NEURIPS 2026

RAIL

Rethinking Auditory Intelligence in Large Audio-Language Models with a CHC-Grounded Benchmark

Hongyu Jin[1]*, Siyi Wang[1]*, Yang Xiao[1]*, Jiaheng Dong[1]*, Shihong Tan[4], Kaiyuan Peng[1], Georgiana Juravle[2], Shanquan Chen[3], Gongping Huang[4], Hong Jia[5], Eun-Jung Holden[1], James Bailey[6], Ting Dang[1]†

[1]The University of Melbourne · [2]Alexandru Ioan Cuza University of Iași · [3]The University of Hong Kong
[4]Wuhan University · [5]The University of Auckland · [6]Monash University

*Equal contribution · †Corresponding author

MOTIVATION

Imagine walking into a busy café

Perception

Pick out the voice you care about from clattering cups, a hissing espresso machine and music.

Reasoning

Work out who is talking to whom, and why.

Memory

Recognise a voice as the colleague you met last week.

Humans do all three at once, without effort: a natural yardstick for audio AI.

THE PROBLEM

Audio benchmarks score tasks, not abilities

01 **No cognitive theory behind task design**
Tasks such as transcription or captioning are defined by datasets, not by abilities.

02 **Data skewed to task-specific formats**
Real listening means, e.g., tracking a list as it is updated across a conversation.

03 **No human-aligned evaluation**
End-task accuracy can't tell which ability failed, or how a model compares to people.

When a model fails, is it perception, memory, reasoning, knowledge or speed?

OUR APPROACH

We ground audio evaluation in CHC theory

NOT CHOSEN

Bloom's taxonomy

Grades the difficulty of educational tasks, not the abilities underneath.

NOT CHOSEN

Multiple intelligences

Gardner's theory lacks standardised measurement and empirical validation.

