MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence

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, Siddhi Patil, Soham Deshmukh, Lasha Koroshinadze, Yao Liu, Leibny Paola Garcia Perera, Eleni Zanou, Themos Stafylakis, Joon Son Chung, David Harwath, Chao Zhang, Dinesh Manocha, Alicia Lozano-Diez, Santosh Kesiraju, Sreyan Ghosh, Ramani Duraiswami

Spatial QA

Question: Who's order does the waiter take first?
Options:
(A) The person to the left of the mic holder
(B) The person to the right of the mic holder
(C) The person in front of the mic holder
Answer: (A) The person to the left of the mic holder

Multicultural Music

Question: What *raag* is this *bandish* composed in?

Options:
(A) Bhimpalasi
(B) Kannada
(C) Durga
(D) Malkaunse
(E) Bhairavi
(F) Yaman Kalyan

Answer: (C) Durga

Voice QA

Question: Answer the question in the audio.
Options:
(A) It sounds like this decision carries a lot of weight...
(B) Perhaps the best approach is to systematically...
(C) Your tranquil state suggests you have a high degree of mental clarity right now. This is an excellent time to trust your judgment, as it's likely unclouded by....
(D) Making major decisions during emotional peaks,..
Answer: Your tranquil state suggests you have a high degree of mental clarity right now. This is an excellent time to trust your judgment, as it's likely unclouded by emotional turmoil.

Long Audio

Question: What did the winning team get at the end of the game?
Options:
(A) They choose the next vacation destination
(B) The other team has to get rid of their rooster
(C) They win the other team's apartment
(D) The game ends in a tie and nothing changes
\

Answer: They win the other team's apartment

Multi-Audio

Question: What effect needs to be applied to the first recording to achieve sound of the second recording?

Options:
(A) echo
(B) distortion
(C) phaser
(D) reverb

Answer: (D) reverb

Sound

Question: What trend can be observed in the weight of the cloths thrown in the audio?
Options:
(A) Increasing
(B) Decreasing
(C) Remains constant
(D) None of these options
Answer: (A) Increasing

Perceptual Skills: Acoustic Trend Estimation
Reasoning Skills: Temporal Reasoning

MMAU-PRO

Voice STEM QA

Audio Transcript: What is the packing efficiency, in percentage, of a solid where Atom X occupies the face-centered cubic lattice sites as well as alternate tetrahedral voids of the same lattice?
Question: Choose the correct option that answers the question in the audio from the options given below.
Options:
(A) 25%
(B) 35%
(C) 55%
(D) 75%
Answer: (B) 35%

Speech

Question: In the audio with roughly the same phrase being repeated, explain how the different tone effects the meaning of each of the 4 phrases and in what order they occur.
Options:
(A) Bored → enthusiastic → questioning → angry
(B) Cheerful → playful → serious → stern
(C) Sarcastic → sincere → doubtful → irritated
(D) genuine → sarcastic → questioning → frustrated.
Answer: (D) He can remove 0 bones
Perceptual Skills: Speech Activity, Turn-Taking and Overlap Detection
Reasoning Skills: Quantitative Reasoning (Counting/Arithmetic Comparison)

Multimodal Instruction Following

Audio Transcript: Instruction: Explain what's happening in this audio. Your answer must contain a title, wrapped in double angular brackets, such as <<poem of joy>>.
Question: What effect needs to be applied to the first recording to achieve sound of the second recording?

Correct Answer: <<Harmonica's Ascending and Descending Scale>> This audio clip features a harmonica playing the C major scale. The musician first plays the scale in an ascending order, moving from the lowest note to the highest.

Wrong Answer: This is an audio clip of a person playing the C major scale on a harmonica, first ascending and then descending.

Music

Question: Can you guess the singer in this song?
Options:
(A) Jeff Beck
(B) Tenacious B
(C) Jimmy Hendricks
(D) Jack Black
Answer: (D) Jack Black

Perceptual Skills: Timbre Perception and Instrument Recognition
Reasoning Skills: Musicological Knowledge

Open-ended QA

Question: What is hyper-foreignism with respect to pronunciation according to the clip?

