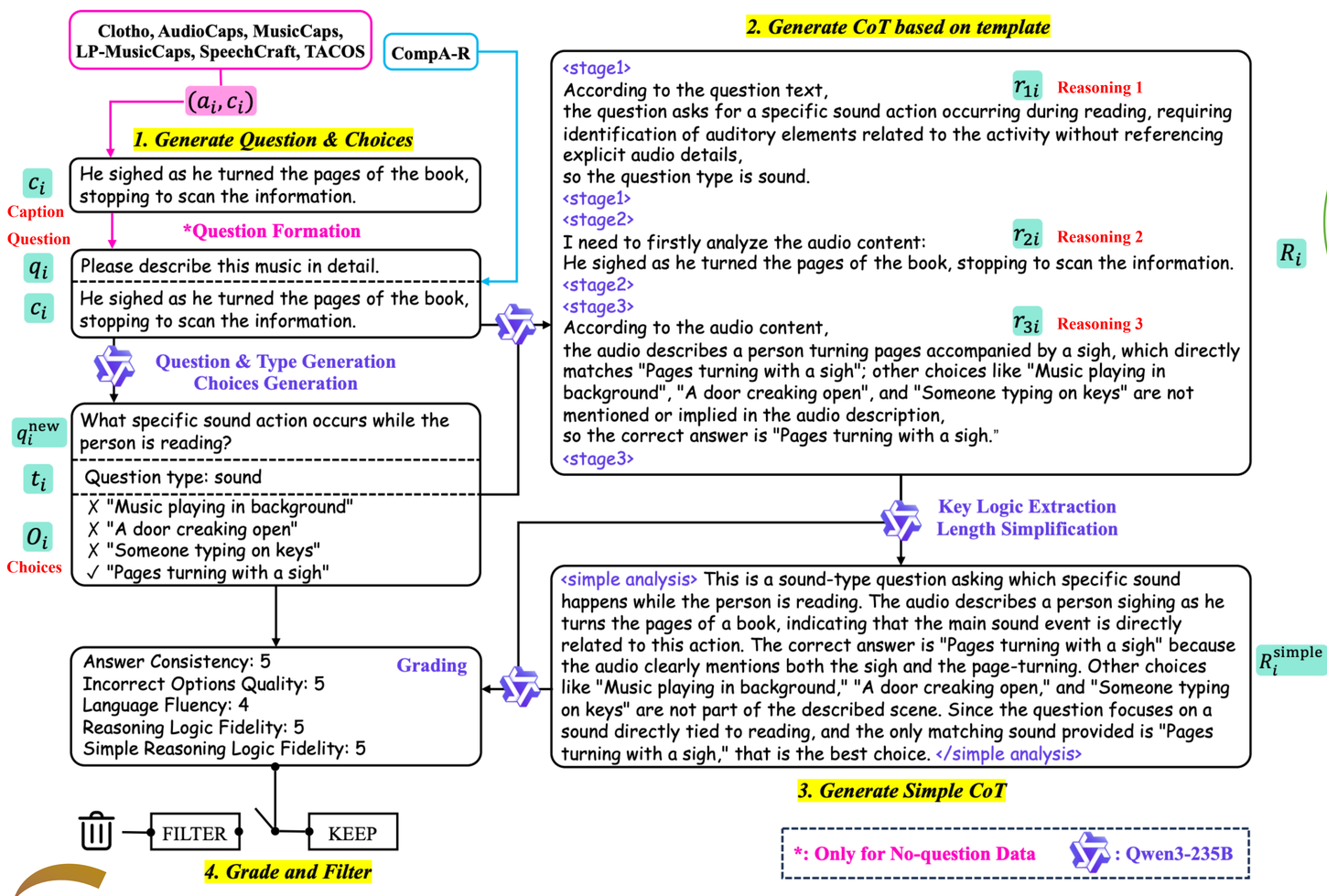
Measuring Audio's Impact on Correctness: Audio-Contribution-Aware Post-Training of Large Audio Language Models

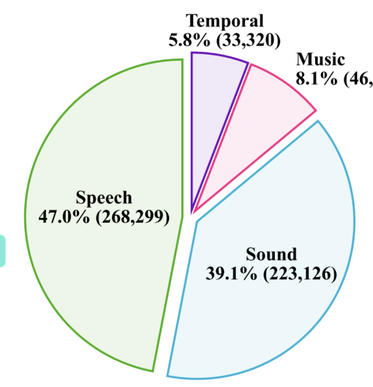
Haolin He[1,3,*], Xingjian Du[2,*], Renhe Sun[3,*], Zheqi Dai[1], Yujia Xiao[1], Mingru Yang[3], Jiayi Zhou[3], Xiquan Li[4], Zhengxi Liu[1], Zining Liang[1], Chunyat Wu[1], Qianhua He[5], Tan Lee[1], Xie Chen[4], Wei-Long Zheng[4], Weiqiang Wang[3], Mark Plumbley[6], Jian Liu[3,†], Qiuqiang Kong[1,†]

