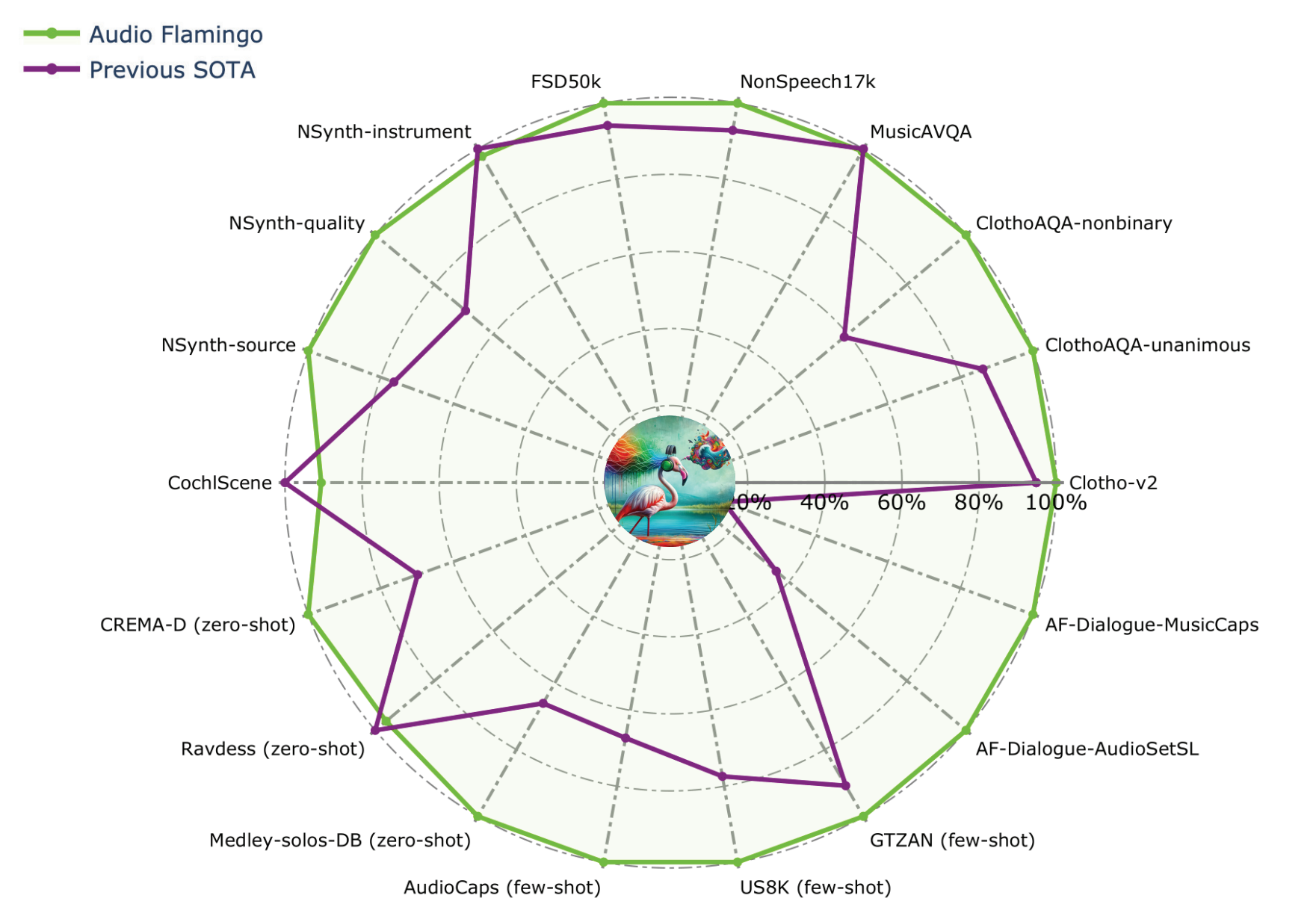
NEURAL INFORMATION PROCESSING SYSTEMS
Spotlight

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

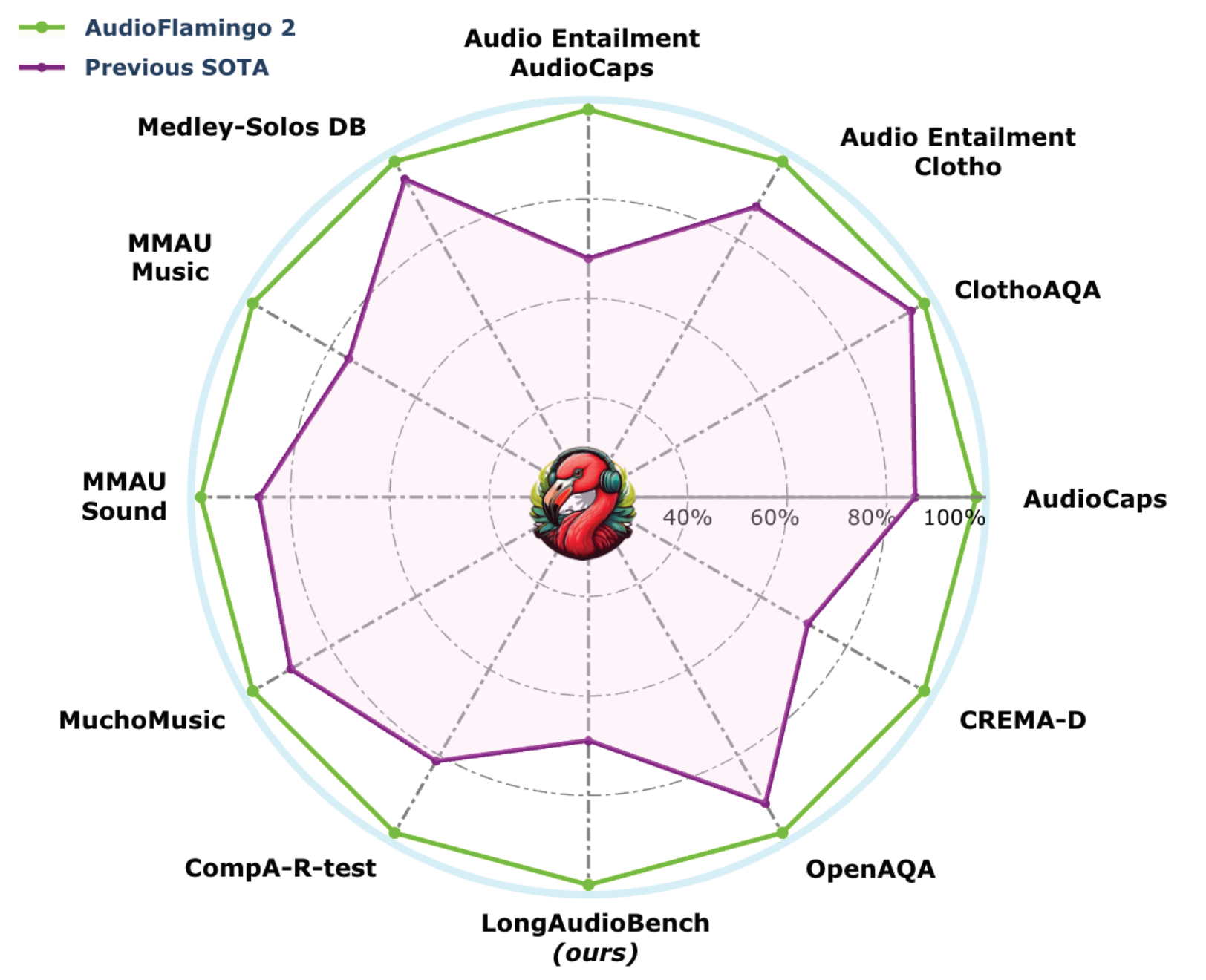
Sreyan Ghosh*, Arushi Goel*, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, Bryan Catanzaro

