Evaluating Multimodal Large Language Models on Core Music Perception Tasks

Brandon J. Carone[1], Iran R. Roman[2], & Pablo Ripollés[1]

[1] Department of Psychology, New York University, New York; Center for Language, Music, and Emotion (CLaME), New York; Music and Audio Research Laboratory (MARL), New York

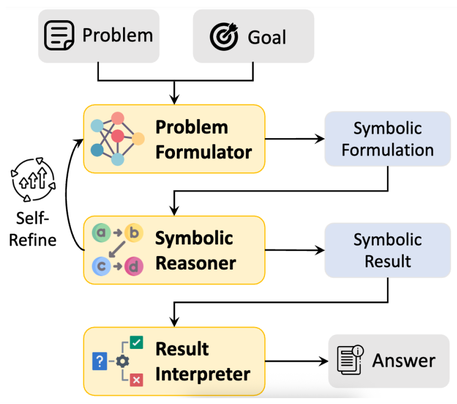
[2] Department of Electronic Engineering and Computer Science, Queen Mary University of London; Centre for Digital Music, London

Introduction

- Multimodal Large Language Models (LLMs) like Google's Gemini and Alibaba's Qwen claim "musical understanding", but their audio capabilities remain poorly characterized.
- Existing audio benchmarks for audio/music often evaluate models on tasks like classification and captioning, but this can conflate ***listening*** with superficial ***score reading.***
- Thus, the present study aimed to evaluate the structural understanding (rhythm, melody, harmony) of audio LLMs by isolating perception from reasoning.

LogicLM

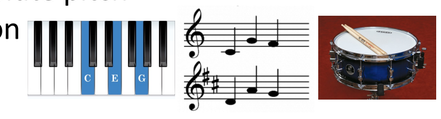
- We adapt LogicLM, forcing models to transcribe audio into machine-checkable symbolic schemas (e.g., pitch lists) before a deterministic solver calculates the answer, thereby exposing "unfaithful reasoning."



Methods

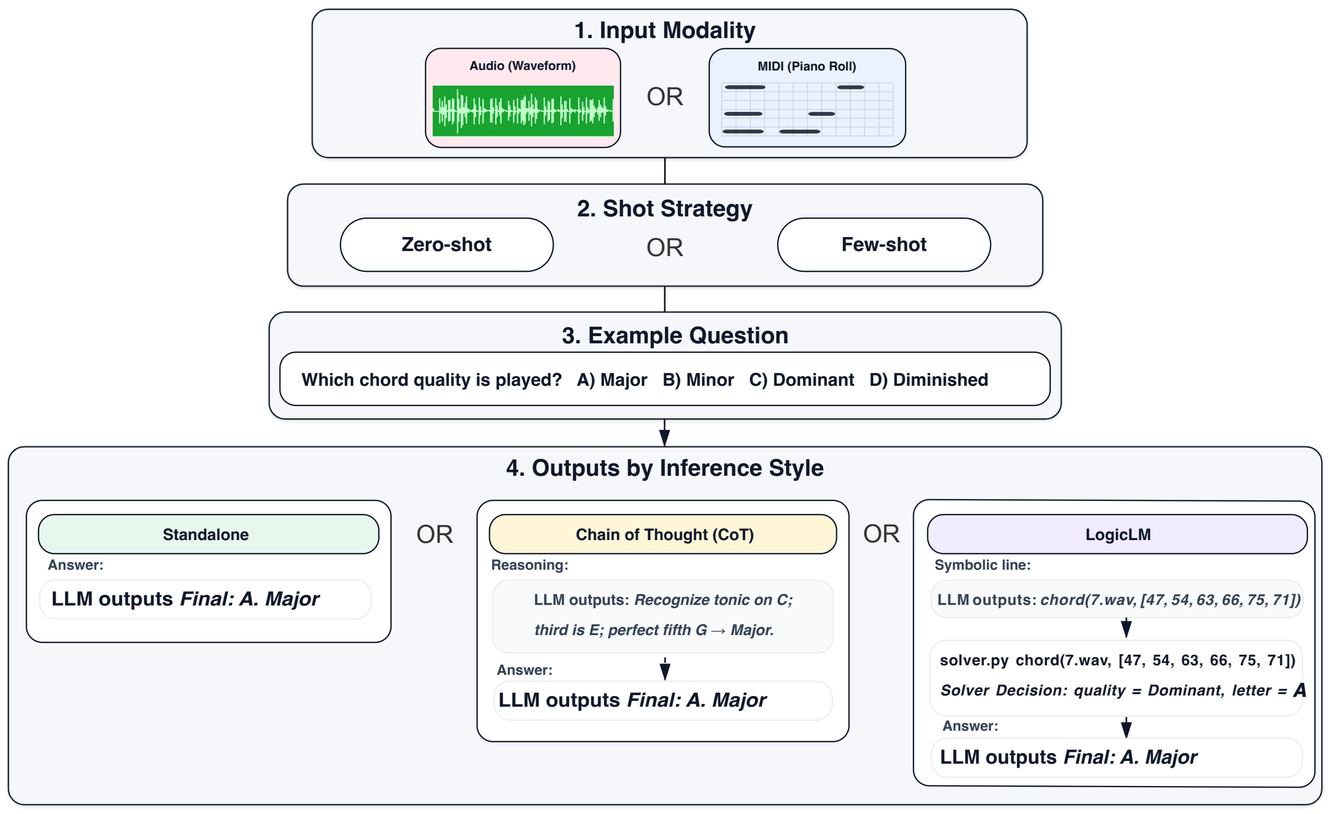
Tasks:

