CESAR: Consistent, Effective and Scalable Audio Reasoners

Incentivizing Reasoning Capability in Audio LLMs via Process Rewards

Jiajun Fan[1], Roger Ren[2], Jingyuan Li[2], Rahul Pandey[2], P.G. Shivakumar[2], I. Bulyko[2], A. Gandhe[2], Ge Liu[1], Yile Gu[2]

[1]UIUC [2]Amazon

ICLR 2026

MMAU Test-Mini: 77.10% (#1 Open-Source) | **MMAU-Pro: 56.4% (#1 Open-Source)**

The Problem: Reasoning Makes Audio LLMs *Worse*

Test-Time Inverse Scaling (Our Finding)

Chain-of-thought **degrades** Audio LLM accuracy.
Longer chains ⇒ progressively worse results.

Root Cause: Four Failure Modes

- × Factual hallucination in reasoning trace
- × Flawed causal / logical inference
- × Reasoning–answer inconsistency
- × Redundant overthinking

Root cause: Not reasoning itself —
outcome-only RLVR never supervises *how* to reason.

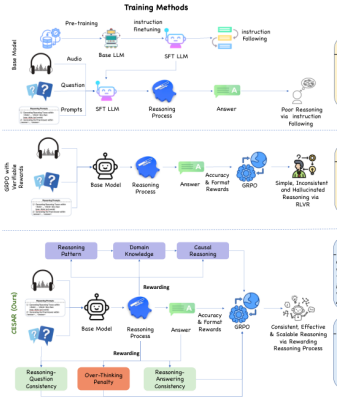
Hallucination: Q: Source of speech? (man/woman. . .)
Reasoning: *"the voice is male. Options don't include 'man'. . . "*
Answer: robot ×

Reasoning–Answer Mismatch: Q: How many times does phone ring?
Reasoning: *"rings three times. . . "*
Answer: 2 ×

CESAR fixes both: grounds reasoning in audio evidence,
ensures conclusion matches stated answer. ✓

Takeaway: Outcome-only rewards teach *what* to answer, not *how* to reason. CESAR targets the reasoning process directly.

CESAR: Shifting from Outcome to Process Rewards



GRPO + Multi-Faceted Process Reward Suite

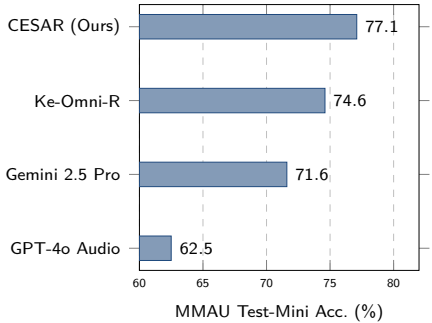
$$R_{\text{total}} = \alpha_1 R_{\text{acc}} + \alpha_2 R_{\text{fmt}} + \alpha_3 R_{\text{con}} + \alpha_4 R_{\text{kw}} + \alpha_5 R_{\text{OT}}$$

Reward	What It Incentivizes
$R_{\text{acc}} + R_{\text{fmt}}$	Correctness and proper XML reasoning format
R_{con} (Consistency)	Semantic alignment: reasoning $\leftrightarrow$ answer $\leftrightarrow$ question
R_{pattern} (Keywords)	Structured patterns: elimination, comparison, causal chains
R_{domain} (Keywords)	Audio domain terminology and knowledge integration
R_{OT} (Overthinking)	**Penalty** for redundant/circular reasoning

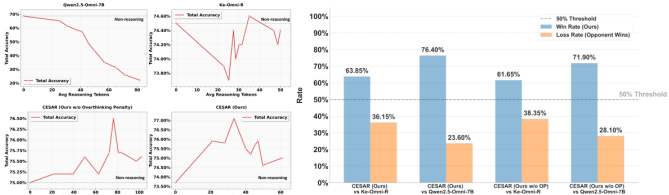
Key: All rewards are **rule-based** (no LLM judge). Applies to any GRPO-trained audio model — fully **scalable**.

Takeaway: CESAR = GRPO + 5 process rewards. Each reward targets a specific failure mode of outcome-only training.

Results: SOTA + Test-Time Scaling Unlocked



+2.5pp vs. Ke-Omni-R **+5.5pp** vs. Gemini 2.5 Pro



Test-Time Scaling: Inverse Scaling ⇒ Sweet Spot

Before CESAR: reasoning degrades accuracy (inverse scaling).
After CESAR: peaks at ∼35–40 tokens — **77.1%** with reasoning.
MMSU Semantic Reasoning: 88.72% vs. human 82.16% (super-human).

Takeaway: CESAR resolves test-time inverse scaling: reasoning becomes a reliable, scalable asset.

What Better Reasoning Looks Like

Before CESAR (Ke-Omni-R)

Q: How many times does the telephone ring?
Thinking: *"rings three times..."* **Answer: 2** ×
reasoning–answer mismatch

Q: What is the chord from 0:02–0:03?
Thinking: *"progression... progression... progression..."*
× Redundant overthinking, wrong answer

After CESAR (Ours)

Q: How many times does the telephone ring?
Thinking: *"clearly three distinct rings."*
Answer: 3 ✓ Consistent and grounded

Q: What is the chord from 0:02–0:03?
Thinking: *"A:min/P5 fits — minor chord, the others don't."*
✓ Concise, domain-aware, correct

Three Key Contributions

1. **Identify** test-time inverse scaling in Audio LLMs and diagnose its root cause as inadequate process training.
2. **CESAR**: process reward suite that resolves inverse scaling and enables test-time scaling gains.
3. **SOTA**: 77.1% MMAU Test-Mini; near-/super-human MMSU semantic reasoning.

Takeaway: Process rewards turn reasoning from a liability into an asset — framework generalizes to any GRPO-based audio model.

Thank You!

CESAR — Consistent, Effective and Scalable Audio Reasoners

MMAU Test-Mini	**77.10%**	#1 open-source (+5.5pp vs Gemini 2.5 Pro)
MMAU-Pro	**56.4%**	#1 open-source
MMSU Semantic	**88.72%**	vs. human 82.16% (super-human)
Test-time scaling	Inverse scaling resolved; sweet spot at ∼35–40 tokens	

jiajunf3@illinois.edu | ICLR 2026

Questions?