The Audio Flamingo Series



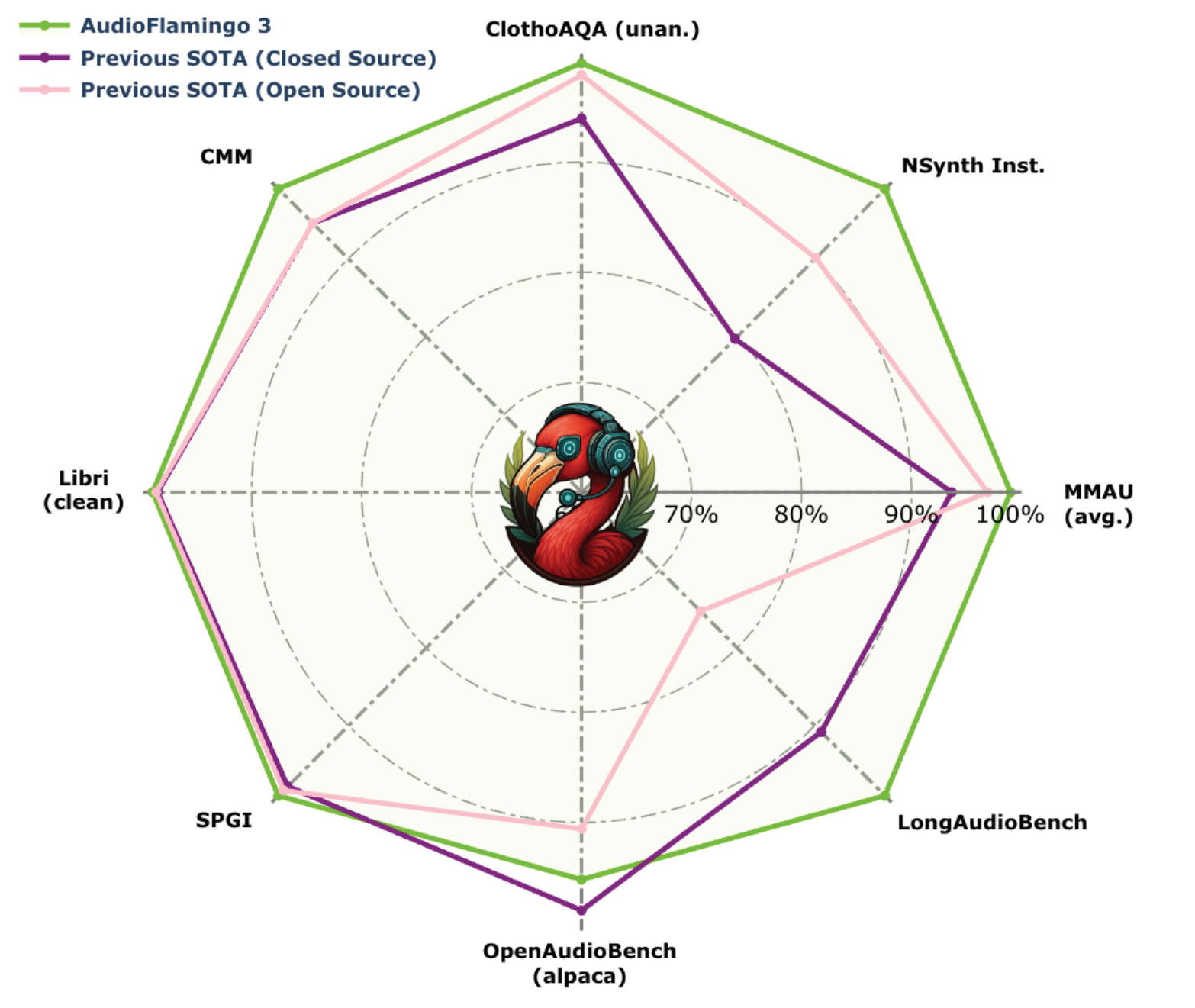
Audio Flamingo 1

- 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



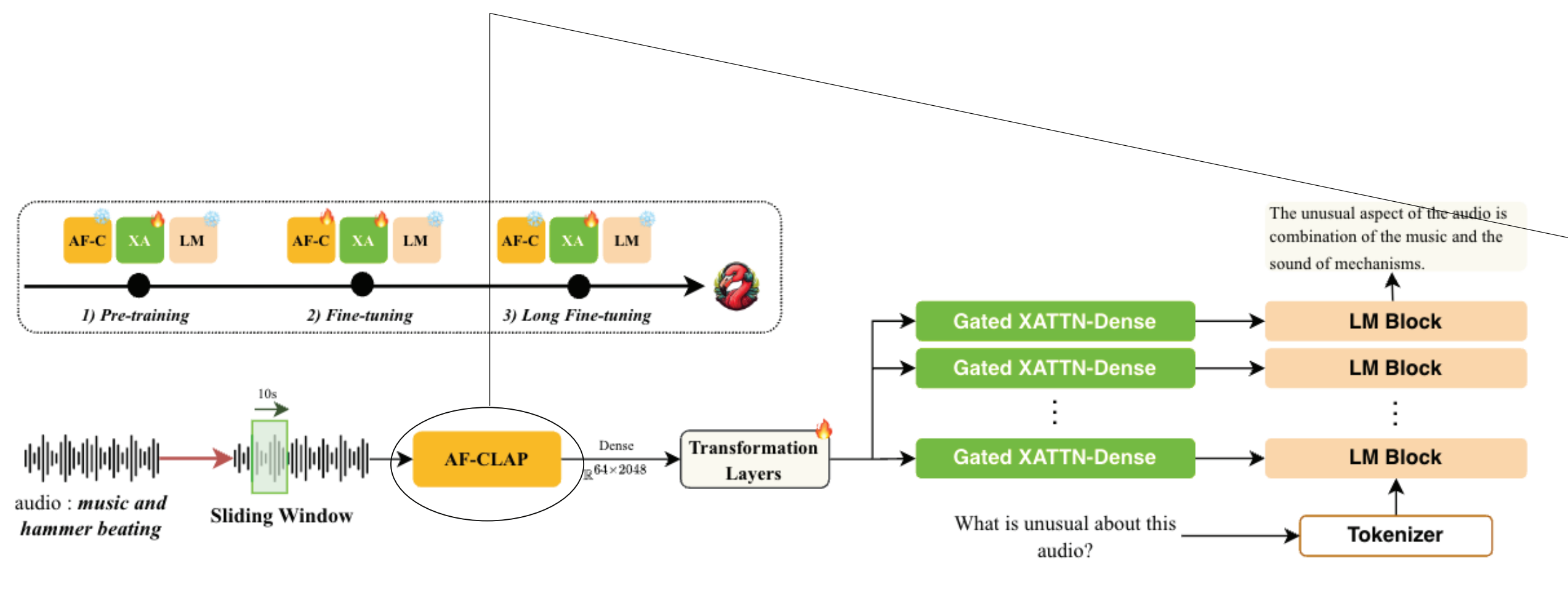
Audio Flamingo 3

- 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

