SAR-LM: SYMBOLIC AUDIO REASONING WITH LARGE LANGUAGE MODELS

Termeh Taheri, Yinghao Ma, Emmanouil Benetos

Introduction

Motivation

- Humans easily reason about sounds
- Large language models (LLMs) struggle with audio reasoning, unlike text and vision.
- Current methods rely on dense audio embeddings → limited accuracy, poor interpretability.

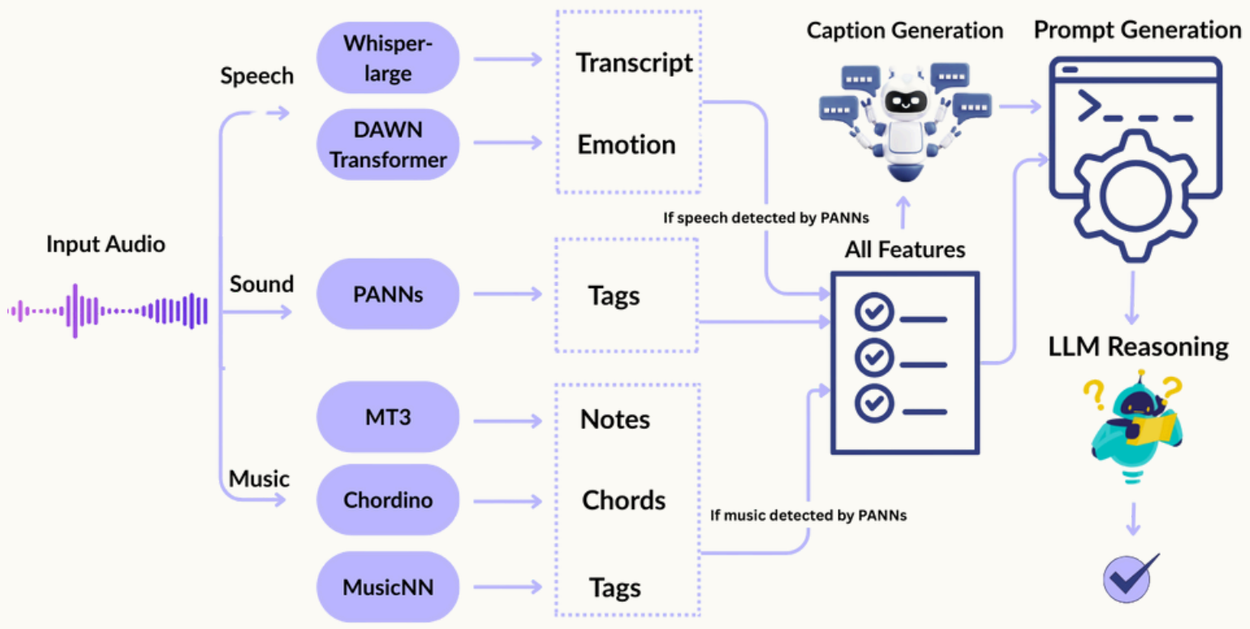
Our Work

- We propose SAR-LM, a symbolic audio reasoning pipeline.
- Converts raw audio into human-readable features: transcripts, emotions, sound events, music notes/chords.
- Enables LLMs to reason over structured inputs, not opaque embeddings.

Key Contributions

- Modular pipeline for symbolic audio reasoning.
- Evaluation on MMAU [1] and MMAR [2] benchmarks with competitive performance.
- Interpretability first: exposes why models fail, enabling detailed error analysis.

SAR-LM Pipeline



Experiments & Results

Setup

Datasets

- MMAU: 10k clips, 27 task types (speech, music, environment). Mini-test set (1k) used with public labels.
- MMAR: 1k audio–QA pairs with multi-step reasoning across speech, music, and mixed audio.

Models Tested

- Qwen2.5-Omni [3] – unstable outputs.
- Qwen3-Instruct [4] – more stable, moderate accuracy.
- Gemini 2.5 Pro [5] – best overall for both captioning & reasoning.

Dynamic Feature Selection

- SAR-LM has many possible features (transcripts, events, chords, tags). Including all can add noise.
- We use a GPT-style agent to pick only the relevant ones → Gemini 2.5 Pro gave stable, meaningful selections, improving reasoning accuracy.

