MMAU

A Massive Multi-Task Audio Understanding and Reasoning Benchmark

정승규

BrAIn Lab.
Brain and Artificial Intelligence Lab.

MMAU

BrAIn Lab.
Brain and Artificial Intelligence Lab.

ICLR 2025 (Spotlight)

S Sakshi

팔로우

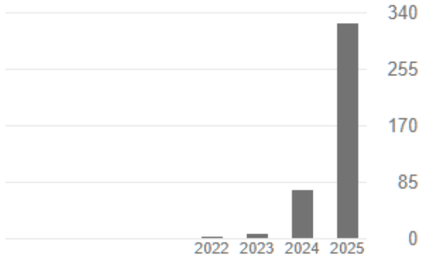
Ph.D. in CS at University of Maryland, College Park
umd.edu의 이메일 확인됨 - 홈페이지

Machine Learning Natural Language Processing Audio Processing

제목	인용	연도
MMAU: A massive multi-task audio understanding and reasoning benchmark S Sakshi, U Tyagi, S Kumar, A Seth, R Selvakumar, O Nieto, ... arXiv preprint arXiv:2410.19168	114	2024
Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities S Ghosh, S Kumar, A Seth, CKR Evuru, U Tyagi, S Sakshi, O Nieto, ... arXiv preprint arXiv:2406.11768	113	2024
Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities S Ghosh, Z Kong, S Kumar, S Sakshi, J Kim, W Ping, R Valle, D Manocha, ... arXiv preprint arXiv:2503.03983	55	2025
Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances S Ghosh, S Lepcha, S Sakshi, RR Shah, S Umesh arXiv preprint arXiv:2110.07592	36 *	2021
Compa: Addressing the gap in compositional reasoning in audio-language models S Ghosh, A Seth, S Kumar, U Tyagi, CK Evuru, S Ramaneswaran, ... arXiv preprint arXiv:2310.08753	29	2023
Dale: Generative data augmentation for low-resource legal nlp S Ghosh, CK Evuru, S Kumar, S Ramaneswaran, S Sakshi, U Tyagi, ... arXiv preprint arXiv:2310.15799	27	2023
M-MELD: a multilingual multi-party dataset for emotion recognition in conversations S Ghosh, S Ramaneswaran, U Tyagi, H Srivastava, S Lepcha, S Sakshi, ... arXiv preprint arXiv:2203.16799	6	2022

인용

	전체	2020년 이후
서지정보	409	409
h-index	6	6
i10-index	6	6



공동 저자

- Sreyan Ghosh — Ph.D. in CS at University of Mary...
- Sonal Kumar — Computer Science at University ...
- Dinesh Manocha — Distinguished University Profess...
- Utkarsh Tyagi — University of Maryland, College ...
- Chandra Kiran Reddy Evuru — ServiceNow
- Ashish seth — Computer Science at University

Introduction

BrAIn Lab.
Brain and Artificial Intelligence Lab.

- 기존 오디오 벤치마크의 한계 – NLP 나 VISION 분야와 달리 AUDIO AI 평가는 기초 작업에 편중되어 인간수중의 복잡한 추론 능력은 측정하지 못함
- 제안하는 MMAU 데이터셋은 음성 (Speech), 소리 (Sound), 음악 (Music) 3가지 핵심 오디오 도메인에 걸쳐 10,000개 이상의 오디오-질의응답 쌍으로 구성
- 단순 정보 검색을 넘어 27개의 고급 추론 능력을 요구하는 작업들로 구성
- 현존 모델의 명확한 한계 확인

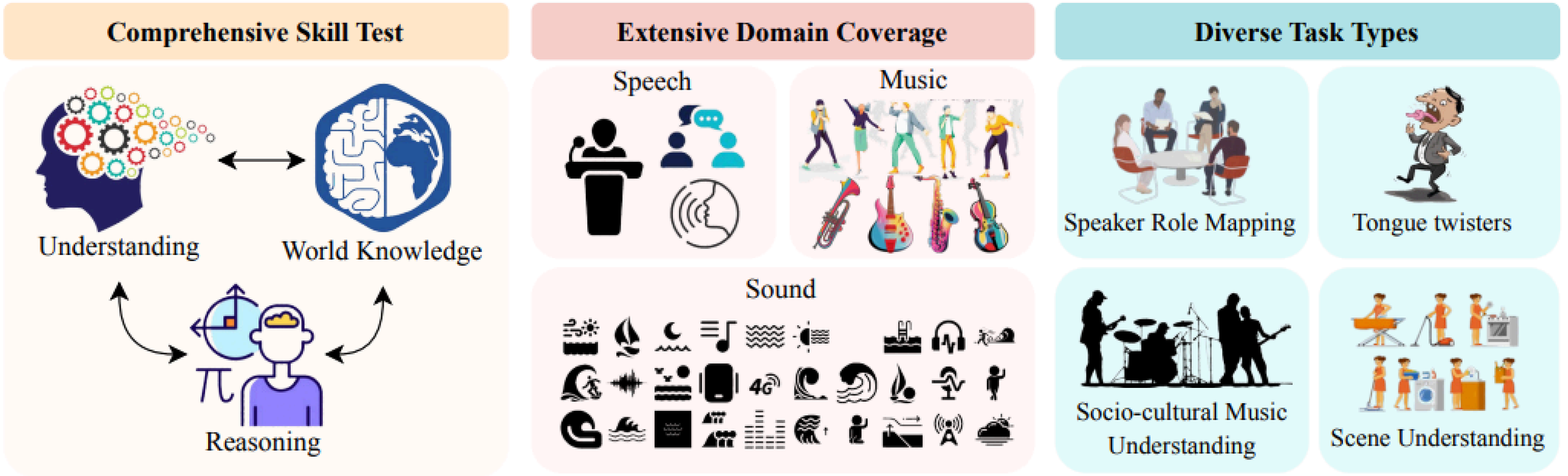
Related Work

BrAIn Lab.
Brain and Artificial Intelligence Lab.

- 초기에는 CLAP, AudioCLIP 과 같이 공유 표현을 학습하는 형태
- 최근에는 LLM의 추론 능력을 통합한 거대 오디오-언어 모델 (LALM) 형태로 발전
- 지금까지의 오디오 벤치마크데이터는 비교적 단순한 task, 해당 벤치마크는 고급 추론평가를 위해 생성

MMAU

BrAIn Lab.
Brain and Artificial Intelligence Lab.