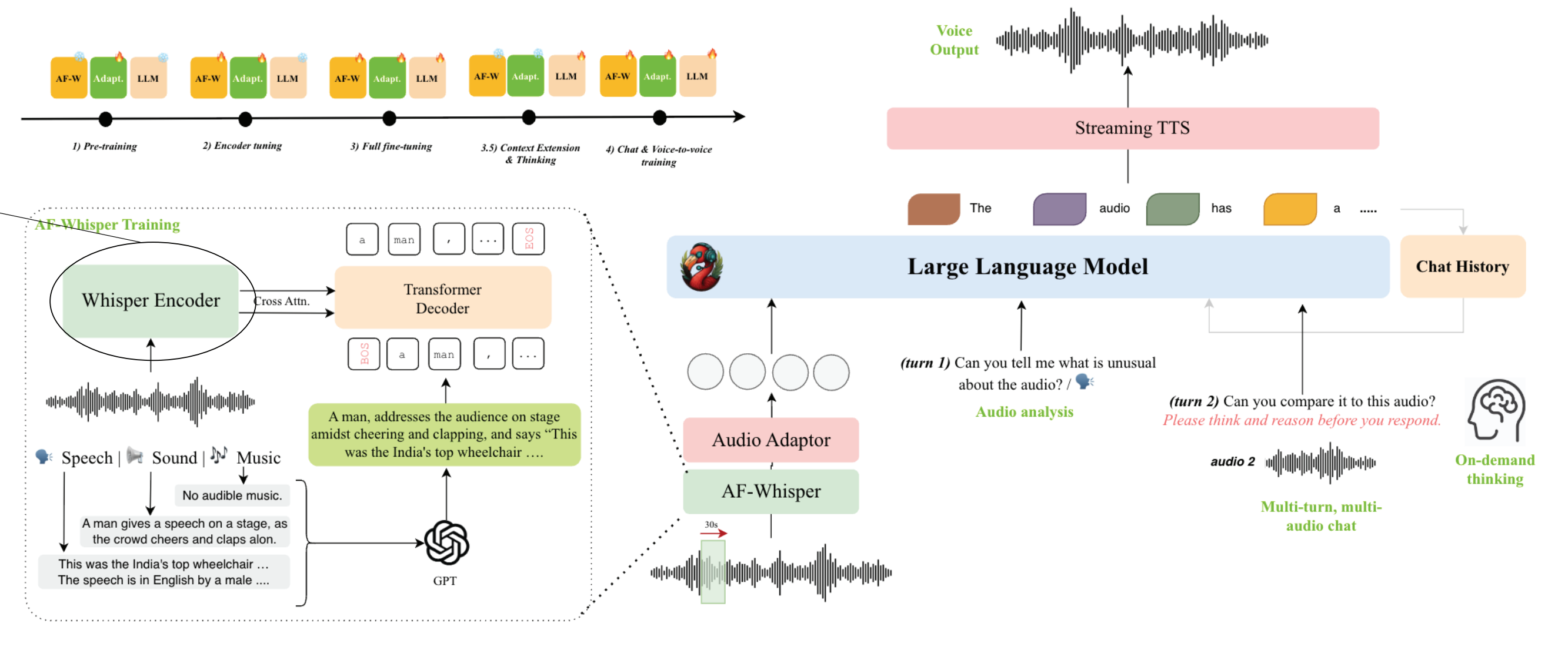
Music & Audio-Visual Flamingo

- Focused on deep understanding of full-length songs.
- With Omni Flamingo, we bring audio-intelligence to visual understanding models

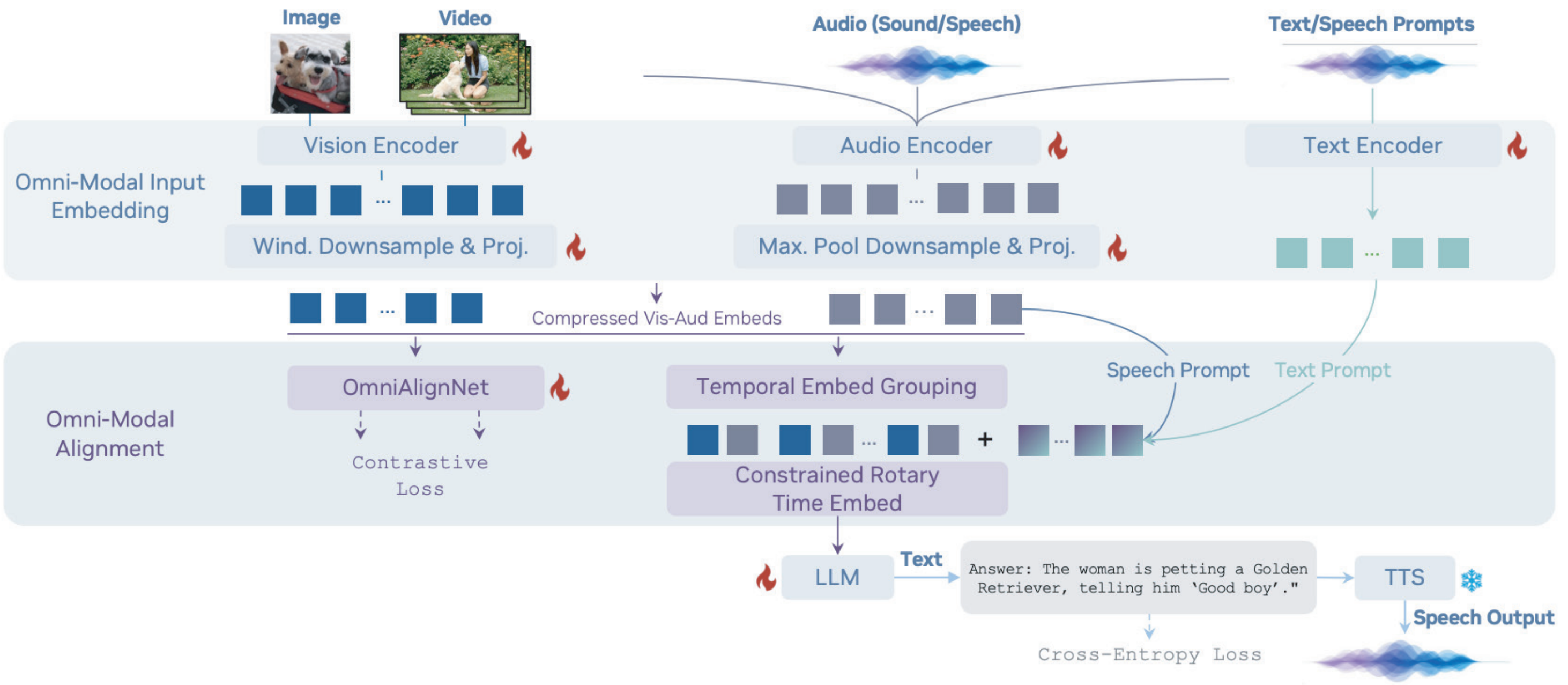
Improved Perception = Improved Performance!!