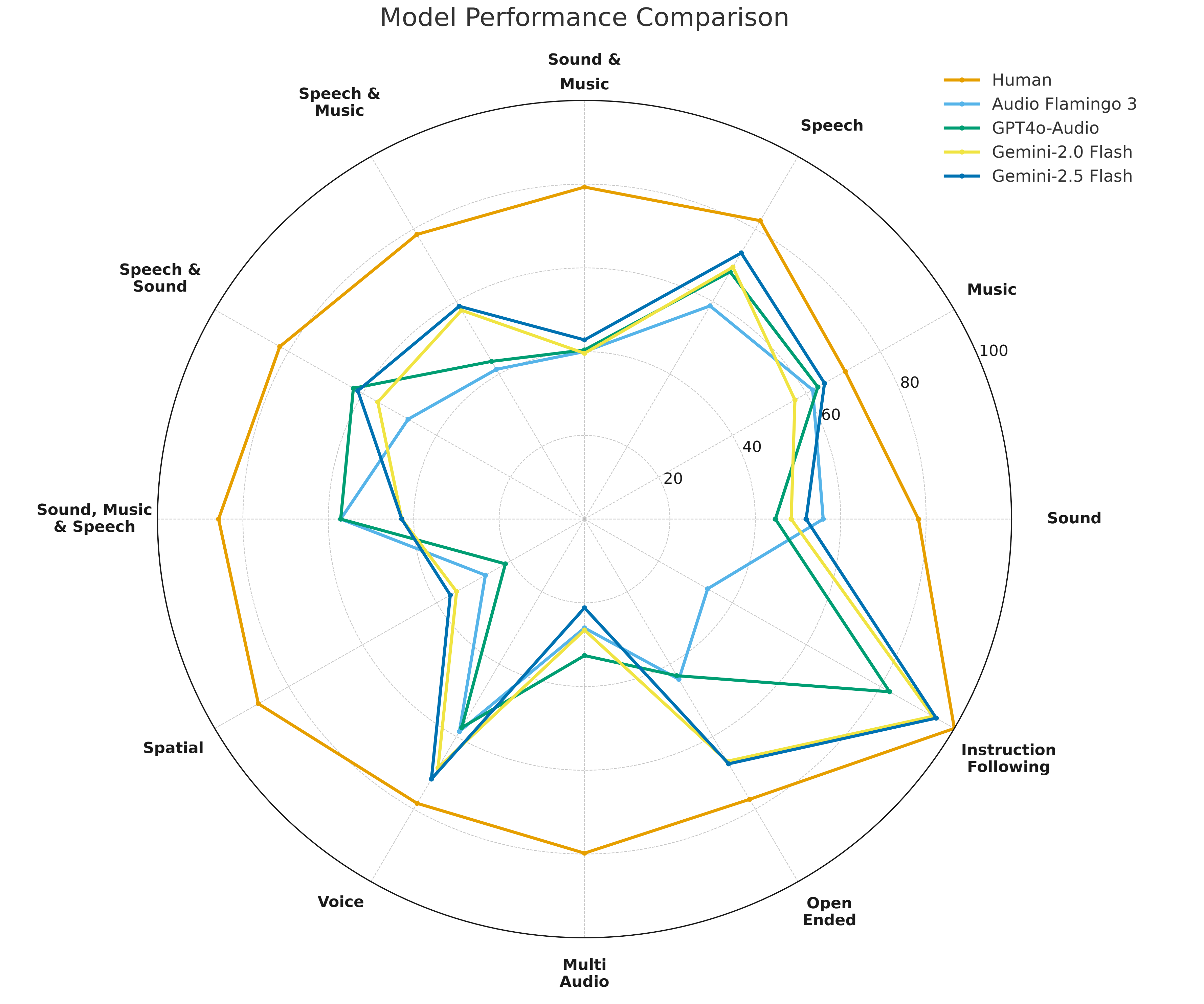
Answer: When a speaker changes the way they say a word to sound more like that of the stereotype they hold for a foreign language