- ***Syncopation Scoring*** tests sensitivity to rhythmic expectancy violations and metric displacement
- ***Transposition Detection*** tests melody identification invariant to absolute pitch
- ***Chord Quality Identification*** tests musical interval pattern recognition



Stimuli:

- Audio recordings (.wav) from **the MUSE Benchmark**
- Human-played, symbolic representations of the original audio (MIDI)
- 20 stimuli for both Syncopation Scoring and Transposition Detection; 44 stimuli for Chord Quality ID



Results

Robust Modality Gap:

- Gemini models achieve near-ceiling performance on symbolic MIDI, but their accuracy drops by ~40-45 percentage points on audio inputs (see Figure A below).

Prompting Effects:

- Advanced prompting (Few-shot, CoT) fails to compensate for upstream perceptual errors, showing no significant gains over zero-shot baselines (Figure B below).

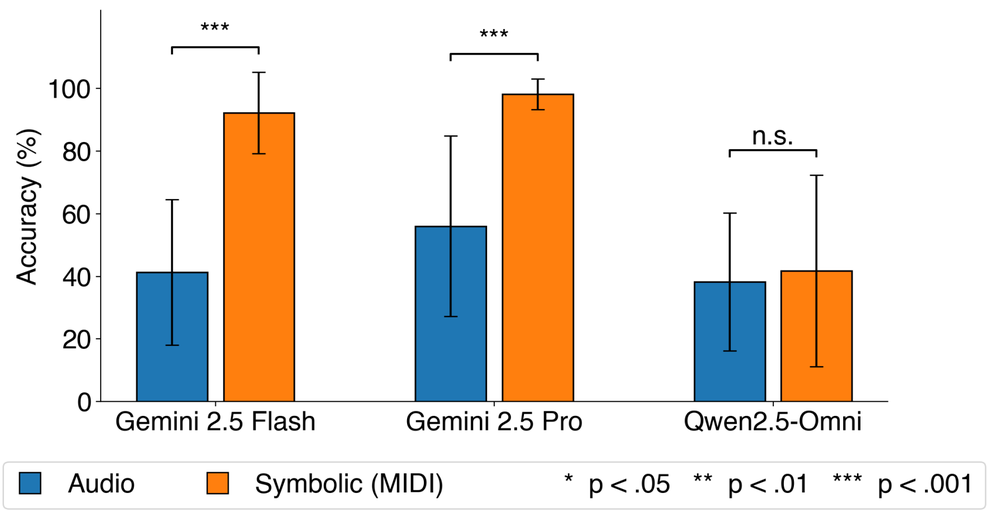
LogicLM Fragility:

- LogicLM often aids symbolic reasoning, but it severely degrades audio performance (e.g., Chord ID drops to < 10%) due to transcription errors.