AF-C
XA
LM
1) Pre-training
2) Fine-tuning
3) Long Fine-tuning
audio : music and hammer beating
10s
Sliding Window
AF-CLAP
Dense
ℝ64×2048
Transformation Layers
Gated XATTN-Dense
LM Block
The unusual aspect of the audio is combination of the music and the sound of mechanisms.
What is unusual about this audio?
Tokenizer

Cross-Attention for Audio Flamingo 1 & 2


AF-W
Adapt.
LLM
1) Pre-training
2) Encoder tuning
3) Full fine-tuning
3.5) Context Extension & Thinking
4) Chat & Voice-to-voice training
AF-Whisper Training
Whisper Encoder
Cross Attn.
Transformer Decoder
a
man
,
...
EOS
BOS
Speech | Sound | Music
No audible music.
A man gives a speech on a stage, as the crowd cheers and claps alon.
This was the India's top wheelchair ...
The speech is in English by a male ....
A man, addresses the audience on stage amidst cheering and clapping, and says "This was the India's top wheelchair ....
GPT
Voice Output
Streaming TTS
The
audio
has
a
.....
Large Language Model
Chat History
Audio Adaptor
AF-Whisper
30s
(turn 1) Can you tell me what is unusual about the audio? /
Audio analysis
(turn 2) Can you compare it to this audio?
Please think and reason before you respond.
audio 2
Multi-turn, multi-audio chat
On-demand thinking

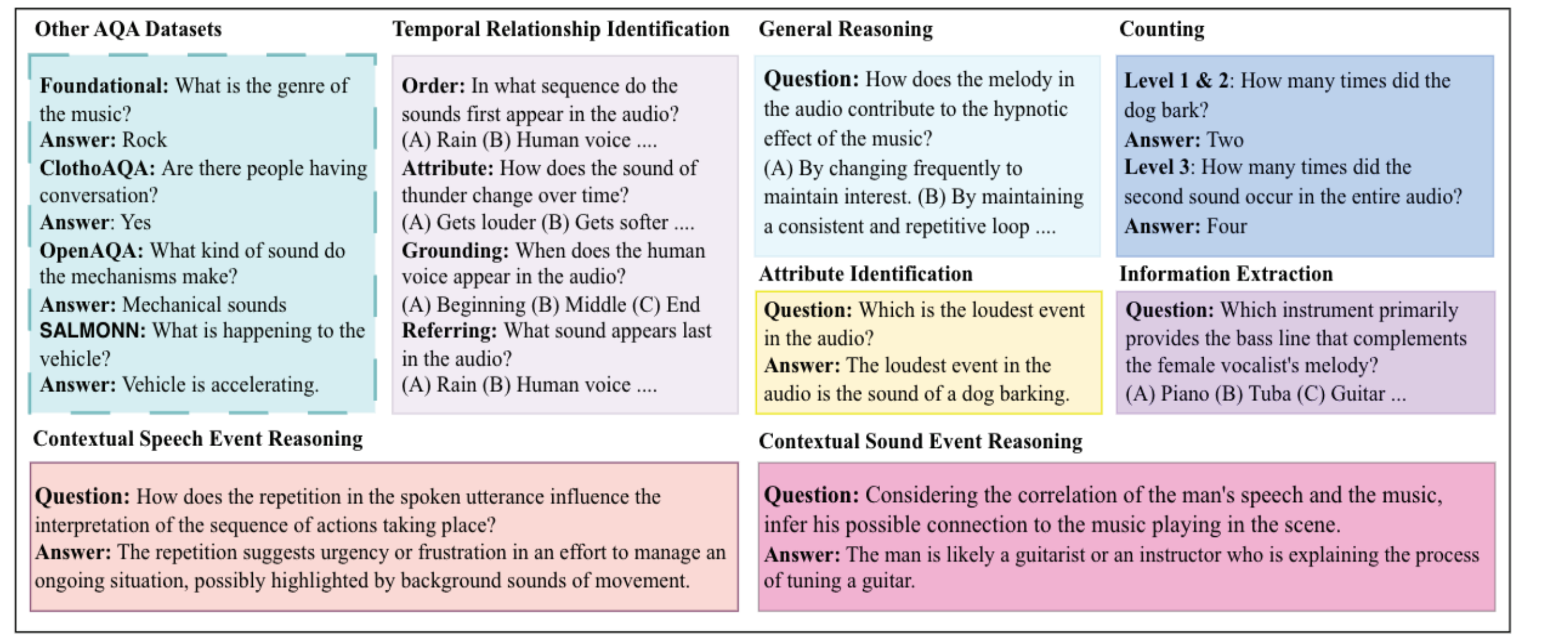
Prefix for Music & Audio Flamingo 3



Audio-Visual Flamingo

Data is All You Need!

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 ....

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

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.

Information Extraction

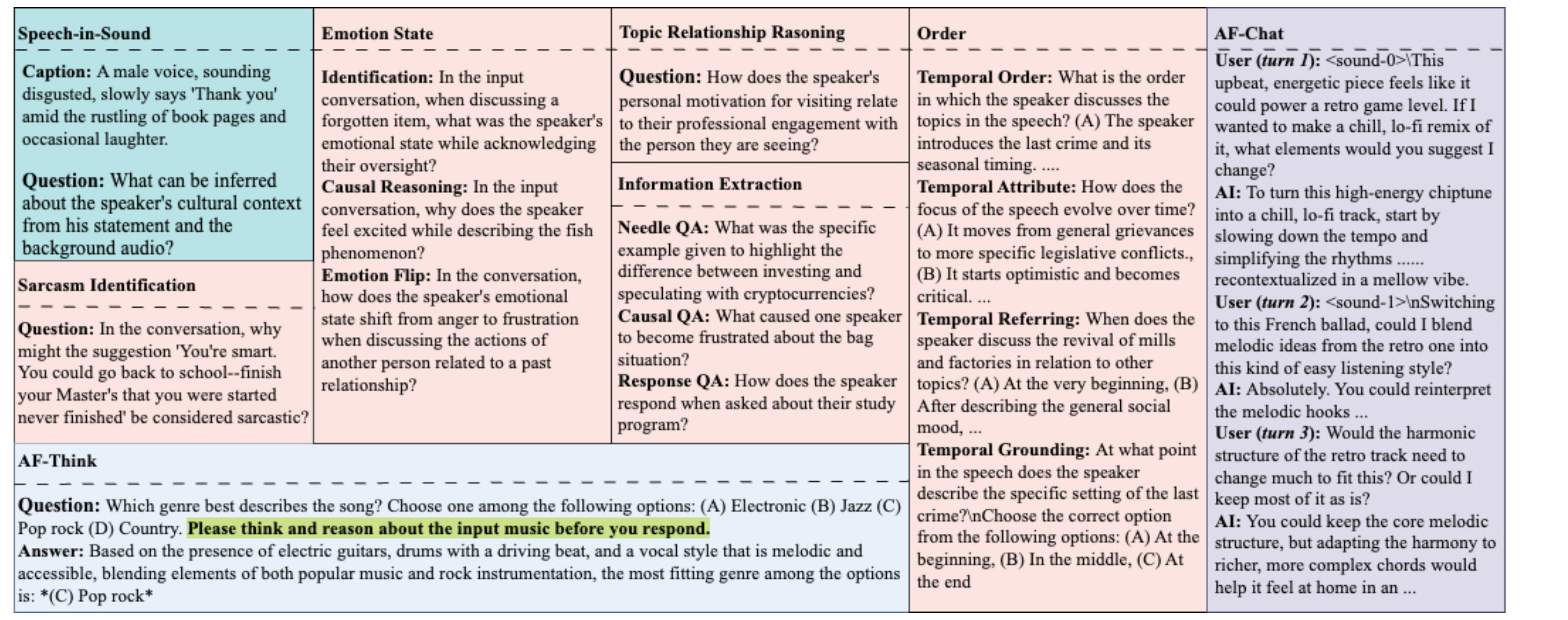
Question: Which instrument primarily provides the bass line that complements the female vocalist's melody? (A) Piano (B) Tuba (C) Guitar ...

