Echo: Towards Advanced Audio Comprehension via Audio-Interleaved Reasoning

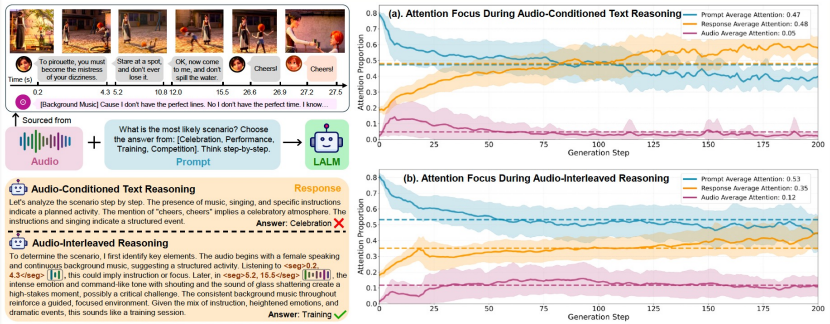
Daiqing Wu[1,7] · Xuan Zhang[3] · Dongbao Yang[1] · Jiashu Yao[3] · Longfei Chen[5] · Qingsong Liu[3] · Sicheng Zhao[6] · Can Ma[1] · Yangyang Kang[2,3] · Yu Zhou[4]

[1] IIE, Chinese Academy of Sciences, [2] Zhejiang University, [3] ByteDance China, [4] Nankai University, [5] ShanghaiTech University, [6] Tsinghua University, [7] University of Chinese Academy of Sciences

MOTIVATION

The Bottleneck in Audio-Conditioned Text Reasoning

- Rely on **one-time encoding** of audio content, which forces all reasoning to proceed from a static representation.
- Risk losing subtle details after the initial pass, creating a fundamental information bottleneck.



Audio-Interleaved Reasoning

- Treats audio as **active reasoning components**.
- Enables sustained audio engagement and perception-grounded analysis through **dynamic re-listening**, emulating human auditory cognition.

Main Contributions

1. Propose audio-interleaved reasoning, a new paradigm for LALMs.
2. Present Echo, a LALM with dynamic segment re-listening.
3. SOTA on 3 benchmarks, surpassing GPT-4o & Gemini-2.0-Flash.

TRAINING FRAMEWORK

Two-Stage Training: SFT → RL

Stage 1 — SFT

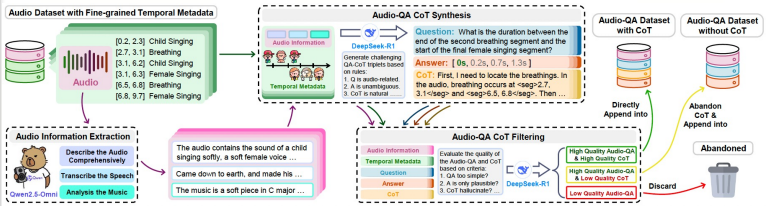
- Teaches the model to localize and reference salient audio segments via <seg> tags with CoT annotations.

Stage 2 — RL

- Incentivizes proficient and accurate re-listening behavior using Group Relative Policy Optimization.

DATA PIPELINE

Structured Data Generation Pipeline



- Qwen2.5-Omni captions → DeepSeek-R1 QA-CoT synthesis.
- Quality filtering by DeepSeek-R1 under detailed criteria.
- EAQA-SFT: 75.9k w/ CoT | EAQA-RL: 21.9k w/o CoT.

MAIN RESULTS

State-of-the-Art on 3 Audio Benchmarks

MMAR (Expert-Level Reasoning, 1k tasks)

- **Echo: 69.99%** (Best in Music, Sound+Music & Sound+Speech)
- Surpasses Audio-Thinker (67.25%) and Gemini-2.0-Flash (67.90%).

MMAU-mini / MMAU (General Audio, 1k/9k tasks)

- **Echo: 80.41% / 76.61%** (Best except for Sound, Music of MMAU)
- Surpasses Audio-Thinker (75.39%) and Step-Audio-2 (73.86%).

ANALYSIS

Training Framework Effectiveness & Skill Evolution

	Model	SFT Data	RL Data	Reasoning Format	MMAR (Avg)		
					Acc (%)	Length (word)	Latency (s)
A	Base Model	-	-	Audio-Conditioned Text	51.80	67.95	1.18
B	Cold-Start Model	EAQA-SFT	-	Audio-Grounded	56.77	117.06	2.04
B'	Unadpted RL Model	EAQA-SFT	EAQA-RL	Audio-Grounded	64.63	98.12	1.97
C	Cold-Start Model	EAQA-SFT	-	Audio-Interleaved	52.26	101.73	2.16
D	Echo	EAQA-SFT	EAQA-RL	Audio-Interleaved	69.99	107.40	2.12
D'	Echo-AVQA	EAQA-SFT	AVQA	Audio-Interleaved	67.58	120.18	2.37
E	Direct RL Model	-	EAQA-RL	Audio-Conditioned Text	63.15	104.42	2.06
E'	Direct RL Model	-	AVQA	Audio-Conditioned Text	59.51	111.14	2.11

- **Training Effectiveness (A→B→D):** SFT +4.97% acc → RL peaks at 69.99%. Controlled latency overhead (2.12s vs 1.18s baseline).
- **Audio Involvement (E→B'→D) :** Accuracy improves monotonically with audio engagement, confirming intrinsic benefit.
- **Data Quality (D→D', E→E'):** EAQA-RL (21.9k) outperforms 40.4k AVQA samples, validating the quality of constructed data.
- **Broad Skill Gains:** +37% multi-speaker mapping, +20.8% event-based sound reasoning, +20.5 emotion state summarization.