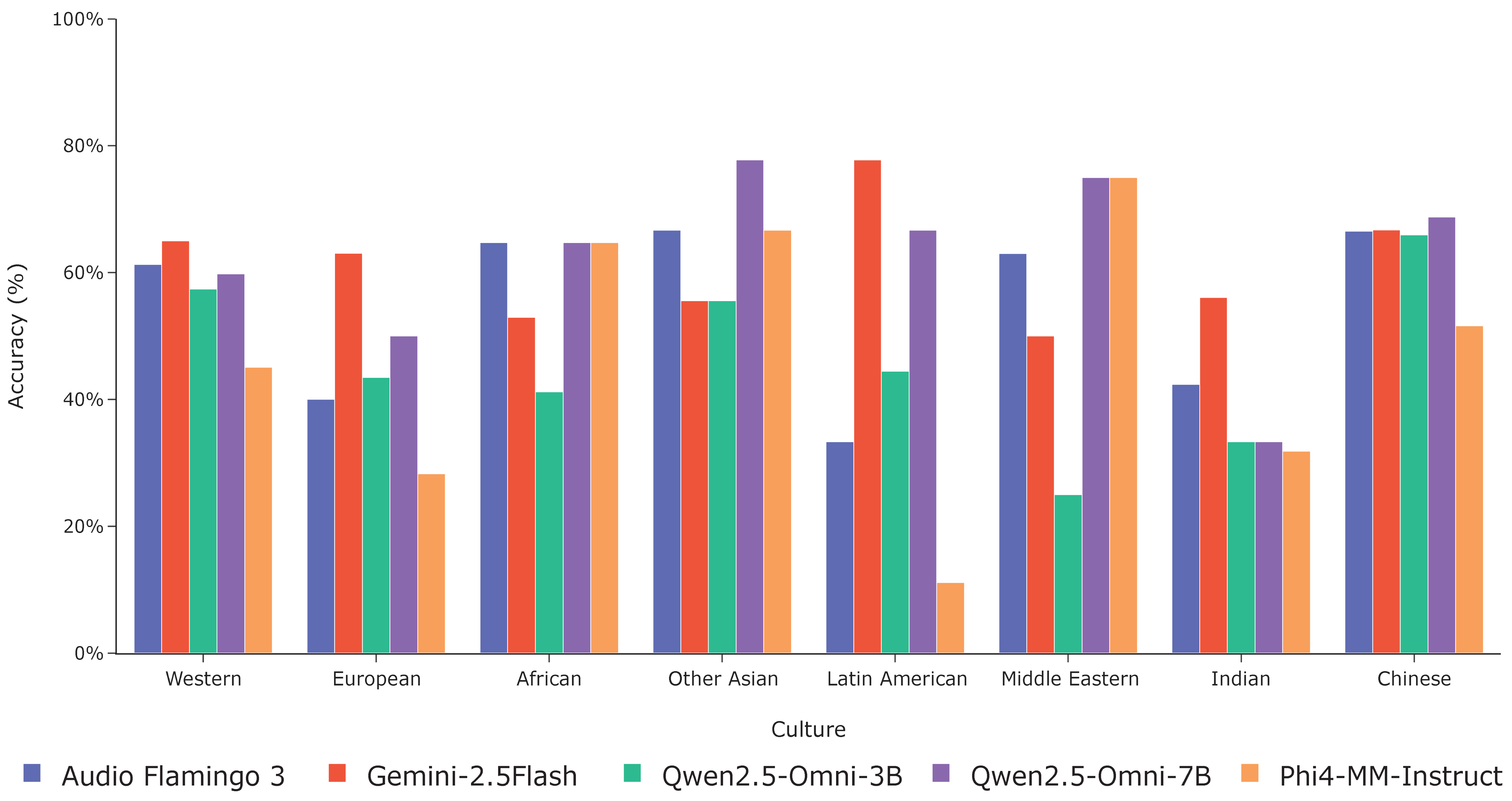
Speech-Sound-Music Mix

Question: Which of the following songs is made according to the speaker?