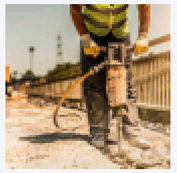


MMAU

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Sound

Info-extraction - Eco-Acoustic Knowledge
Question: What natural environment is most likely represented by the audio?



A. A serene forest
B. A quiet library
C. A construction site
D. A peaceful beach
Answer: C. **A construction site**

Reasoning - Temporal Event Reasoning
Question: For the given audio, identify which of the following sounds can be heard for the longest duration.

A. Video game sound
B. Music
C. Sound effect
D. Background noise
Answer: A. **Video game sound**

Speech

Info-extraction - Phonological Sequence Decoding
Question: For the given tongue twister identify which word appears first?

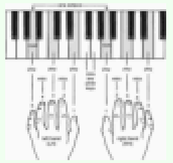
A. iron
B. aluminiuming
C. copperbottoming
D. none of these
Answer: B. **aluminiuming**

Reasoning - Multi Speaker Role Mapping
Question: Identify the role of the first and the second speaker in the conversation

A. Parent and child
B. Teacher and student
C. Doctor and patient
D. Coach and athlete
Answer: C. **Doctor and patient**

Music

Info-extraction - Harmony and Chord Progressions
Question: Which chord progression is used in the audio?



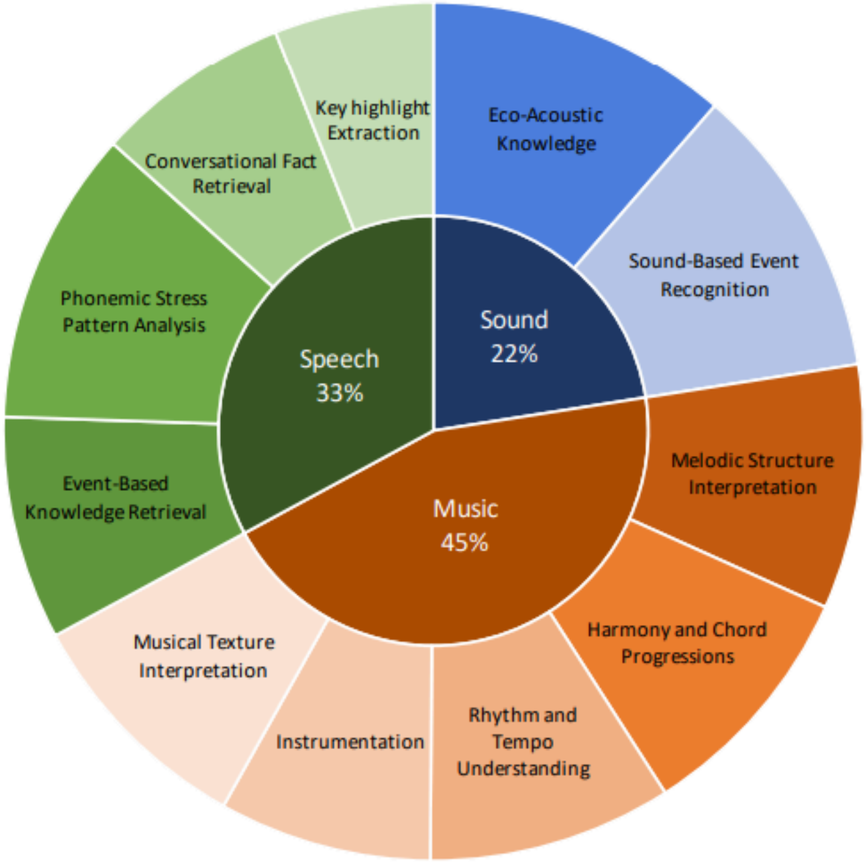
A. G, Em, B7, C6, E7, Am7
B. C, G, Am, F, Dm, E7
C. D, A, Bm, G, E, F#m,
D. A, E, F#m, D, Bm, C#m
Answer: A. **G, Em, B7, C6, E7, Am7**

Reasoning - Emotional Tone Interpretation
Question: What is the overall emotional atmosphere created by the combination of instruments in the audio?

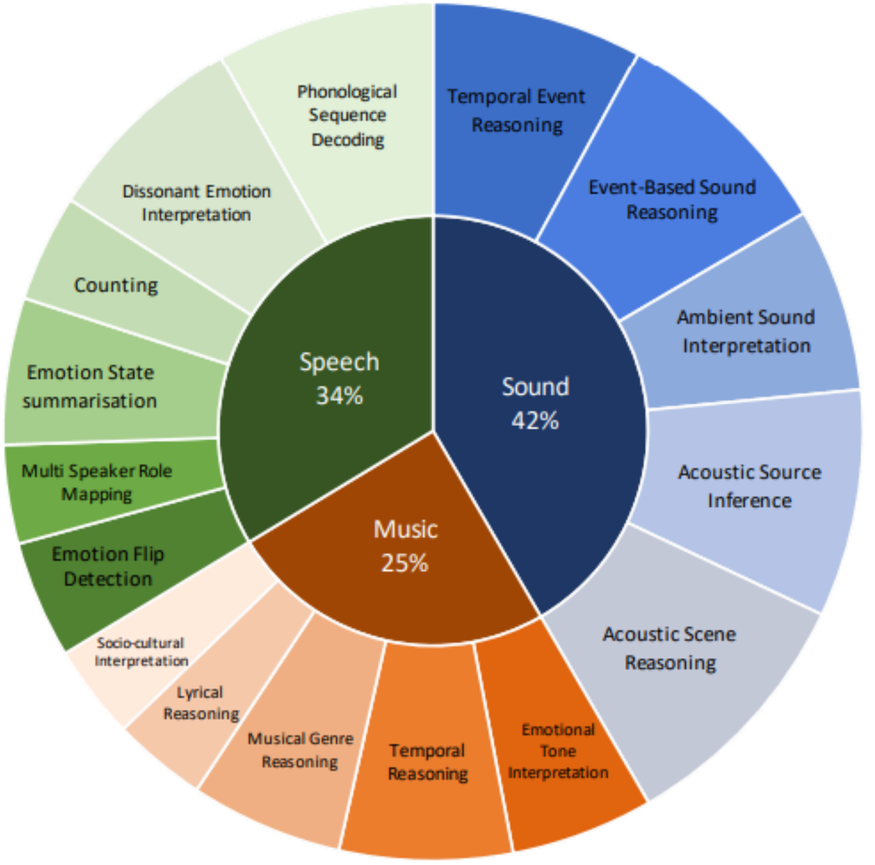
A. Ordinary and dull
B. Unique and heart-touching
C. Chaotic and confusing
D. Energetic and fast-paced
Answer: B. **Unique and heart-touching**

MMAU

BrAIn Lab.
Brain and Artificial Intelligence Lab.