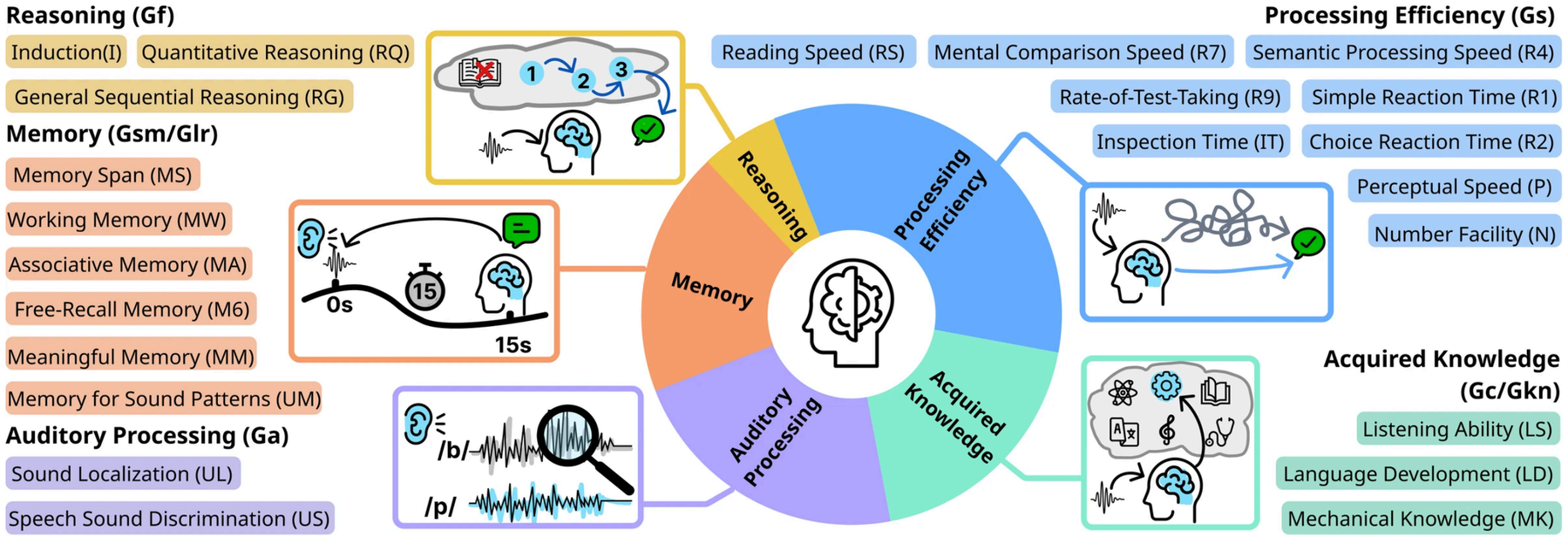
CHOSEN

Cattell–Horn–Carroll

Data-driven and measurable. Broad abilities split into narrow ones, so each can be tested on its own.

Framework chosen together with cognitive scientists, as stage 1 of the RAIL pipeline.

Five capabilities, 32 subcapabilities



Auditory · 7 | Reasoning · 3 | Memory · 6 | Efficiency · 9 | Knowledge · 7 = 32 subcapabilities

TASK DESIGN

Every answer must come from the audio

Two principles: **auditory dependence**, so answers rest on audio cues, and **capability independence**, so no ability is a prerequisite for measuring another.

Ga · Phonetic coding

Audio: two overlapping phonemes, slightly offset

Q: Which two phonemes are present?

A: /æ/ as in “cat” and /uː/ as in “boot”

Gf · Sequential reasoning

Audio: cough · 1 s silence · sneeze

Rules: cough = swap first and last; sneeze = sort A→Z

A: Mango, Apple, Cherry
→ Cherry, Apple, Mango
→ Apple, Cherry, Mango

Gsm · Working memory

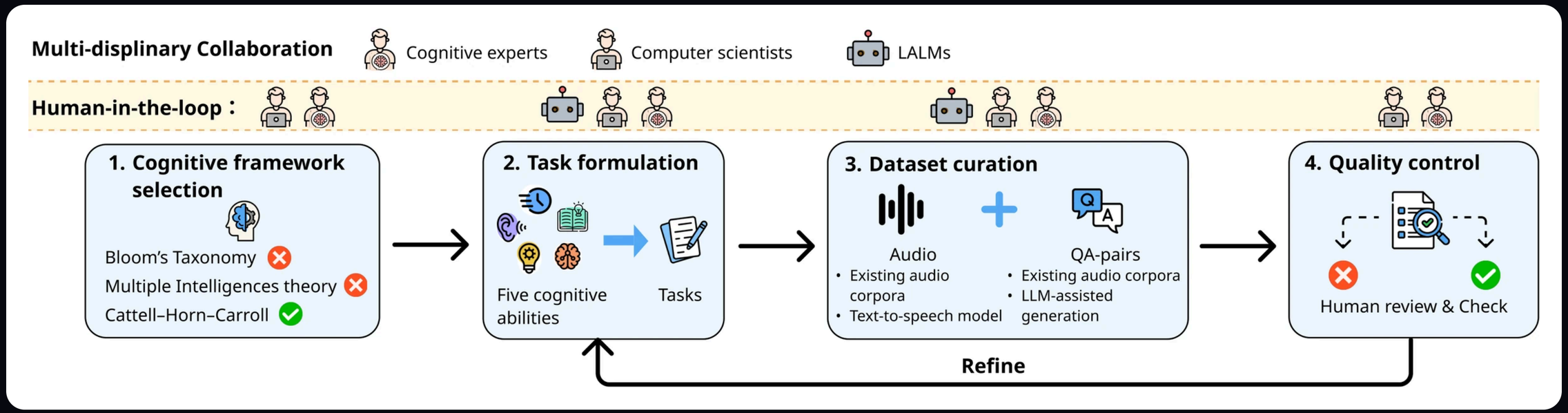
Audio: a chat that adds, drops and reorders shopping-list items

Q: After all updates, what is the final list?

A: spinach, eggs, olive oil, coffee beans, butter

DATA

Built with cognitive experts in the loop



5,306
audio question–answer samples

30.6 h
of speech, music and environmental sound

68.1%
newly constructed for RAIL

640
items with a human baseline from 24 listeners

EVALUATION

26 models, one protocol

26

large audio-language models

21 open-source · 5 closed-source

167M – 33.5B parameters

Speech LLMs, general audio LLMs, omni-modal models, GPT and Gemini APIs

Strict accuracy

Correct only if the final answer matches the reference exactly.