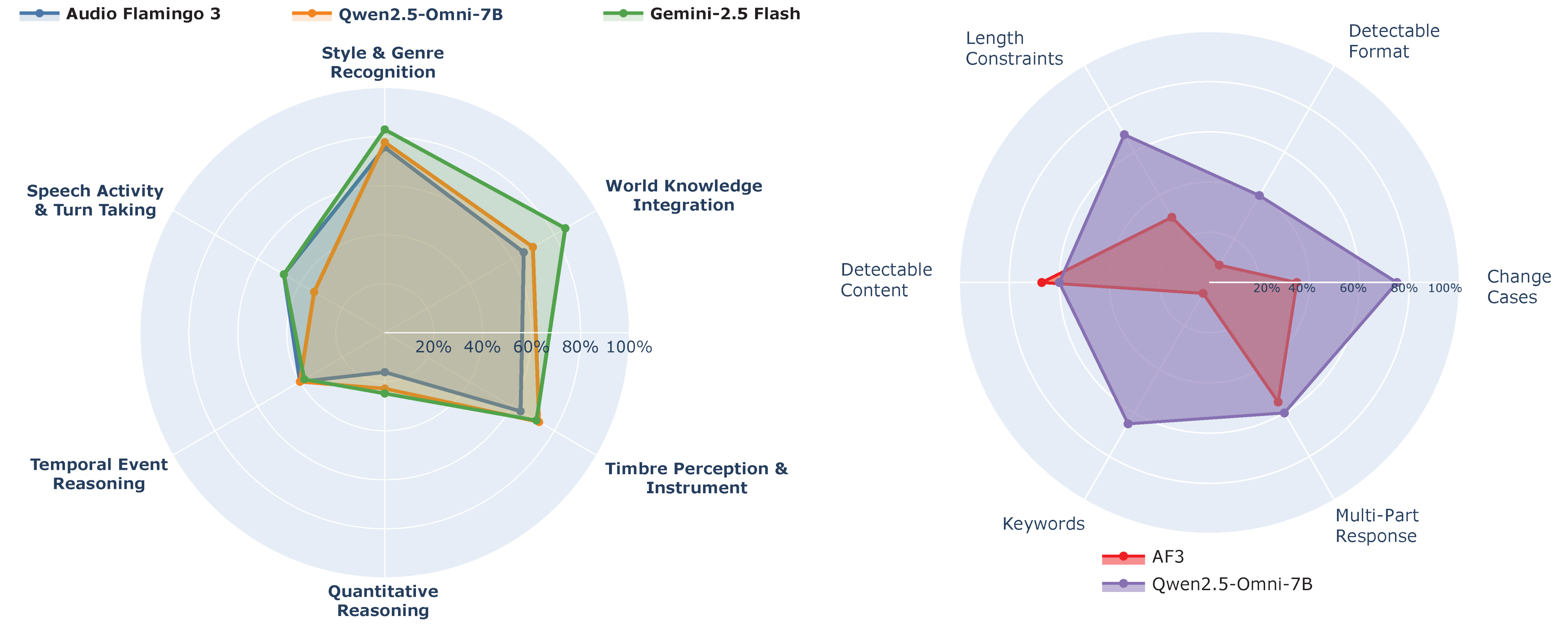
Options:
(A) Not Like Us
(B) Star Biy
(C) All the stars
(D) HUMBLE

Answer: (D) HUMBLE





Accuracy by music culture for five LALMs on the MMAU-Pro benchmark. Each bar group shows per-culture performance for AF3, Gemini-2.5 Flash, Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and Phi4-MM-Instruct, highlighting significant drops on



(Left.) Comparison of top 3 best and worst performing skills.
(Right.) Performance comparison of instruction following capabilities.

- We introduce MMAU-Pro, a comprehensive new benchmark for auditory intelligence featuring **5,305 expert-annotated Q&A pairs**, 49 skills, and challenging new tasks like spatial and multi-clip reasoning using long-form audio (up to 10 minutes).

- Top performers like Gemini 2.5 Flash only achieving 59.2% accuracy, and the best open-source models (Audio Flamingo 3, Qwen2.5-Omni) scoring just 51.7% and 52.2%, respectively.