Information Extraction



Reasoning Questions

MMAU

BrAIn Lab.
Brain and Artificial Intelligence Lab.

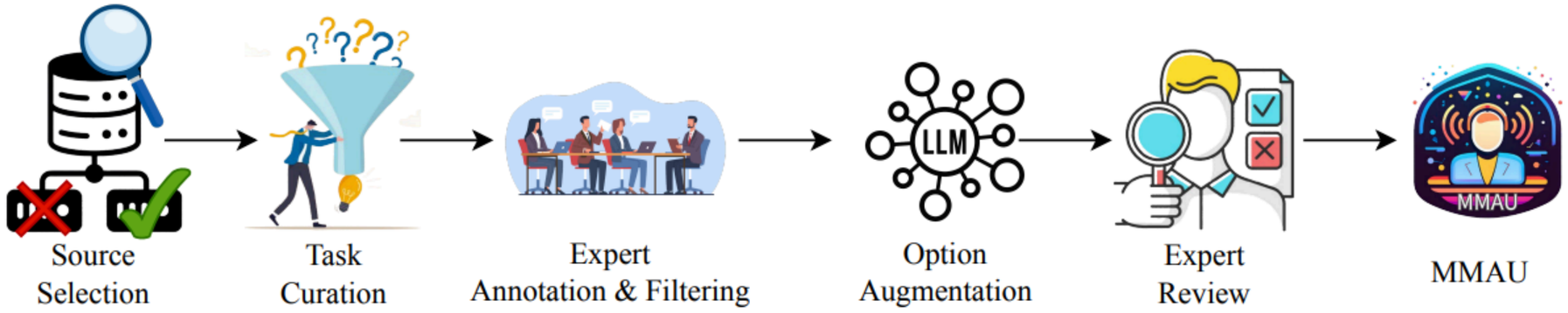
Statistics	Number
Total Questions	10,000
Audio Domains	3
Domain Categories (Speech:Music:Sound)	10:10:7
Difficulties (Easy: Medium: Hard)	22%:56%:22%
Splits (test-mini: test)	1000:9000
Information Extraction Based Questions	3499 (34.99%)
Reasoning Based Questions	6501 (65.74%)
Average question length	9.28 words
Average option length	5.23 words
Average audio length	10.14 sec

MMAU

BrAIn Lab.
Brain and Artificial Intelligence Lab.

데이터 생성 과정

1. 소스 선택 : 13개의 데이터셋에서 선택 (테스트 데이터에서 선택)
2. 작업 큐레이션 : 초기 90개 작업에서 -> 27개로 축소 (전문가와 협업)
3. 전문가 주석 : 각 오디오 클립에 대한 고품질의 질문과 답변 제작
4. 전문가 필터링 : 별도의 전문가 팀이 엄격하게 검토
5. 선택지 증강 : GPT-4를 이용하여 추가적인 답변 선택지를 증강. 무작위 증강이 아닌 오디오의 맥락과 질문을 기반으로 생성. 증강 후 전문가 필터링
6. 전문가 검토
7. MMAU 최종화 : 10,000개의 데이터 선택. 평가를 위한 1,000개는 test-mini, 나머지는 test



MMAU

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Benchmark	Size	Domain			Tasks				Expert Comments	Difficulty Level
		Speech	Sound	Music	Info Extraction		Reasoning			
CompA	600	×	✓	×	0	×	0.6k	✓	Requires only sound event sequence understanding. Limited in reasoning depth and knowledge scope.	2.0
CompA-R	1.5k	×	✓	×	0	×	1.5k	✓	Restricted to sounds and only contextual event understanding. Limited in knowledge scope.	3.0
MuChin	1k	×	×	×	0	×	0	×	Restricted to music with minimal reasoning depth. Limited in knowledge scope.	2.5
MusicBench	0.4k	×	×	✓	0	×	0	×	Restricted to music with minimal reasoning depth. Limited in knowledge scope.	2.5
MuChoMusic	1.2k	×	×	✓	0.7k	✓	0.4k	✓	Restricted to music with open-ended answers. Limited in knowledge scope.	3.5
OpenASQA	8.8k	✓	✓	×	8.8k	✓	0	×	Requires limited acoustic scene understanding. Does not require external or expert knowledge.	3.0
AudioBench	100k+	✓	✓	✓	5k	✓	0	×	Basic acoustic information retrieval with minimal reasoning depth and complexity. Does not require external or expert knowledge.	3.5
AIR-Bench	19k	✓	✓	✓	1.2k	✓	0.8k	✓	Basic acoustic information retrieval with minimal reasoning depth and complexity. Does not require external or expert knowledge.	2.5
MMAU *(ours)*	10K	✓	✓	✓	4.5k	✓	5.2k	✓	Requires fine-grained audio understanding with expert-level, multi-step reasoning and specialized knowledge across a broad range of topics.	4.5

Result

BrAIn Lab.
Brain and Artificial Intelligence Lab.