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?

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 Rasoning

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?\nChoose the correct option from the following options: (A) At the beginning, (B) In the middle, (C) At the end

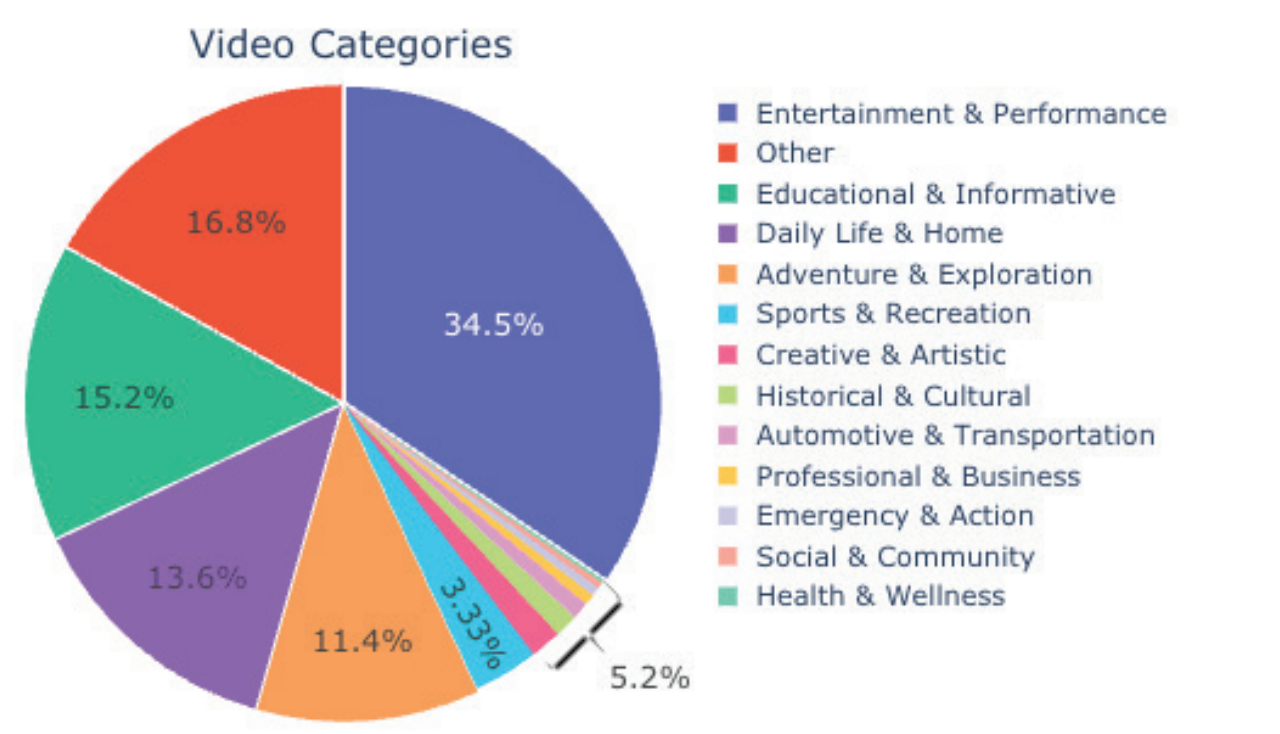
AF-Chat

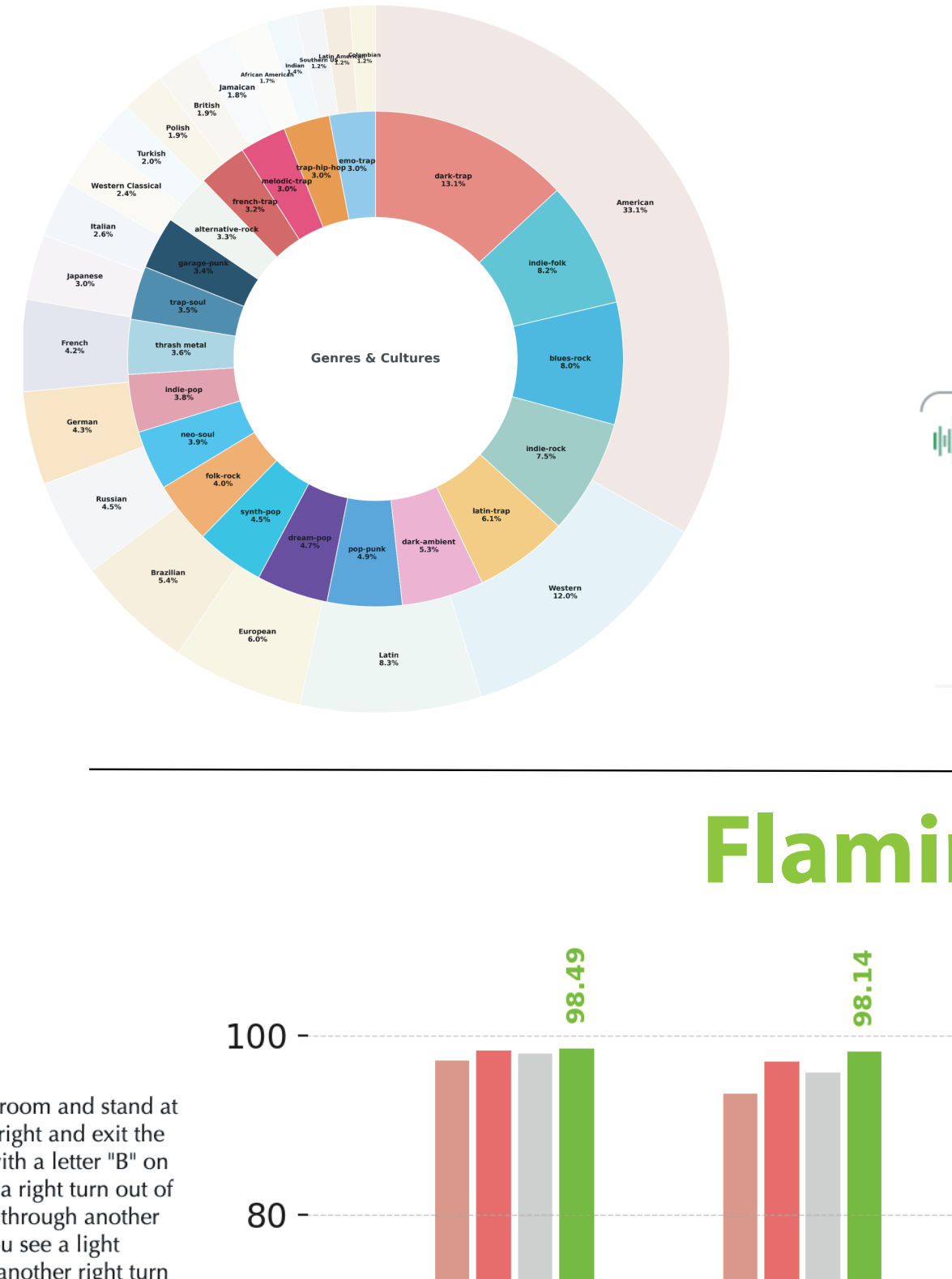
User (*turn 1*): <sound-0>\This upbeat, energetic piece feels like it could power a retro game level. If I wanted to make a chill, lo-fi remix of it, what elements would you suggest I change?
AI: To turn this high-energy chiptune into a chill, lo-fi track, start by slowing down the tempo and simplifying the rhythms ...... recontextualized in a mellow vibe.
User (*turn 2*): <sound-1>\nSwitching 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 ...