LLM-as-judge

GPT-5.4 checks semantic equivalence with the reference.

B-AUC for processing efficiency

Area under the accuracy curve as the reasoning-token budget grows: right answers with short reasoning score high.

$$\text{B-AUC} = \frac{1}{B} \cdot \sum_b \tfrac{1}{2} \left[\text{Acc}(\leq b) + \text{Acc}(\leq b+1)\right]$$

FINDING 1 · CAPABILITY PROFILE

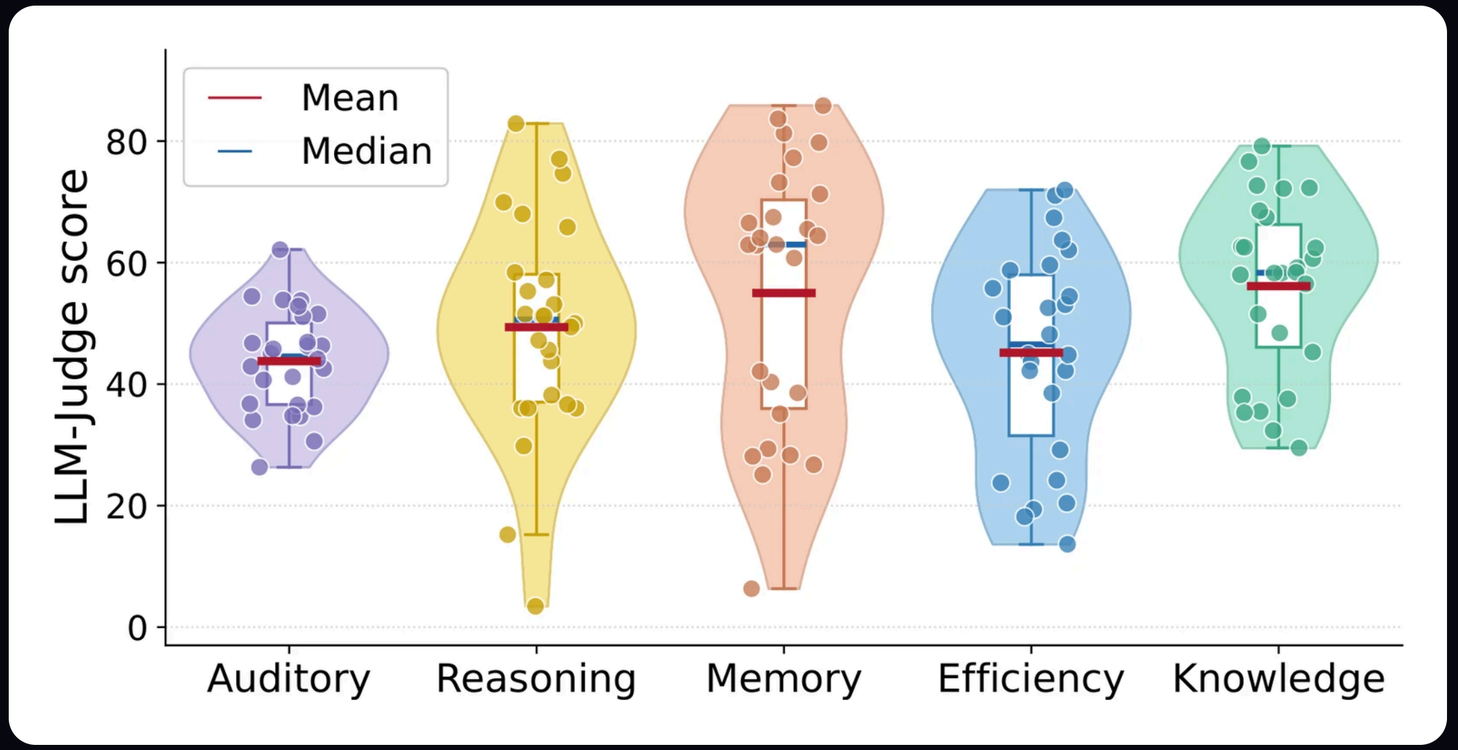
Models know more than they hear

56.21
Knowledge: highest mean

55.05
Memory: high mean, widest spread

43.83
Auditory processing: lowest mean

Mean LLM-judge score over 26 models. Closed-source 65.10 vs open-source 46.27.