평가 대상 모델

1. LALM (Large Audio Language Model) : Qwen2-Audio, GAMA 와 같은 Open source + Gemini-Pro 같은 폐쇄형 모델 포함
2. LLM (Large Language Model) : GPT-4o, LLaMA 등

LLM의 경우 Qwen2-Audio 가 생성한 캡션을 입력으로 제공

평가 방식 : Micro-Averaged Accuracy 사용

현재 LALM은 개방형 응답을 생성하도록 구현 되어있어, 문자열 매칭을 사용

Result

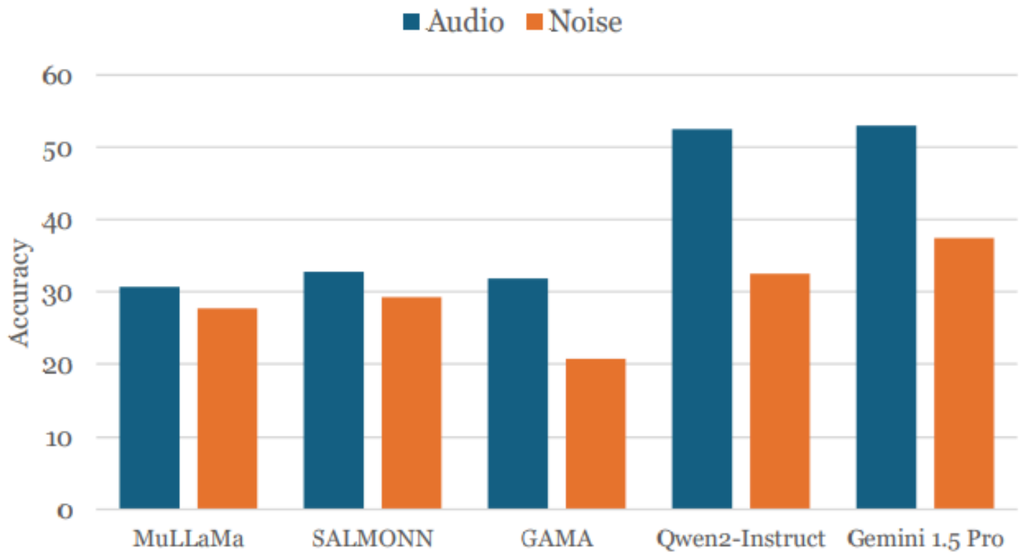
BrAIn Lab.
Brain and Artificial Intelligence Lab.

Models	Size	{So, Mu, Sp}	Sound		Music		Speech		Avg	
			Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Random Guess	-	-	26.72	25.73	24.55	26.53	26.72	25.50	26.00	25.92
Most Frequent Choice	-	-	27.02	25.73	20.35	23.73	29.12	30.33	25.50	26.50
Human (test-mini)	-	-	86.31	-	78.22	-	82.17	-	82.23	-
Large Audio Language Models (LALMs)										
Pengi	323M	✓ ✓ ×	06.10	08.00	02.90	03.05	01.20	01.50	03.40	04.18
Audio Flamingo Chat	2.2B	✓ ✓ ×	23.42	28.26	15.26	18.20	11.41	10.16	16.69	18.87
LTU	7B	✓ ✓ ×	22.52	25.86	09.69	12.83	17.71	16.37	16.89	18.51
LTU AS	7B	✓ ✓ ✓	23.35	24.96	9.10	10.46	20.60	21.30	17.68	18.90
MusiLingo	7B	× ✓ ×	23.12	27.76	03.96	06.00	05.88	06.42	10.98	13.39
MuLLaMa	7B	× ✓ ×	40.84	44.80	32.63	30.63	22.22	16.56	31.90	30.66
M2UGen	7B	× ✓ ×	03.60	03.69	32.93	30.40	06.36	04.53	14.28	12.87
GAMA	7B	✓ ✓ ×	41.44	45.40	32.33	30.83	18.91	19.21	30.90	31.81
GAMA-IT	7B	✓ ✓ ×	43.24	43.23	28.44	28.00	18.91	15.84	30.20	29.02
Qwen-Audio-Chat	8.4B	✓ × ×	55.25	56.73	44.00	40.90	30.03	27.95	43.10	41.86
Qwen2-Audio	8.4B	✓ ✓ ✓	07.50	08.20	05.14	06.16	03.10	04.24	05.24	06.20
Qwen2-Audio-Instruct	8.4B	✓ ✓ ✓	54.95	45.90	50.98	53.26	42.04	45.90	49.20	52.50
SALAMONN	13B	✓ ✓ ✓	41.00	40.30	34.80	33.76	25.50	24.24	33.70	32.77
Gemini Pro v1.5	-	-	**56.75**	54.46	49.40	48.56	**58.55**	55.90	54.90	52.97
Gemini 2.0 Flash	-	-	56.46	**61.73**	**58.68**	**56.53**	51.65	**61.53**	**55.60**	**59.93**
Large Language Models (LLMs)										
GPT4o + weak cap.	-	-	39.33	35.80	39.52	41.9	58.25	68.27	45.70	48.65
GPT4o + strong cap.	-	-	**57.35**	**55.83**	49.70	**51.73**	**64.86**	**68.66**	**57.30**	**58.74**
Llama-3-Ins. + weak cap.	8B	-	34.23	33.73	38.02	42.36	54.05	61.54	42.10	45.87
Llama-3-Ins. + strong cap.	8B	-	50.75	49.10	**50.29**	48.93	55.25	62.70	52.10	53.57