Mod.	Shot	Cond.	Syncopation Flash	Syncopation Pro	Syncopation Qwen	Transposition Flash	Transposition Pro	Transposition Qwen	Chord ID Flash	Chord ID Pro	Chord ID Qwen
Audio	ZS	Stand.	30.00	25.00	20.00	55.56	94.74	75.00	31.82	**47.73**	31.82
		CoT	**35.00**	25.00	20.00	76.92	**95.00**	65.00	31.82	43.18	31.82
		LogicLM	20.00	20.00	20.00	65.00	80.00	50.00	11.36	18.18	6.82
	FS	Stand.	31.58	63.16	40.00	**94.74**	90.00	90.00	25.00	40.91	31.82
		CoT	40.00	**65.00**	40.00	63.16	90.00	60.00	25.00	**52.27**	34.09
		LogicLM	40.00	55.00	20.00	60.00	90.00	35.00	6.82	13.64	18.18
MIDI	ZS	Stand.	84.21	95.00	25.00	**100.00**	**100.00**	85.00	50.00	97.73	22.73
		CoT	94.74	**100.00**	35.00	95.00	**100.00**	20.00	**100.00**	**100.00**	25.00
		LogicLM	90.00	80.00	20.00	**100.00**	**100.00**	10.00	93.18	**100.00**	**100.00**
	FS	Stand.	88.89	**100.00**	35.00	**100.00**	**100.00**	90.00	70.45	**100.00**	29.55
		CoT	95.00	**100.00**	25.00	**100.00**	**100.00**	60.00	97.73	**100.00**	29.55
		LogicLM	**100.00**	95.00	25.00	**100.00**	**100.00**	15.00	**100.00**	**100.00**	**100.00**
	Chance		20.00			50.00			25.00		

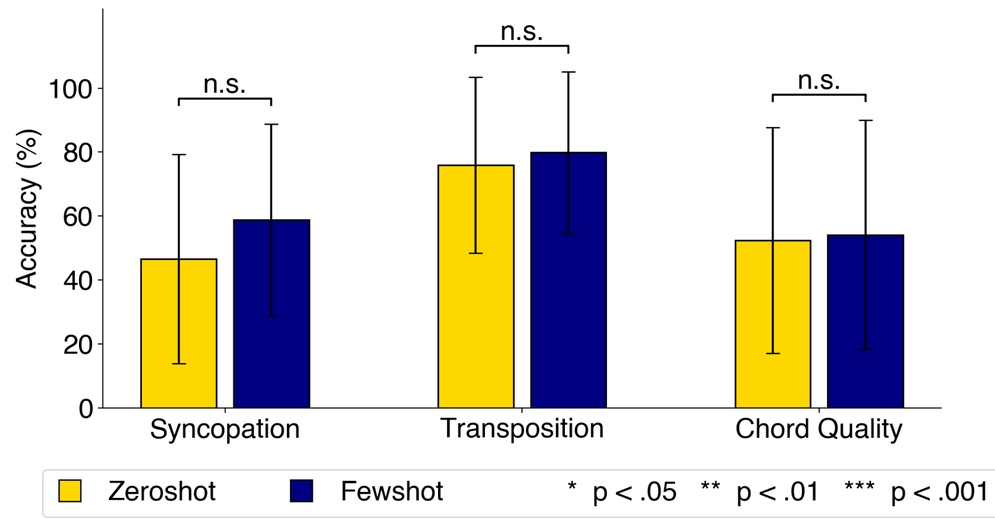
= Near or below chance = Picking up signal, but unreliable = Reliable

A. The Modality Gap (Audio vs. MIDI)



Gemini + MIDI = reliable; Audio lags by 40-45%

B. Impact of Few-Shot Examples



Few-shots do not affect accuracy

Conclusions

- Current state-of-the-art multimodal LLMs reason effectively over symbolic music data, yet still fail to truly ***"listen"*** with audio.
- Ceiling performance on symbolic data (MIDI) should not be mistaken for audio-native competence.
- This work makes the perception-reasoning boundary explicit and offers actionable guidance for building more robust music systems.
- ***Progress in this field will depend on developing stronger audio front-ends and may be bolstered by propagating uncertainty from perception into downstream solvers. Ultimately, for models to acquire genuine musical understanding, they must first learn how to listen.***

Paper, Stimuli, and Contact

Full Paper

2510.22455

Listen to the Stimuli

@brandoncarone

Personal Website

bcarone@nyu.edu