FINDING 2 · AUDITORY PROCESSING

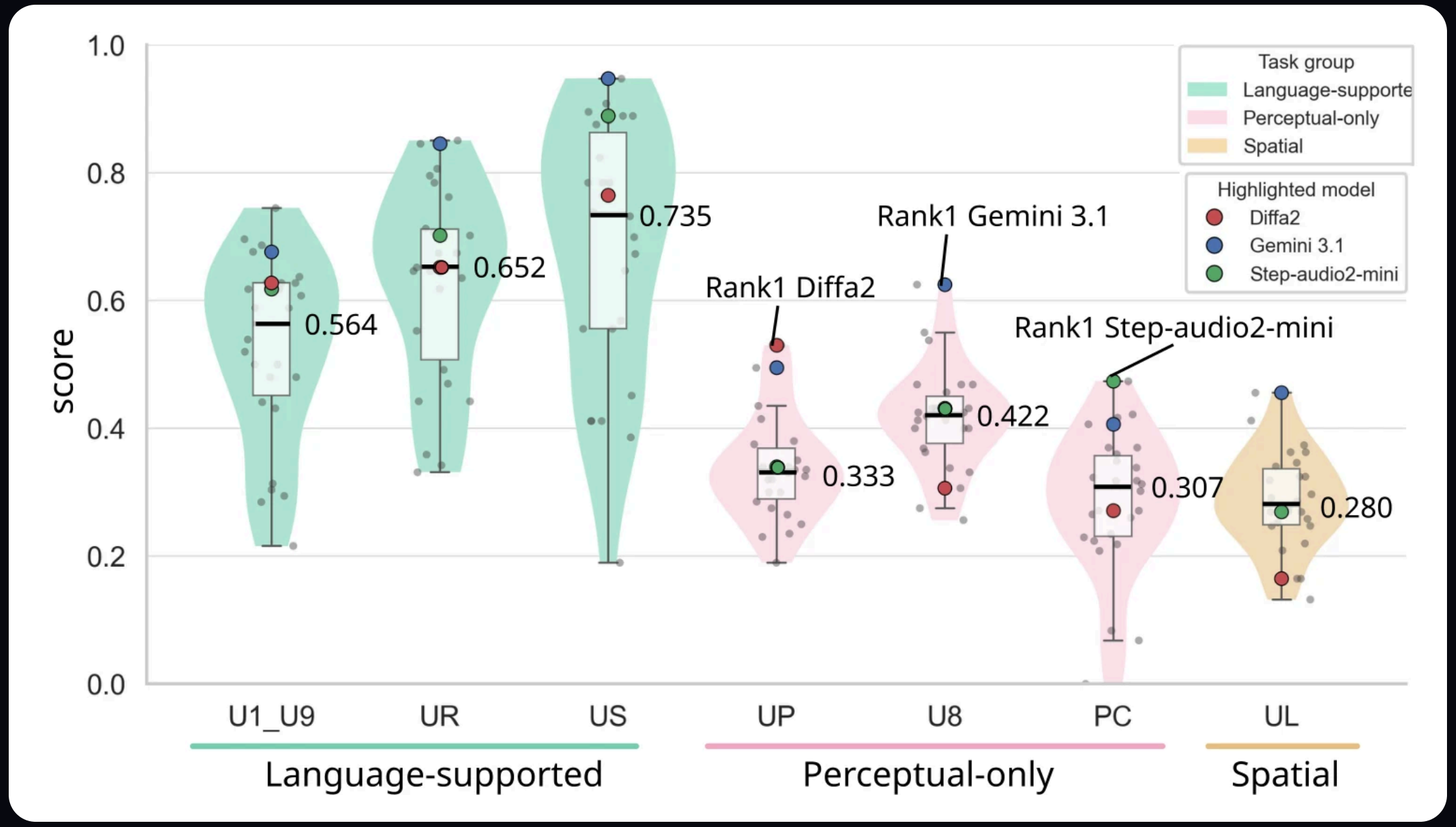
Language cues mask weak listening

0.56–0.74
Language-supported tasks

0.31–0.42
Perceptual-only: pitch, rhythm, phonemes

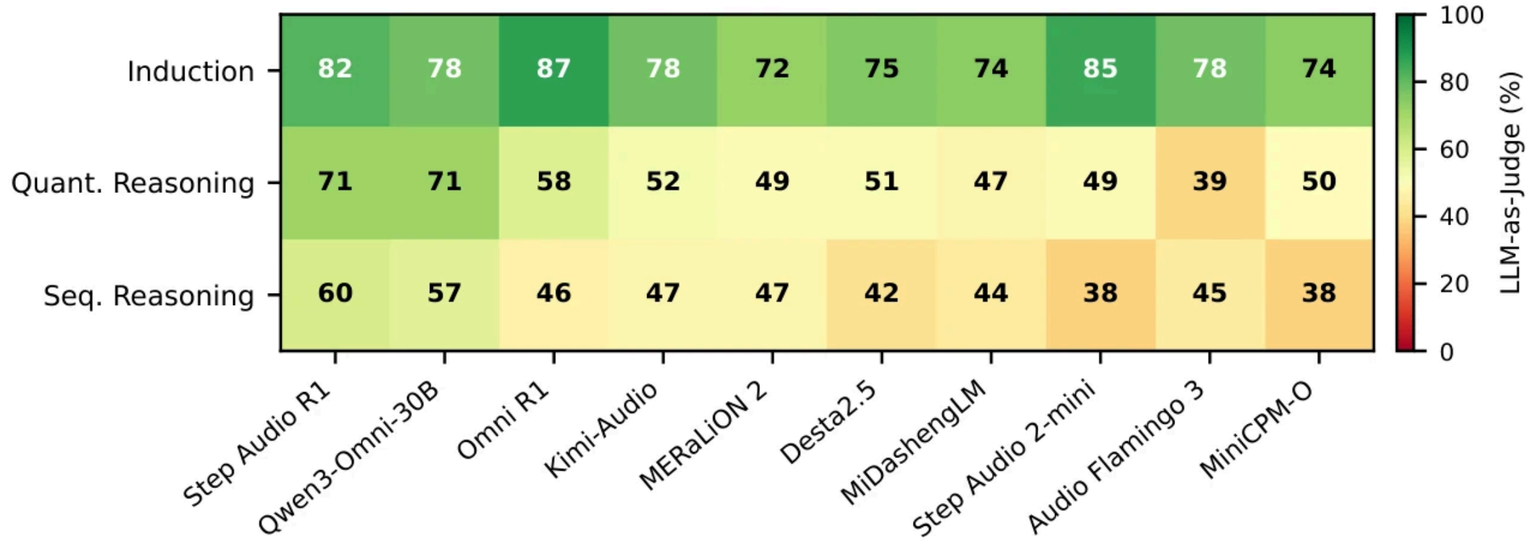
0.28
Sound localization

Median score per task. The gap holds in 18 of 26 models.



FINDING 3 · REASONING

Sequential reasoning is the weakest link



Most open-source models score below 50%: apply a rule, update the state, repeat.

Text-side reasoning training doesn't transfer: similar backbones, very different audio scores.