Result

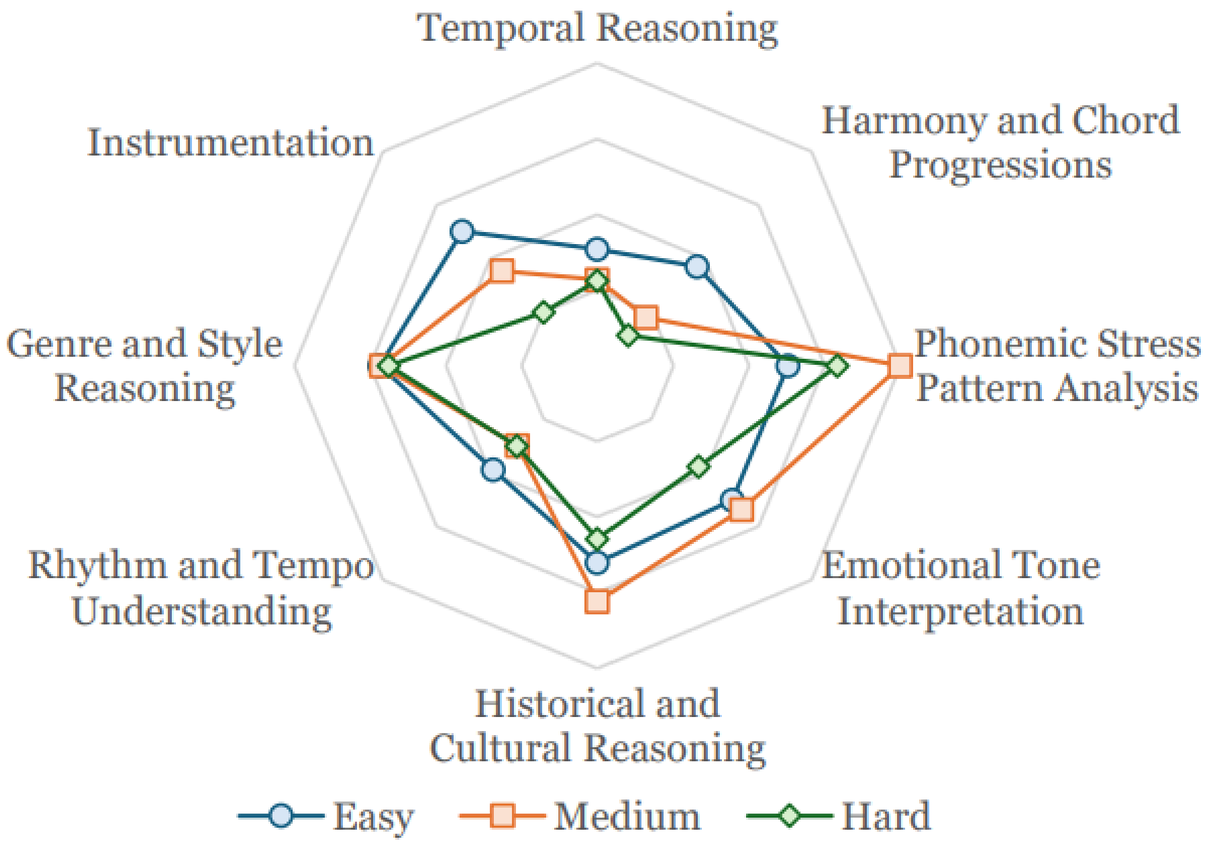
BrAIn Lab.
Brain and Artificial Intelligence Lab.

Models	Size	Sound	Music	Speech	Avg
CompA-CLAP	416M	42.66	38.20	27.98	36.28
ReCLAP	416M	47.43	34.83	**29.51**	37.26
LAION-CLAP	416M	45.10	35.19	25.61	35.30
MS CLAP 2023	159M	**52.10**	**40.00**	28.78	**40.29**