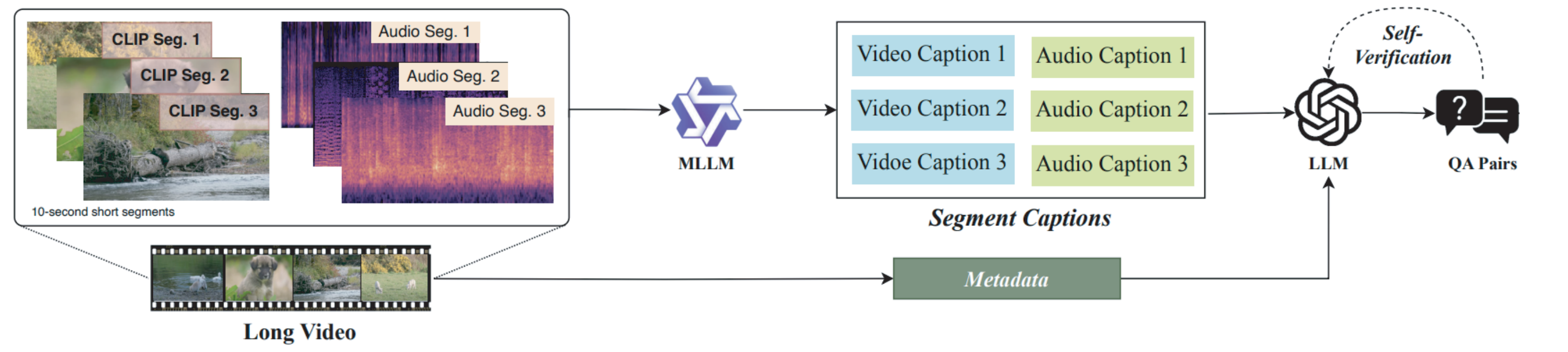
AF-Think

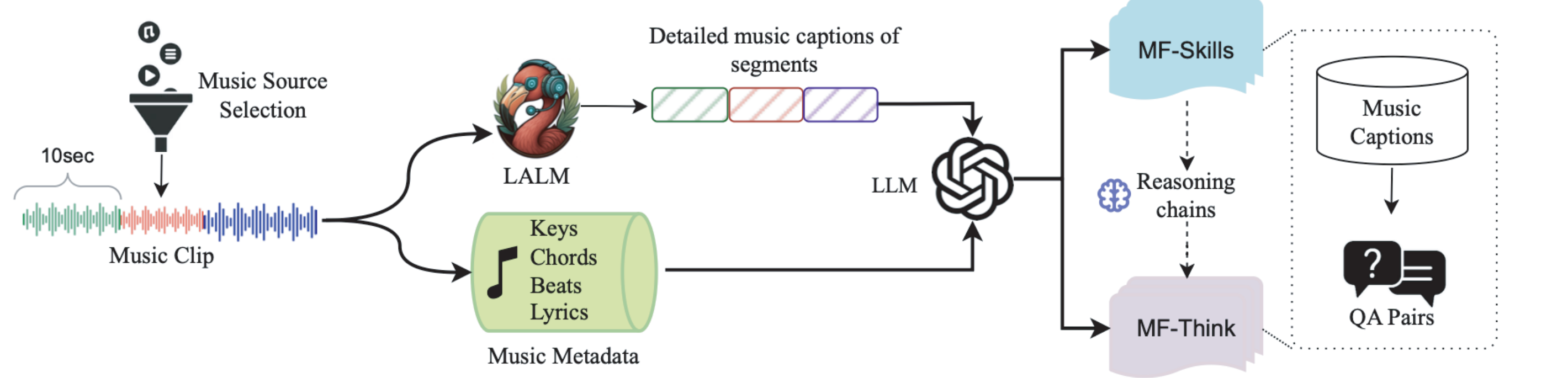
Question: Which genre best describes the song? Choose one among the following options: (A) Electronic (B) Jazz (C) Pop rock (D) Country. **Please think and reason about the input music before you respond.**
Answer: Based on the presence of electric guitars, drums with a driving beat, and a vocal style 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*

Open datasets focused on Long Audios and Full-length Songs









Audio Analysis + Think + Voice Chat (streaming)

AF3 Response: It is tennis.

Voice Output

+ **What sport is being played in the audio?**

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 usedto hit a ball, which is more indicative of tennis than badminton. Therefore, the most likely sport is tennis.

Multi-turn Multi-Audio Chat

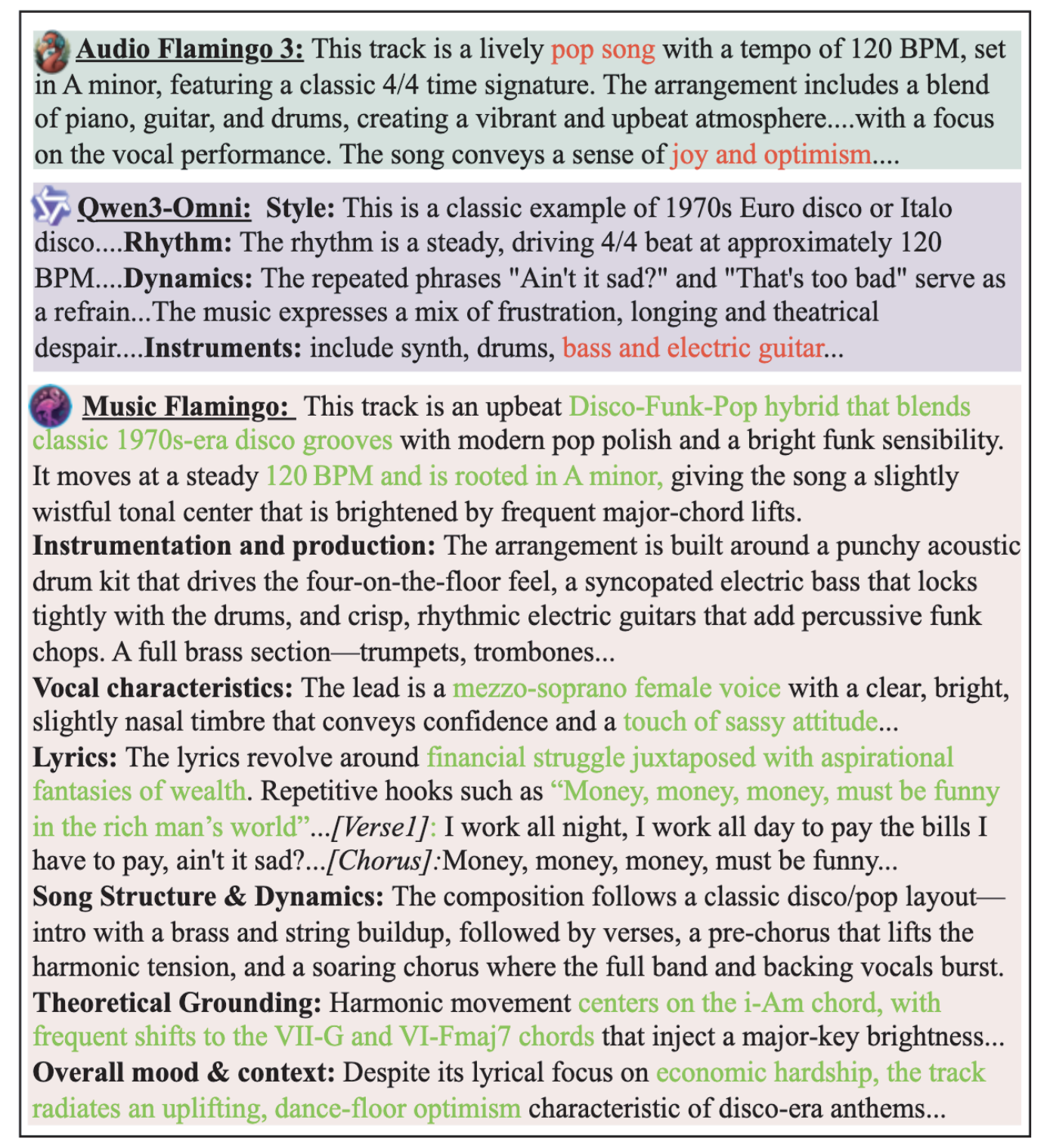
What elements of the track make it feel so calming and meditative?

