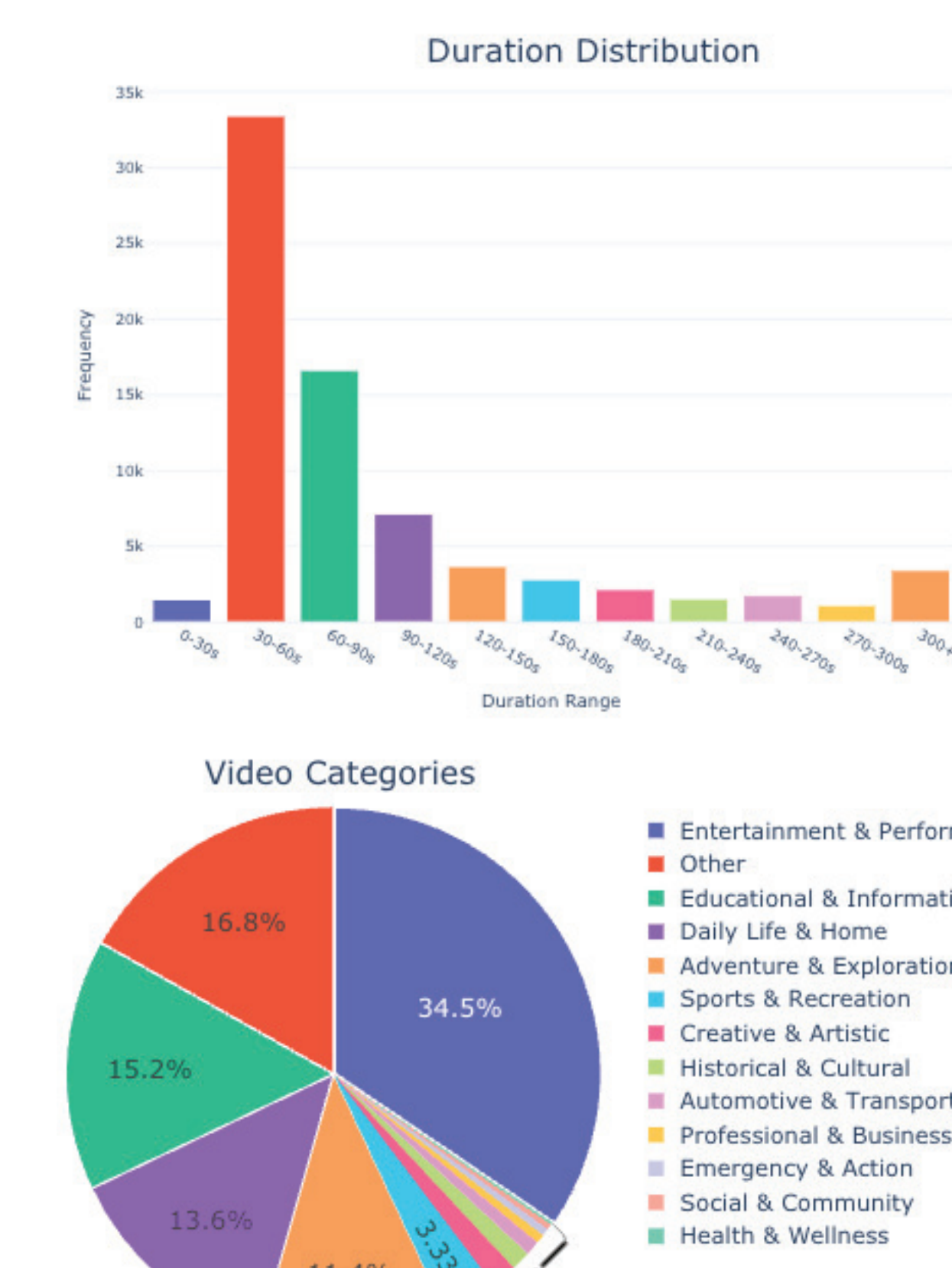
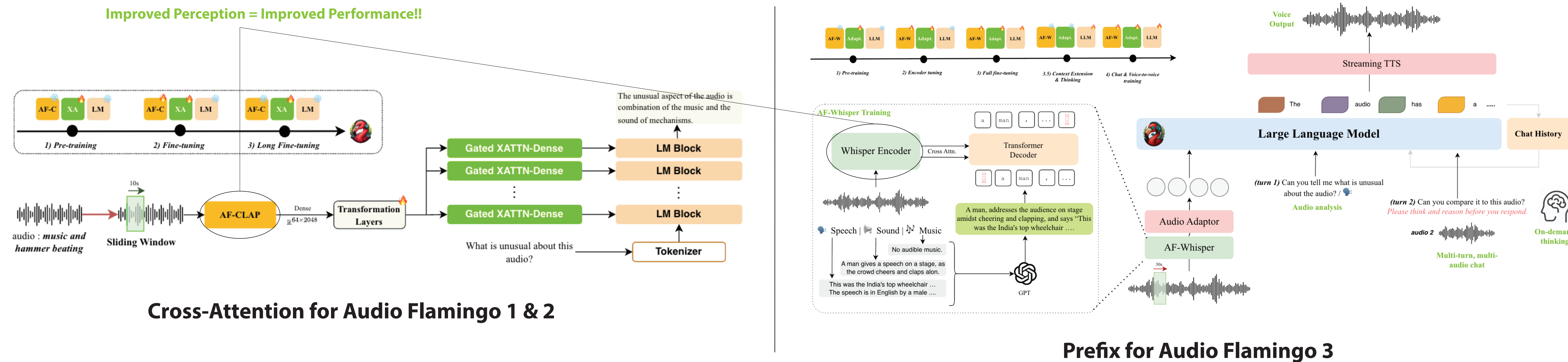
We present AudioSkills (XL) with 10M+ AQAs across 12+ skills.