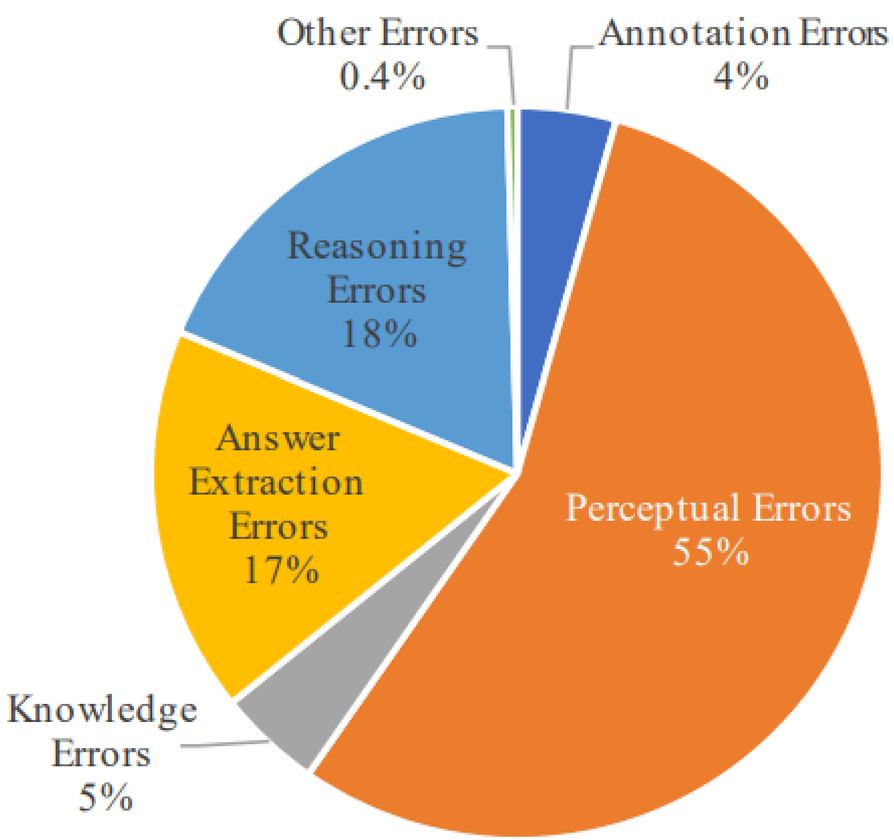
Result

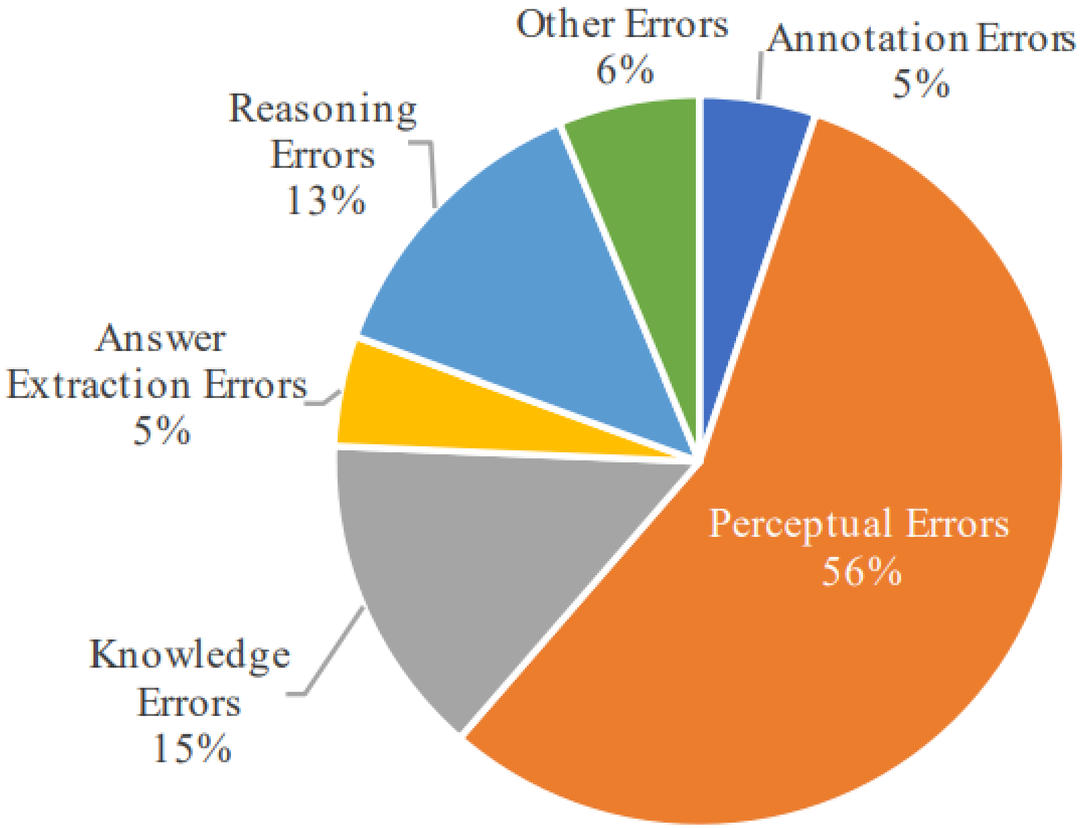
BrAIn Lab.
Brain and Artificial Intelligence Lab.



Result

BrAIn Lab.
Brain and Artificial Intelligence Lab.





Result

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Error Type	Definition	Question	Prediction	Reason
Perceptual Error	The model fails to perceive the audio correctly.	Based on the given audio, identify the source of the following sound. **Choices**: **A. Stream** B. Faucet C. Waterfall D. Rain	Waterfall	Misinterpreted the sound
Knowledge Error	The model understands the audio but lacks the knowledge to answer.	What is the typical frequency range of the instrument playing in the background? **Choices**: **A. The bass typically ranges from 40 Hz to 400 Hz.** B. The bass typically ranges from 400 Hz to 4 kHz. C. The bass typically ranges from 20 Hz to 200 Hz. D. The bass typically ranges from 4 kHz to 40 kHz.	20-200 Hz	Lacked specific frequency knowledge

Result

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Reasoning Error	The model struggles with logical reasoning.	What weather condition is indicated by the audio? **Choices**: **A. Windy** B. Calm C. Humid D. Rainy	Humid	Incorrect reasoning about sound

Conclusion

BrAIn Lab.
Brain and Artificial Intelligence Lab.

- MMAU 는 현재 최고 수준의 모델조차 59%정도의 정확도를 기록하며 LALM 모델의 한계 입증
- 향후 계획 1 : 현재 분리되어있는 task(정보추출, 추론) 기술을 모두 요구하는 복합적인 작업 추가
- 향후 계획 2 : 현재는 전부 객관식인데 향후 주관식 문답으로 확장 현재 40초인 입력 제한을 늘려 긴 오디오 처리 능력 평가

Leaderboard: MMAU-v05.15.25

Open-Source | Open-Access | Proprietary | Fine-tuned

		Sound		Music		Speech		Avg	
Name	Size	Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Audio-Thinker	8.4B	81.98	78.8	74.25	73.8	76.88	75.16	77.7	75.98
Step-Audio-2	-	84.04	80.60	73.56	68.23	75.15	72.75	77.58	73.86
MiMo-Audio	7B	81.68	77.2	74.25	69.73	68.17	70.77	74.7	72.59
Audio Flamingo 3	8.2B	79.58	75.83	73.95	74.47	66.37	66.97	73.30	72.42
Qwen2.5-Omni	8.2B	78.10	76.77	65.90	67.33	70.60	68.90	71.50	71.00
Step-Audio-2-mini	8.3B	79.30	75.57	68.44	66.85	68.16	66.49	72.73	70.23
Gemini 2.5 Pro	-	75.08	70.63	68.26	64.77	71.47	72.67	71.60	69.36
Gemini 2.5 Flash	-	73.27	69.50	65.57	69.40	76.58	68.27	71.80	67.39
Gemini 2.0 Flash	-	71.17	68.93	65.27	59.30	75.08	72.87	70.50	67.03
DeSTA2.5-Audio	8B	70.27	66.83	56.29	57.10	71.47	71.94	66.00	65.21
Kimi-Audio	8.2B	75.68	70.70	66.77	65.93	62.16	56.57	68.20	64.40
Audio Reasoner	8.2B	67.87	67.27	69.16	61.53	66.07	62.53	67.70	63.78
Phi-4-multimodal	5.5B	65.47	62.67	64.37	61.97	67.27	63.80	65.70	62.81
Gemini 2.5 Flash Lite	-	63.06	62.50	63.47	54.87	72.07	67.47	66.20	61.61
Audio Flamingo 2	3B	71.47	68.13	70.96	70.20	44.74	44.87	62.40	61.06
GPT-4o Audio	-	64.56	63.20	56.29	49.93	66.67	69.33	62.50	60.82
Qwen2-Audio-Instruct	7B	67.27	61.17	56.29	55.67	55.26	55.37	59.60	57.40

Plot Development reasoning task

Question: Based on these two audio segments, which of the following inferences are reasonable?

A. Before the new medical team was assigned tasks, the hospital was already operating beyond capacity.
B. The medical teams from Guangzhou and Shanghai were dispatched as expert reinforcement forces to urgently take over a newly expanded intensive care unit.
C. The initial meeting in the first audio segment is a standard welcome ceremony, aimed at giving the visiting team time to rest and familiarize themselves with the environment before starting work.
D. Director Zhang announced in the second audio segment that the newly renovated ICU was handed over to the new team, indicating that the hospital's patient intake pressure had been alleviated.

Answer: A B
A. Before the new medical team was assigned tasks, the hospital was already operating beyond capacity.
B. The medical teams from Guangzhou and Shanghai were dispatched as expert reinforcement forces to urgently take over a newly expanded intensive care unit.

Multi-Character Interaction Reasoning task

Question: Based on these two audio segments, which of the following inferences are reasonable?