FINDING 4 · MEMORY

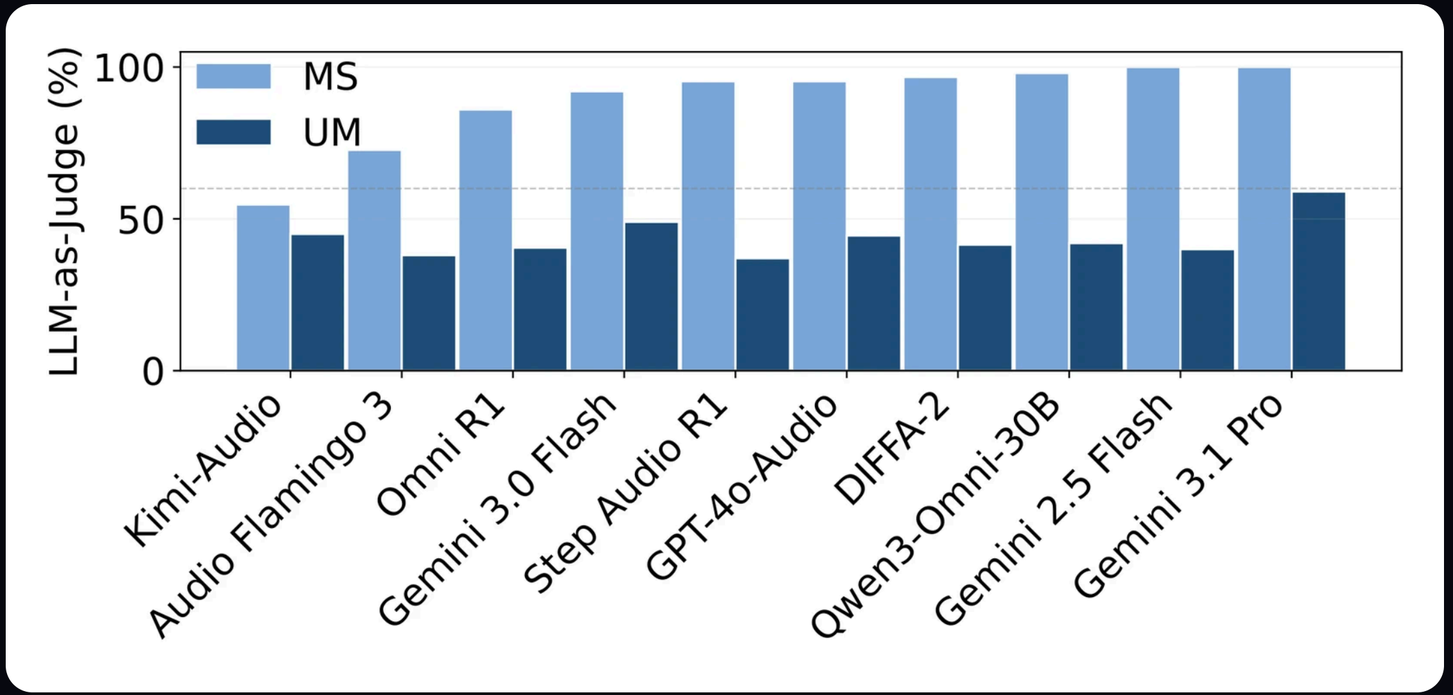
Speech memory is not sound memory

100
Memory span for speech (MS), best models

59
Best memory for sound patterns (UM): no model reaches 60

6 vs 6
Free recall: six models above 87, six below 10

LLM-judge scores (%). Sound memory covers environmental sounds and prosody.