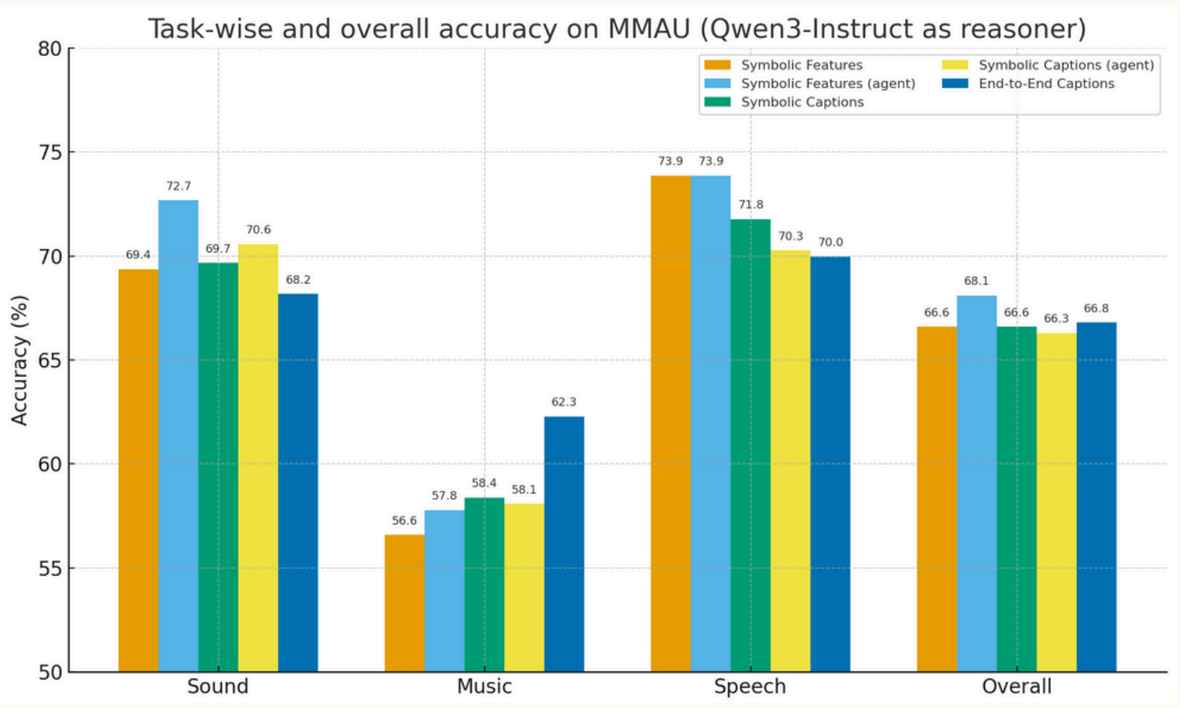
Results

We compare our best configuration (Gemini 2.5 Pro + symbolic features) with reported baselines, showing clear gains, on both MMAU and MMAR

Method	Sound	Music	Speech	Overall
MMAU (Best)	57.35	49.70	64.86	57.30
Audio-CoT	62.16	55.99	56.16	58.10
Audio-Reasoner	60.06	64.30	60.70	61.71
Ours (Gemini + Symbolic)	**73.27**	**64.97**	**82.28**	**73.5**

Method	Sound	Music	Speech	Overall
MMAR (Best)	**61.21**	50.97	72.11	65.6
Ours (Gemini + Symbolic)	52.73	**56.31**	**80.95**	**69.3**

On MMAU, symbolic features give strong speech and sound performance; agent-based selection further boosts accuracy, while captions help slightly in music tasks.



Error Analysis

- Example temporal reasoning question: “What was the order of the sounds?”
- Correct order: light switch → boiling water → doorbell → clock.
- Symbolic pipeline failed (missed first two sounds due to PANNs).
- End-to-end captioner succeeded (had full waveform access).
- Takeaway: Symbolic reasoning is only as strong as feature extraction.

Conclusion & Future Work

- SAR-LM: a modular pipeline for symbolic audio reasoning with LLMs.
- Achieves competitive performance on MMAU and MMAR while providing interpretability to diagnose errors (e.g., missed temporal events).
- Limitations: feature extraction is computationally heavy; errors in transcripts or music transcription can propagate.

Future work:

- Improve feature extraction (e.g., universal sound recognition).
- Integrate stronger pretrained encoders (e.g., MERT).
- Move toward unified feature extraction for both accuracy and interpretability.

Scan QR for CV & Portfolio

Open to PhD & research opportunities

termeh.taheri.dev@gmail.com

References

[1] Sakshi Sakshi, Utkarsh Tyagi, Saurav Kumar, Abhishek Seth, Rohan Selvakumar, Oriol Nieto, and Dinesh Manocha, “Mmau: A massive multi-task audio understanding and reasoning benchmark,” arXiv preprint arXiv:2410.19168, 2024.
[2] Ziyang Ma, Yuxuan Ma, Yifan Zhu, Chengming Yang, Yu-Wei Chao, Rongjie Xu, and Xia Chen, “Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix,” arXiv preprint arXiv:2505.13032, 2025.
[3] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin, “Qwen2.5-omni technical report,” arXiv preprint arXiv:2503.20215, 2025.
[4] Qwen Team, “Qwen3 technical report,” 2025.
[5] Google DeepMind, “Gemini 2.5 pro,” https://deepmind.google/technologies/gemini/#gemini-25, 2024, Accessed: 2025-08-14.