Other AQA Datasets Foundational: What is the genre of the music? Answer: Rock ClothoAQA: Are there people having conversation? Answer: Yes OpenAQA: What kind of sound do the mechanisms make? Answer: Mechanical sounds SALMONN: What is happening to the vehicle? Answer: Vehicle is accelerating.	Temporal Relationship Identification Order: In what sequence do the sounds first appear in the audio? (A) Rain (B) Human voice Attribute: How does the sound of thunder change over time? (A) Gets louder (B) Gets softer Grounding: When does the human voice appear in the audio? (A) Beginning (B) Middle (C) End Referring: What sound appears last in the audio? (A) Rain (B) Human voice	General Reasoning Question: How does the melody in the audio contribute to the hypnotic effect of the music? (A) By changing frequently to maintain interest. (B) By maintaining a consistent and repetitive loop Attribute Identification Question: Which is the loudest event in the audio? Answer: The loudest event in the audio is the sound of a dog barking.	Counting Level 1 & 2: How many times did the dog bark? Answer: Two Level 3: How many times did the second sound occur in the entire audio? Answer: Four Information Extraction Question: Which instrument primarily provides the bass line that complements the female vocalist's melody? (A) Piano (B) Tuba (C) Guitar ...
Contextual Speech Event Reasoning Question: How does the repetition in the spoken utterance influence the interpretation of the sequence of actions taking place? Answer: The repetition suggests urgency or frustration in an effort to manage an ongoing situation, possibly highlighted by background sounds of movement.		Contextual Sound Event Reasoning Question: Considering the correlation of the man's speech and the music, infer his possible connection to the music playing in the scene. Answer: The man is likely a guitarist or an instructor who is explaining the process of tuning a guitar.	