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.

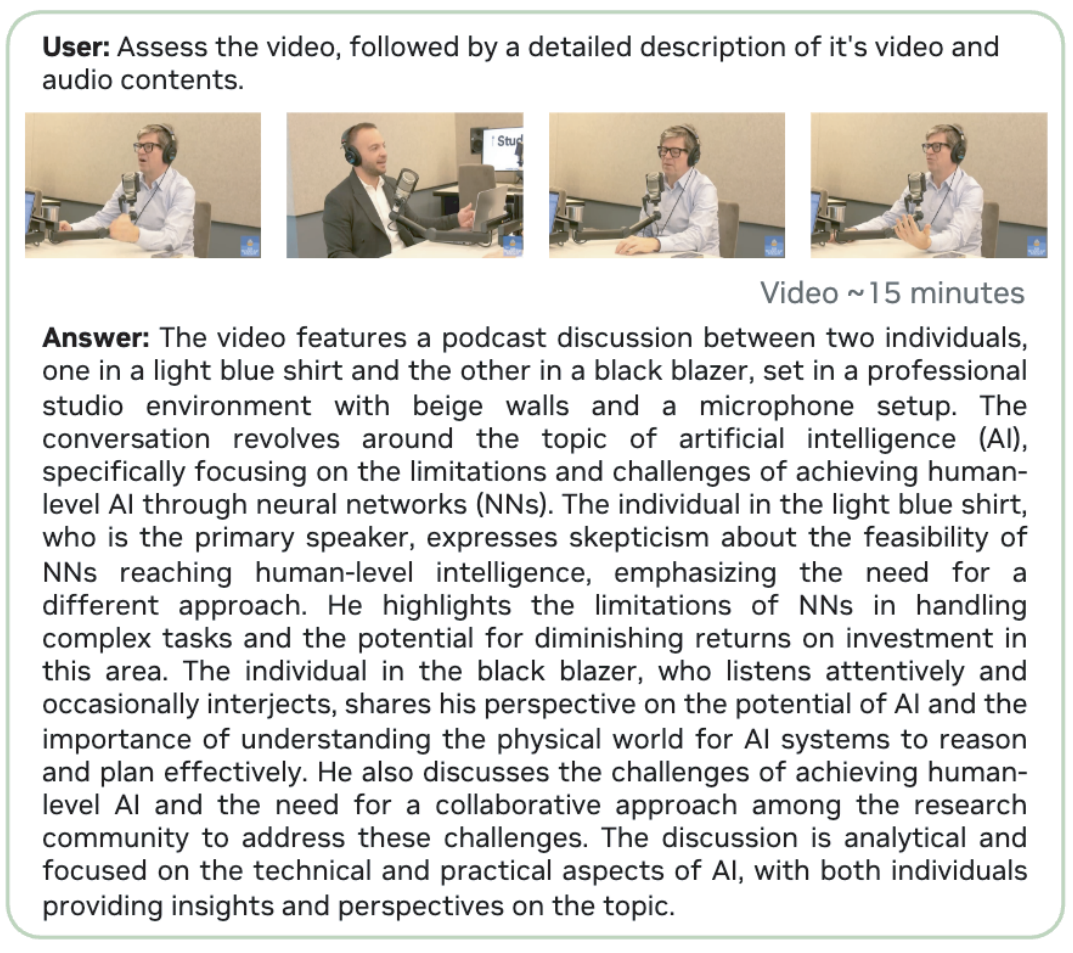
ABBA: Money Money Money (English) - 162M views



Audio Flamingo 3: This track is a lively pop song with a tempo of 120 BPM, set in A minor, featuring a classic 4/4 time signature. The arrangement includes a blend of piano, guitar, and drums, creating a vibrant and upbeat atmosphere....with a focus on the vocal performance. The song conveys a sense of joy and optimism....

Qwen3-Omni: **Style:** This is a classic example of 1970s Euro disco or Italo disco....**Rhythm:** The rhythm is a steady, driving 4/4 beat at approximately 120 BPM....**Dynamics:** The repeated phrases "Ain't it sad?" and "That's too bad" serve as a refrain...The music expresses a mix of frustration, longing and theatrical despair....**Instruments:** include synth, drums, bass and electric guitar...

Music Flamingo: This track is an upbeat Disco-Funk-Pop hybrid that blends classic 1970s-era disco grooves with modern pop polish and a bright funk sensibility. It moves at a steady 120 BPM and is rooted in A minor, giving the song a slightly wistful tonal center that is brightened by frequent major-chord lifts.
Instrumentation and production: The arrangement is built around a punchy acoustic drum kit that drives the four-on-the-floor feel, a syncopated electric bass that locks tightly with the drums, and crisp, rhythmic electric guitars that add percussive funk chops. A full brass section—trumpets, trombones...
Vocal characteristics: The lead is a mezzo-soprano female voice with a clear, bright, slightly nasal timbre that conveys confidence and a touch of sassy attitude...
Lyrics: The lyrics revolve around financial struggle juxtaposed with aspirational fantasies of wealth. Repetitive hooks such as "Money, money, money, must be funny in the rich man's world"...*[Verse1]*: I work all night, I work all day to pay the bills I have to pay, ain't it sad?...*[Chorus]*:Money, money, money, must be funny...
Song Structure & Dynamics: The composition follows a classic disco/pop layout—intro with a brass and string buildup, followed by verses, a pre-chorus that lifts the harmonic tension, and a soaring chorus where the full band and backing vocals burst.
Theoretical Grounding: Harmonic movement centers on the i-Am chord, with frequent shifts to the VII-G and VI-Fmaj7 chords that inject a major-key brightness...
Overall mood & context: Despite its lyrical focus on economic hardship, the track radiates an uplifting, dance-floor optimism characteristic of disco-era anthems...