FINDING 5 · PROCESSING EFFICIENCY

Bigger is not more efficient

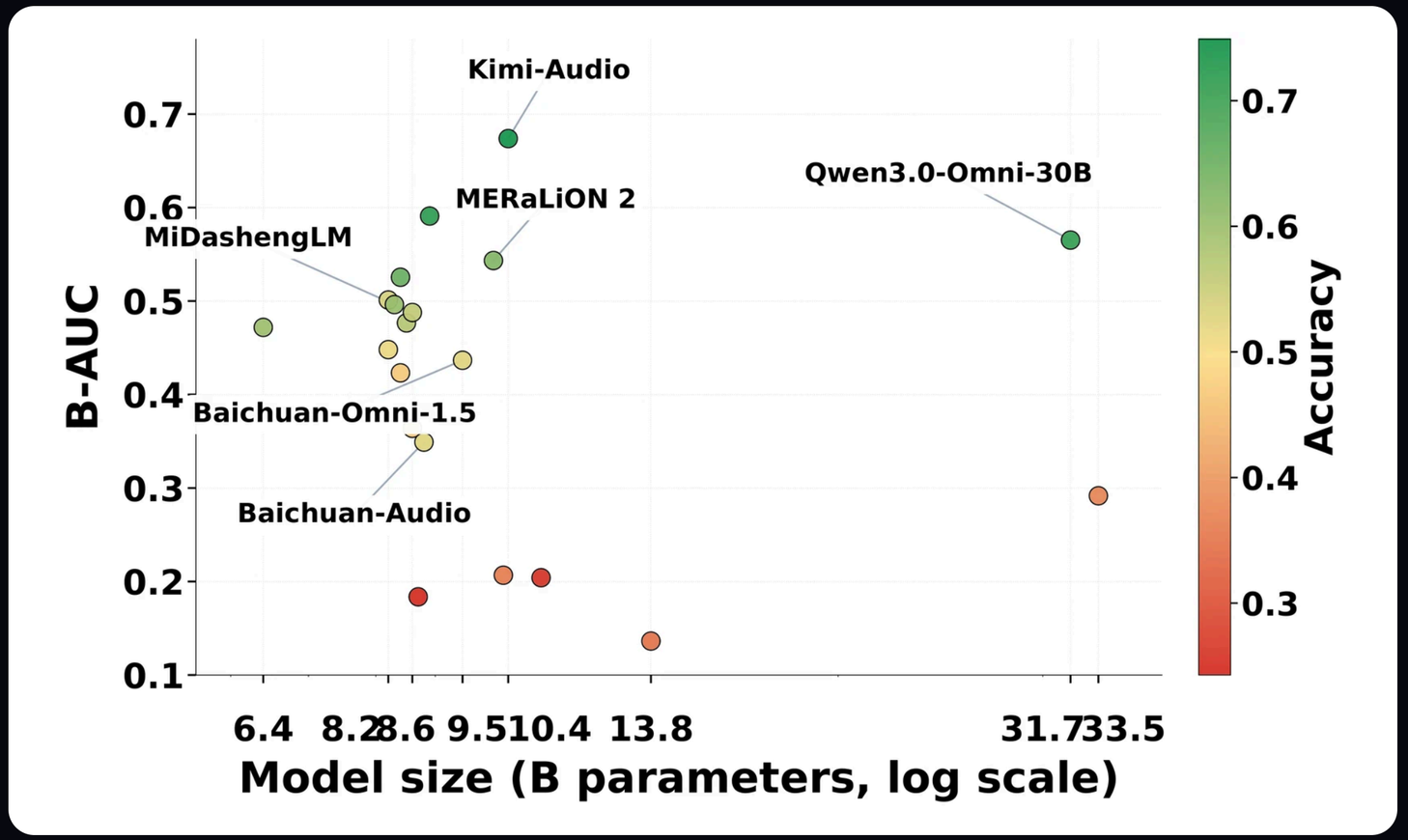
0.2–0.7

B-AUC range among models of nearly the same size (8.4–8.6B)

21

open-source models evaluated (167M–33.5B): size does not predict B-AUC

Generation behaviour drives efficiency: weakly tuned models repeat themselves, larger ones over-reason on simple inputs.