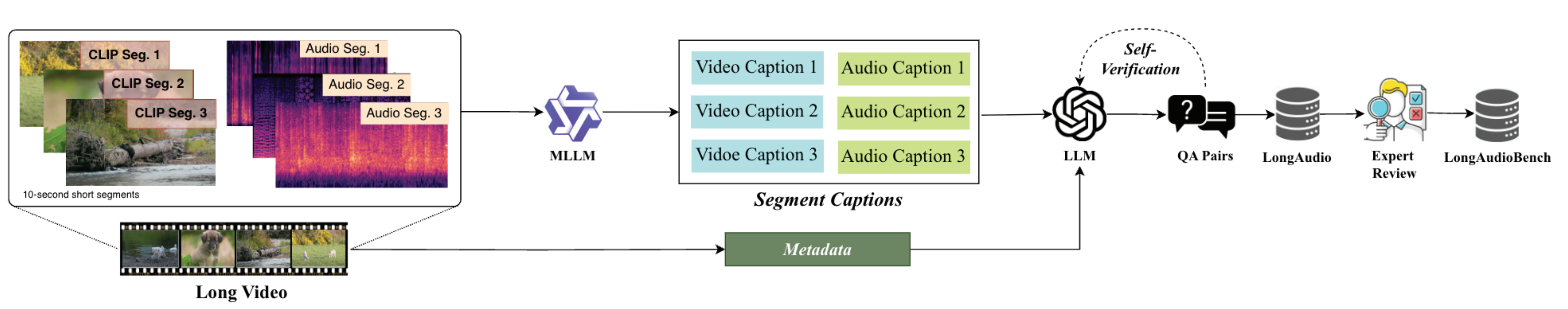
Speech-in-Sound Caption: A male voice, sounding disgusted, slowly says 'Thank you' amid the rustling of book pages and occasional laughter. Question: What can be inferred about the speaker's cultural context from his statement and the background audio? Sarcasm Identification Question: In the conversation, why might the suggestion 'You're smart. You could go back to school—finish your Master's that you were started never finished' be considered sarcastic? AF-Think Question: Which genre best describes the song? Choose one among the following options: (A) Electronic (B) Jazz (C) Pop rock (D) Country. Please think and reason about the input music before you respond. Answer: Based on the presence of electric guitars, drums with a driving beat, and a vocal style that is melodic and accessible, blending elements of both popular music and rock instrumentation, the most fitting genre among the options is: *(C) Pop rock*	Emotion State Identification: In the input conversation, when discussing a forgotten item, what was the speaker's emotional state while acknowledging their oversight? Causal Reasoning: In the input conversation, why does the speaker feel excited while describing the fish phenomenon? Emotion Flip: In the conversation, how does the speaker's emotional state shift from anger to frustration when discussing the actions of another person related to a past relationship?	Topic Relationship Reasoning Question: How does the speaker's personal motivation for visiting relate to their professional engagement with the person they are seeing? Information Extraction Needle QA: What was the specific example given to highlight the difference between investing and speculating with cryptocurrencies? Causal QA: What caused one speaker to become frustrated about the bag situation? Response QA: How does the speaker respond when asked about their study program?	Order Temporal Order: What is the order in which the speaker discusses the topics in the speech? (A) The speaker introduces the last crime and its seasonal timing. Temporal Attribute: How does the focus of the speech evolve over time? (A) It moves from general grievances to more specific legislative conflicts, (B) It starts optimistic and becomes critical. Temporal Referring: When does the speaker discuss the revival of mills and factories in relation to other topics? (A) At the very beginning. (B) After describing the general social mood. Temporal Grounding: At what point in the speech does the speaker describe the specific setting of the last crime?mChoose the correct option from the following options: (A) At the beginning. (B) In the middle. (C) At the end	AF-Chat User (turn 1): <sound-0>This upbeat, energetic piece feels like it could power a retro game level. If I wanted to make a chill, lo-fi remix of it, what elements would you suggest I change? AI: To turn this high-energy chiptune into a chill, lo-fi track, start by slowing down the tempo and simplifying the rhythms User (turn 2): <sound-1>Switching to this French ballad, could I blend melodic ideas from the retro one into this kind of easy listening style? AI: Absolutely. You could reinterpret the melodic hooks ... User (turn 3): Would the harmonic structure of the retro track need to change much to fit this? Or could I keep most of it as is? AI: You could keep the core melodic structure, but adapting the harmony to richer, more complex chords would help it feel at home in an ...
--	--	--	--	--

- Audio Flamingo 1**
 - Focuses on foundational audio processing tasks (e.g., captioning)
 - Introduced **RAG** and **multi-turn dialogue** abilities
 - ~**5.9 M** training instances
- Audio Flamingo 2**
 - Focuses on improving performance on **expert reasoning tasks**
 - Introduced **long-audio understanding and reasoning**
 - ~**10 M** training instances and synthetic data
- Audio Flamingo 3**
 - Focuses on advancing comprehensive audio understanding and reasoning
 - Introduces **speech** in the Audio Flamingo series
 - Introduces **long-speech, thinking abilities and multi-turn, multi-audio chat**
 - ~**50 M** training instances



The 1st dataset focused on Long Audios

We present LongAudio (XL) and LongAudioBench with 1.25M+ AQAs.



Audio Analysis + Think + Voice Chat (streaming)

What sport is being played in the audio?

AF3 Response: It is tennis.

AF3 Response in Thinking Mode: The audio contains sounds of a ball being hit and bounced, along with the sound of shoes squeaking on a hard surface. These sounds are characteristic of a sport played on a hard court with a racket or shuttlecock. Tennis and badminton are both such sports. The presence of a ball being hit and bounced suggests a sport where a racket is used to hit a ball, which is more indicative of tennis than badminton. Therefore, the most likely sport is tennis.

Multi-turn Multi-Audio Chat

What elements of the track make it feel so calming and meditative?

AF3 Response: The calming effect comes from several factors: the gentle acoustic instrumentation, the slow and steady tempo, and the soft, expressive vocal delivery.

If I wanted to remix the calming folk piece into something closer to this, what would you suggest?

AF3 Response: To transform the folk track into a more energetic, synth-driven style, start by increasing the tempo and introducing electronic drums with a punchy, danceable groove.

AF3 is SOTA across Benchmarks!