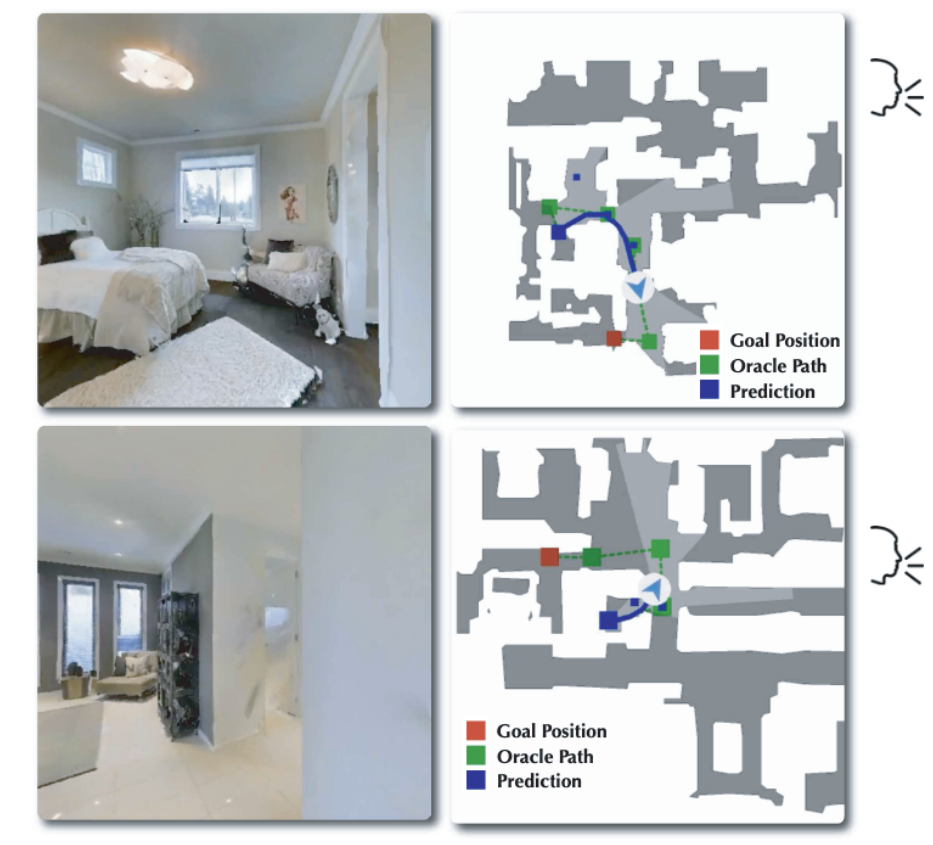


User: Assess the video, followed by a detailed description of it's video and audio contents.

Video ~15 minutes

Answer: The video features a podcast discussion between two individuals, one in a light blue shirt and the other in a black blazer, set in a professional studio environment with beige walls and a microphone setup. The conversation revolves around the topic of artificial intelligence (AI), specifically focusing on the limitations and challenges of achieving human-level AI through neural networks (NNs). The individual in the light blue shirt, who is the primary speaker, expresses skepticism about the feasibility of NNs reaching human-level intelligence, emphasizing the need for a different approach. He highlights the limitations of NNs in handling complex tasks and the potential for diminishing returns on investment in this area. The individual in the black blazer, who listens attentively and occasionally interjects, shares his perspective on the potential of AI and the importance of understanding the physical world for AI systems to reason and plan effectively. He also discusses the challenges of achieving human-level AI and the need for a collaborative approach among the research community to address these challenges. The discussion is analytical and focused on the technical and practical aspects of AI, with both individuals providing insights and perspectives on the topic.

Joint Vision-Audio Perception

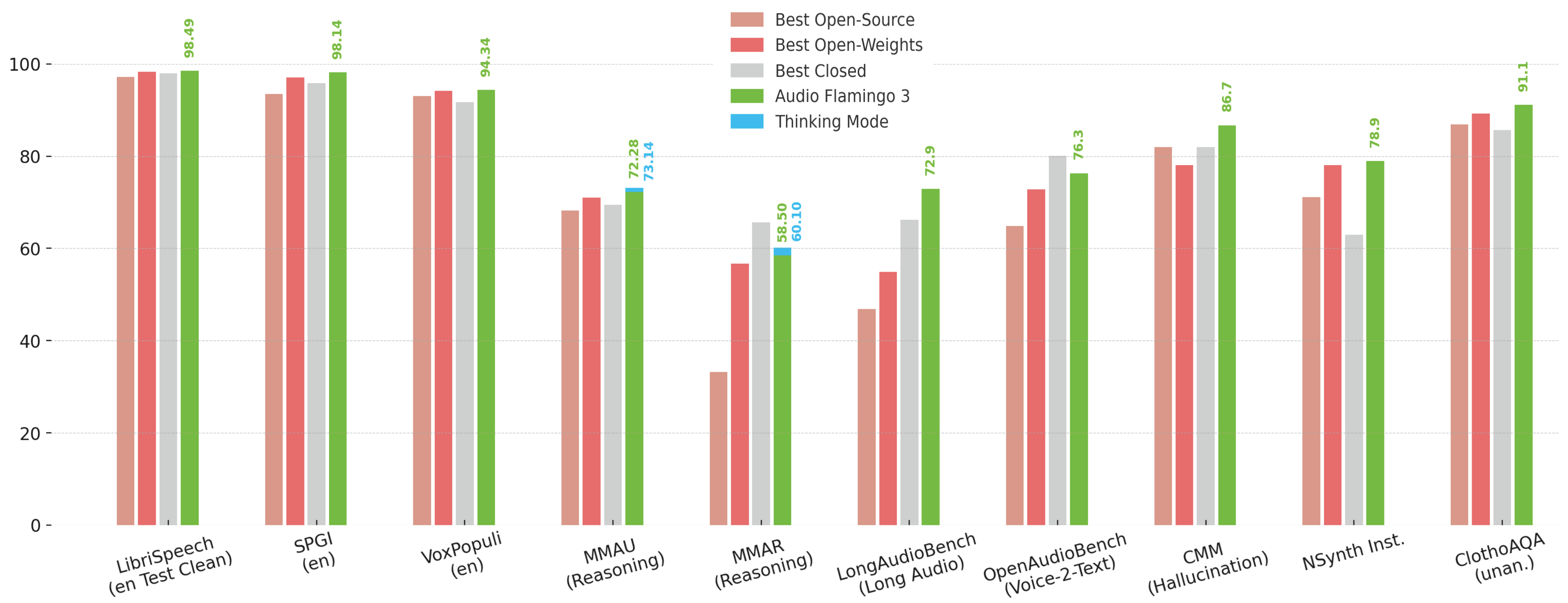


Instruction: Enter the bedroom and stand at the foot of the bed. Walk right and exit the bedroom from the door with a letter "B" on the wall to your left. Take a right turn out of the bedroom and walk in through another bedroom's door where you see a light colored shaggy rug. Take another right turn and stand just through the door after the white little dog on the floor to your left.
Agent: Move forward 75 cm.

Instruction: Step through the bathoom doors, make a left turn and take a step, then a sharp left turn and continue walking until you've made it through the bedroom door. Stop.
Agent: Turn left 15 degrees.

A Snapshot of MF and AVF capabilities!

Flamingo Models are SOTA across Benchmarks!


Best Open-Source
Best Open-Weights
Best Closed
Audio Flamingo 3
Thinking Mode
LibriSpeech (en Test Clean)
SPGI (en)
VoxPopuli (en)
MMAU (Reasoning)
MMAR (Reasoning)
LongAudioBench (Long Audio)
OpenAudioBench (Voice-2-Text)
CMM (Hallucination)
NSynth Inst.
ClothoAQA (unan.)