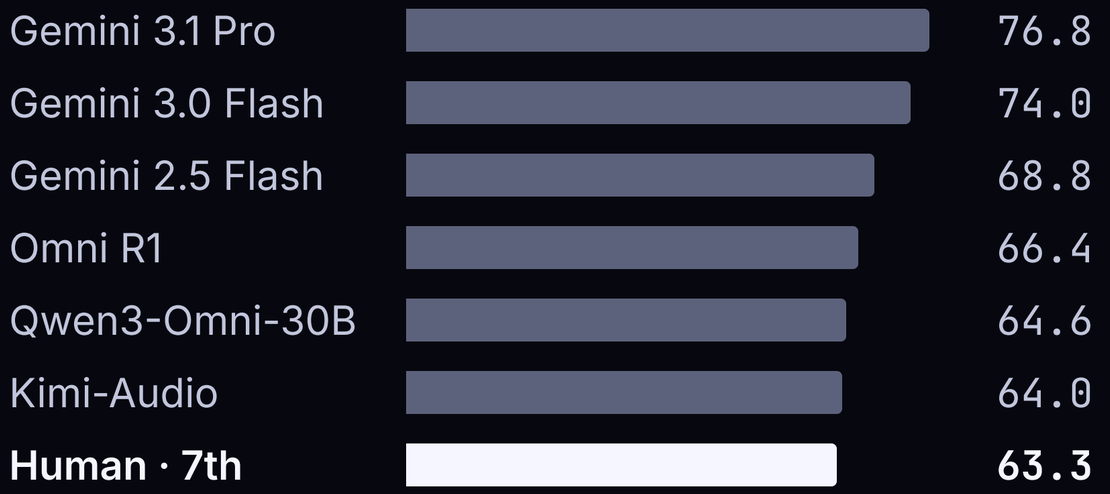


FINDING 6 · HUMAN BASELINE

Humans rank 7th overall, but 1st on listening

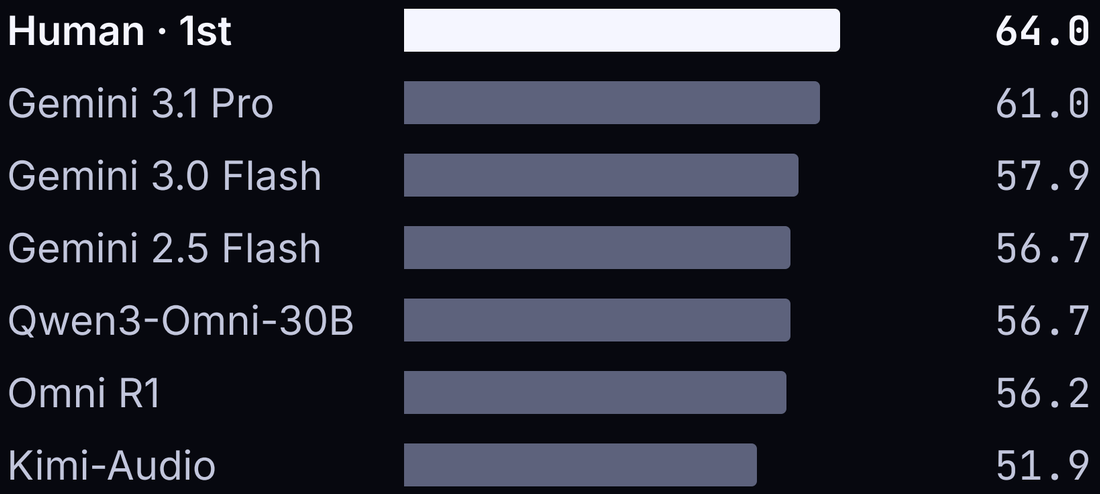
Overall

Mean of five capabilities · top six models



Auditory processing

Same six models · same 640-item subset



LLM-judge scores (%) on the same items for people and models; bars start at 0. Humans also lead processing efficiency (68.7 vs 65.5), but rank 13th on memory and 18th on reasoning out of 26.

IN ONE LINE

Knowledge is inherited; perception is not.

Audio LLMs borrow their knowledge from text pretraining.
Hearing itself still has to be learned from sound.

TAKEAWAYS

What RAIL asks of the next audio model

Auditory	Audio front-ends that keep fine temporal and spectral detail
Reasoning	Stateful, audio-grounded reasoning rather than longer chain-of-thought text
Memory	Training for non-speech, multi-turn and long-form memory
Efficiency	Objectives that adapt reasoning depth to the input
Knowledge	Expert-labelled data for under-covered sounds, such as machines

RAIL gives a capability-level profile, with people as the reference, to track this progress.



Thank you

PAPER

arxiv.org/abs/2606.11260

PROJECT PAGE

aims-rail.github.io/RAIL

DATASET

huggingface.co/datasets/AIMS-RAIL/RAIL

CODE

github.com/AIMS-RAIL/RAIL

Contact: hongyuj1@student.unimelb.edu.au · ting.dang@unimelb.edu.au