[1]The Chinese University of Hong Kong, Hong Kong, China [2]University of Rochester, USA
[3]Ant Group, China [4]Shanghai Jiao Tong University, China
[5]South China University of Technology, China [6]King's College London, UK
rex.lj@antgroup.com qqkong@ee.cuhk.edu.hk

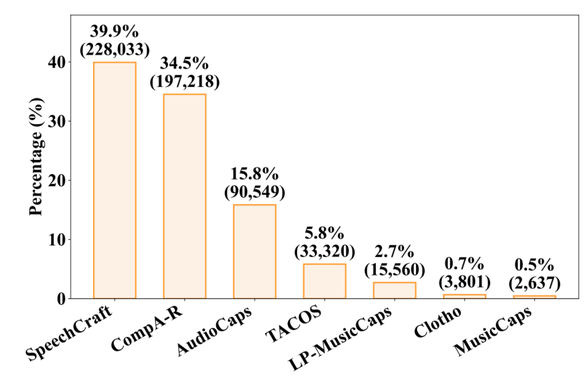
Dataset: https://huggingface.co/datasets/inclusionAI/AudioMCQ

Introducing AudioMCQ Dataset (for Post-training Research of Large Audio Language Models)





(a) Question type distribution.



(b) Source data distribution.

Key Highlights

- **571k high-quality samples** across sound, music, speech, and temporal domains
- **Dual CoT annotations**: Structured and unstructured reasoning paths
- **Audio-Contribution filtering**: Weak (54.8%) and strong (45.2%) splits
- **Pre-trained models available**: Weak-to-Strong and Mixed-to-Strong paradigms

Introducing Zero Audio-Contribution

Table 2: Performance breakdown of LALMs across audio benchmarks **with silent audio input**.

MMAU-test-mini Performance by Subset (%)							
Subset	**Qwen2-Audio**	**A-Flamingo2**	**R1-AQA**	**Kimi-Audio**	**Qwen2.5-Omni**	**Average**	**Random Guess**
Sound	42.0	56.5	56.2	67.6	51.7	**54.8**	25.0
Music	39.8	62.3	49.7	57.2	50.9	**52.0**	25.0
Speech	34.5	41.4	44.1	50.5	42.9	**42.7**	26.7
Overall	**38.8**	**53.4**	**50.0**	**58.4**	**48.5**	**49.8**	**25.5**
MMAR Performance by Subset (%)							
Subset	**Qwen2-Audio**	**A-Flamingo2**	**R1-AQA**	**Kimi-Audio**	**Qwen2.5-Omni**	**Average**	**Random Guess**
Perception	30.9	30.5	36.1	39.1	27.7	**32.9**	27.2
Semantic	34.5	39.6	37.6	40.5	35.4	**37.5**	31.4
Signal	41.9	44.2	37.2	53.5	48.8	**45.1**	33.0
Cultural	32.6	31.9	30.5	40.4	31.9	**33.5**	28.4
Overall	**33.1**	**35.0**	**36.0**	**46.5**	**32.4**	**36.6**	**29.3**
MMSU Performance by Subset (%)							
Subset	**Qwen2-Audio**	**A-Flamingo2**	**R1-AQA**	**Kimi-Audio**	**Qwen2.5-Omni**	**Average**	**Random Guess**
Perception	30.1	28.3	42.0	29.4	26.0	**31.2**	25.0
Reasoning	40.2	43.7	43.3	53.4	42.8	**44.7**	25.0
Overall	**35.3**	**35.8**	**42.7**	**41.0**	**34.1**	**37.8**	**25.0**

Zero audio-contribution refers to the phenomenon where a Large Audio Language Model (LALM) **can still correctly answer a question** even when the original audio is replaced with silent input.

Audio-Contribution Filtering

Table 3: Performance of LALMs on different audio datasets **with silent audio input** and the audio-contribution split ratios of these datasets. AC refers to audio-contribution.