A. After a brief discussion, the people in the first audio finally made a collective decision and took action
B. The dialogue in the second audio is tense due to a heated argument over high heels and snacks
C. The conversation in the second audio mainly revolves around daily trivialities and personal preferences, with a relatively relaxed atmosphere
D. Both audio clips mention behaviors related to "departure" or "departure", but the background and emotional state of the characters are completely different

Answer: A C D
A. After a brief discussion, the people in the first audio finally made a collective decision and took action
C. The conversation in the second audio mainly revolves around daily trivialities and personal preferences, with a relatively relaxed atmosphere
D. Both audio clips mention behaviors related to "departure" or "departure", but the background and emotional state of the characters are completely different

Anomaly Detection and Safety

Question: Based on the events that suddenly occurred at the end of the audio, what is most likely to have happened?

A. The woman jumped into the water and screamed on purpose for dramatic effect.
B. The woman accidentally slipped into the water while speaking excitedly.
C. The audio recording equipment malfunctioned, producing background noise of water sounds and screams.
D. The woman's friend suddenly pushed her into the water as a joke.

Answer: B The woman accidentally slipped into the water while speaking excitedly.

Scene Understanding

Question: Which sports event's final is most likely to be recorded in the audio?

A. Olympic individual archery finals
B. Biathlon (cross-country skiing and rifle shooting)
C. Olympic 10m Air Rifle Final
D. World Cup of Flying Saucer Shooting

Answer: C Olympic 10-meter air rifle final
Sub-category： Scene Element Recognition

Social Relationships and Social Reasoning

Question: Based on the audio, under what scenario is this conversation most likely to take place?

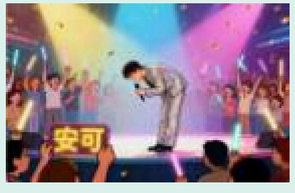
A. At a housewarming party, friends expressed their blessings to the host.
B. At the end of a business meeting, the partners reached a pleasant consensus.
C. At a wedding ceremony, elders give their blessings to the newlyweds.
D. On a solemn occasion of parting due to certain reasons, one party solemnly entrusts the other party.

Answer: D In a solemn occasion of parting due to some reason, one party solemnly entrusts to the other party.
Sub-category： Social Intention Reasoning task

Event Reasoning

Question: Considering audio, what is most likely to happen next?

A. The scene fell into a long silence and contemplation
B. The speaker fainted due to physical exhaustion
C. The audience erupted into even more enthusiastic cheers and applause,
D. Everyone began to calmly discuss the details and risks of the plan

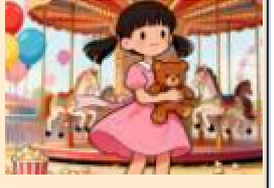


Answer: C The audience erupted into even more enthusiastic cheers and applause.
Sub-category： Event Causal Reasoning Task

Temporal Reasoning

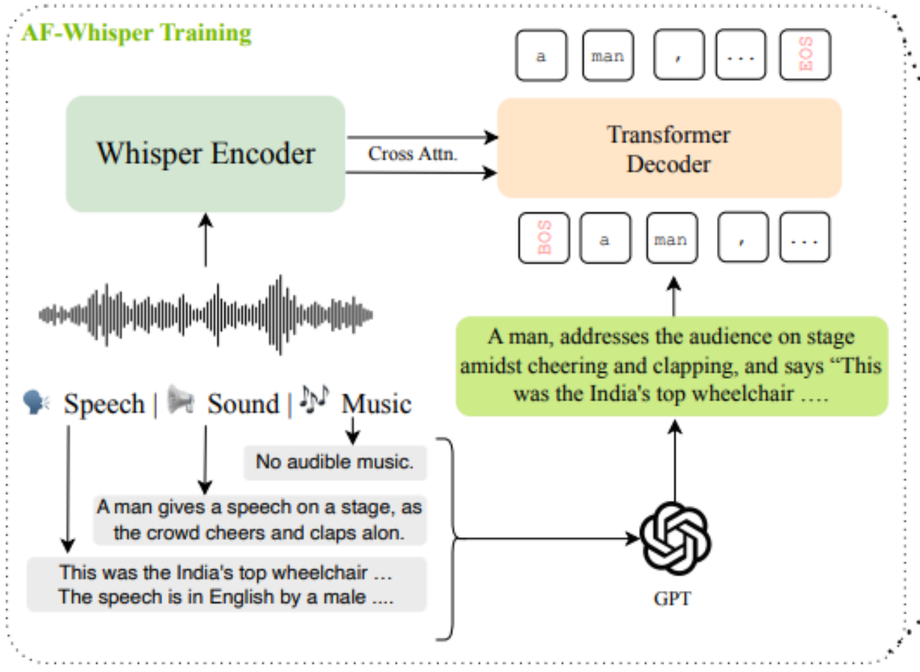
Question: Based on the audio, in which era is this story most likely to have taken place?

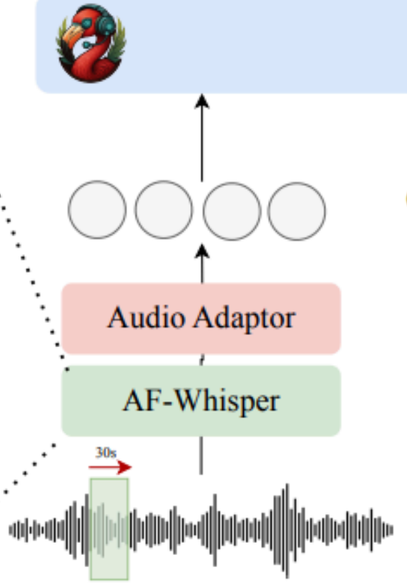
A. In the 1970s, entertainment activities were extremely scarce.
B. In the 1990s or early 21st century, society was undergoing rapid changes.
C. In the 2020s, instant messaging and social media were highly developed.
D. In a future era, people will once again be enthusiastic about physical amusement parks