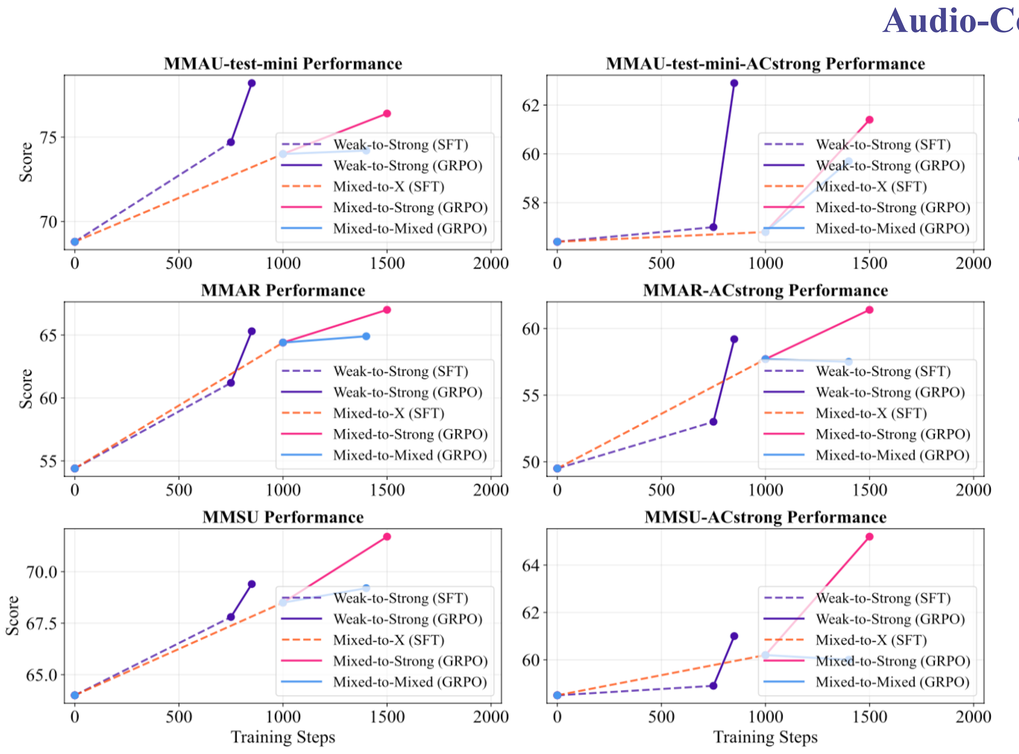
		Model Accuracy w/o Audio (%)			AC Split Ratio(%)	
Source Dataset	**Samples**	**A-Flamingo2**	**R1-AQA**	**Kimi-Audio**	**Weak**	**Strong**
Clotho	3,801	44.9	40.7	58.3	47.4	52.6
AudioCaps	90,549	41.7	38.2	59.0	44.9	55.1
CompA-R	197,218	69.8	64.7	81.6	75.5	24.5
MusicCaps	2,637	48.7	41.5	54.6	46.8	53.2
LP-MusicCaps	15,560	47.5	41.9	60.6	49.2	50.8
SpeechCraft	228,033	35.0	47.1	58.4	45.6	54.4
TACOS	33,320	31.2	34.1	35.4	26.7	73.3
Overall	**571,118**	**48.3**	**50.8**	**65.2**	**54.8**	**45.2**

Table 4: Performance of LALMs on different audio understanding benchmarks **with silent audio input** and the audio-contribution split ratios of these benchmarks. AC refers to audio-contribution.

		Model Accuracy (%)			AC Split Ratio(%)	
Source Dataset	**Samples**	**A-Flamingo2**	**R1-AQA**	**Kimi-Audio**	**Weak**	**Strong**
MMAU-test-mini	1000	53.4	50.0	58.4	53.9	46.1
MMAR	1000	35.0	36.0	40.5	32.9	67.1
MMSU	5000	35.8	42.7	41.0	35.7	64.3

Specifically, **multiple LALMs are evaluated using silent audio input;** if most models can still answer a question correctly, the sample is labeled as ***weak audio-contribution***, otherwise it is labeled as ***strong audio-contribution***. This process distinguishes text-solvable questions from those that genuinely require audio understanding.

Audio-Contribution-Aware Post-Training



Two Training Paradigms

- **Weak-to-Strong**: SFT on weak audio-contribution data, followed by RL on strong audio-contribution data.
- **Mixed-to-Strong**: SFT on mixed data, followed by RL on strong audio-contribution data.

Table 5: Performance comparison across different models.[2]

Method	**MMAU-test-mini**	**MMAU**	**MMAR**	**MMSU**
Audio-Reasoner	67.7	63.8	36.8	49.2
R1-AQA	68.9	68.5	50.8	61.6
Kimi-Audio	68.2	64.4	57.6	59.3
SARI	67.0	–	–	66.0
Qwen2.5-Omni (backbone)	71.5	71.0	56.7	60.6
Audio Flamingo 3	73.3	72.4	60.1	62.3
Omni-R1	77.0	75.0	63.4	–
Audio-Thinker	78.0	75.4	65.3	–
GPT4o-Audio	62.5	60.8	63.5	56.4
Gemini-2.0-Flash	70.5	67.0	65.6	51.0
Our Methods				
All Data SFT	75.2	75.0	64.6	64.0
All Data GRPO	78.1	75.4	63.0	70.2
Mix AC SFT + Mix AC GRPO	74.2	74.4	64.9	69.2
Weak AC SFT + Strong AC GRPO	**78.2**	**75.6**	65.3	69.3
Mix AC SFT + Strong AC GRPO	76.4	75.1	**67.0**	**71.7**

The baseline Mixed-to-Mixed approach **demonstrating suboptimal MMAU performance**.

In contrast, the Weak-to-Strong paradigm **surpasses the previous 1200-step GRPO-only approach**, achieving SOTA performance on MMAU-test-mini and 75.6% on MMAU.

The Mixed-to-Strong approach attains **SOTA performance on MMAR and MMSU**.

TAKE AWAY

- **Existing LALMs suffer from zero audio-contribution, meaning they can answer correctly even without audio.**
- **First, using strong audio-contribution data for RL is important.**
- **Second, SFT data selection should align with specific downstream task characteristics.**