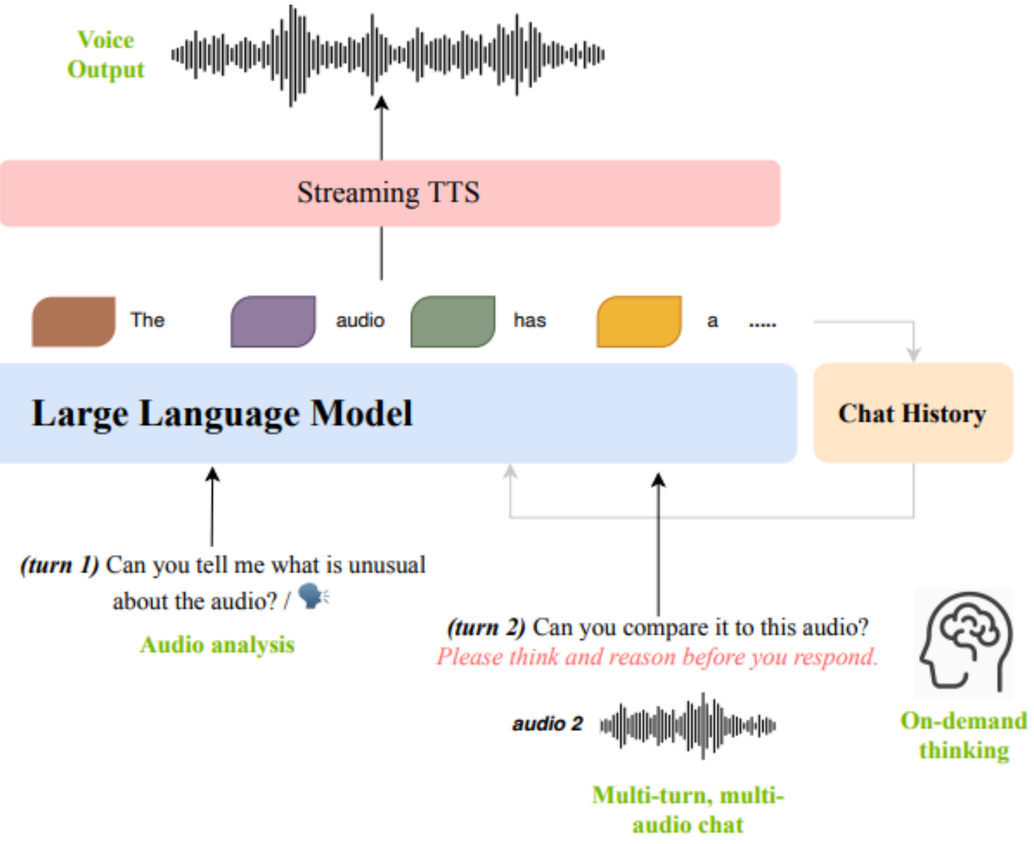


Answer: B In the 1990s or early 21st century, society was undergoing rapid changes


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
A man, addresses the audience on stage amidst cheering and clapping, and says "This was the India's top wheelchair ….
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 ....
GPT


Audio Adaptor
AF-Whisper
30s


Voice Output
Streaming TTS
The
audio
has
a
.....
Large Language Model
Chat History
(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

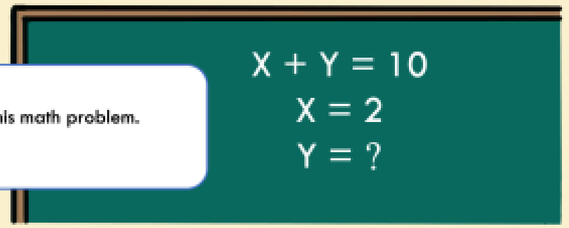
Video-Chat

Query:
What did the person on the left and the person on the right say, respectively?

Response:

(The person on the left says, "Welcome everyone to the show."
The person on the right says,
"I'm very happy that everyone can watch our show.")

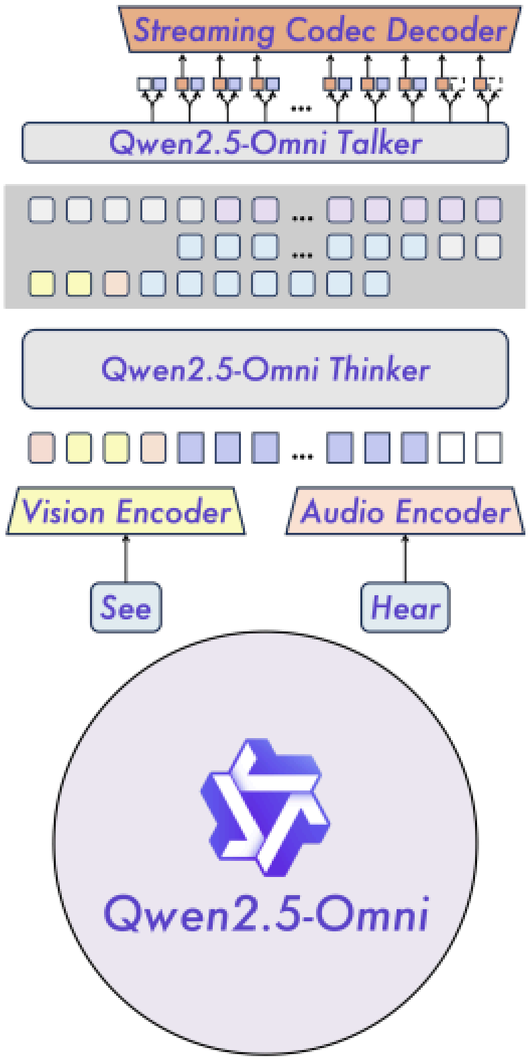
Image-Chat



Query:
Please help me solve this math problem.

Response:

(Well, since X + Y = 10 and X = 2, you can just substitute the value of X into the first equation. So it becomes 2 + Y = 10. Then, if you subtract 2 from both sides, you get Y = 8.So, Y equals 8.)



Text-Chat

Query:
Mother's Day is coming soon. Can you help me write a message for my mom?

Response:

("Mom, Mother's Day is almost here, and I just wanted to tell you how much you mean to me. You've always been there for me, no matter what. Your love has shaped who I am today. Thank you for everything you've done for me. I hope this Mother's Day brings you so much joy and happiness.")

Audio-Chat

Query:
Please provide a brief description of this music.

Response:

(it's a pop song in A major with a 4/4 time signature. The chord progression mainly alternates between A major and D major. It has a tempo of about 90 BPM)