How does fine-tuning multimodal models like OpenPangu-7B-MLA on adversarial paralinguistic datasets like VoxPa

Assignee Research

June 10, 2026

Abstract

Audio large language models (Audio LLMs) demonstrate strong performance on speech understanding tasks, yet their ability to understand paralinguistic information remains limited. To systematically quantify this issue, we introduce VoxParadox, an adversarial benchmark with 2,000 verified examples, spanning 10 paralinguistic tasks, created with controlled speech synthesis to intentionally mismatch transcript claims and speaking style, enabling direct measurement of speech paralinguistic understanding. Evaluation of a diverse set of Audio LLMs reveals consistently low accuracy on acoustic ground

1 Introduction

This paper examines: Do Audio LLMs Listen or Read? Analyzing and Mitigating Paralinguistic Failures with VoxParadox. Research question: How does fine-tuning multimodal models like OpenPangu-7B-MLA on adversarial paralinguistic datasets like VoxParadox impact their performance on MMSU tasks compared to text-only and unaugmented multimodal baselines, measured by accuracy and robustness metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

15 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 2.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Audio Flamingo 2 (AF2) has an age prediction accuracy of 35.00% and a speaker identity recognition accuracy of 99.00%.	×	0.03
Audio Flamingo 3 (AF3) has an emotion recognition accuracy of 24.50% and a pitch comparison accuracy of 11.00%.	×	0.03
Qwen2-Audio-7B-Instruct has a gender prediction accuracy of 3.00% and a volume comparison accuracy of 22.00%.	×	0.02
SALMONN-7B has a speed comparison accuracy of 0.00% and a vocal range comparison accuracy of 0.00%.	×	0.02
Kimi-Audio-7B-Instruct has an intonation perception accuracy of 5.00% and a total speaker count recognition accuracy of	×	0.03
VITA-Audio has a pitch comparison accuracy of 8.50% and a volume comparison accuracy of 8.00%.	×	0.02
MiMo-Audio-7B-Instruct has an emotion recognition accuracy of 0.50% and a total speaker count recognition accuracy of 36	×	0.02
GPT-4o Audio has an age prediction accuracy of 4.50% and a speaker identity recognition accuracy of 1.00%.	×	0.02
Gemini 2.5 Flash has a gender prediction accuracy of 14.50% and a total speaker count recognition accuracy of 92.50%.	×	0.01
Most evaluated Audio LLMs exhibit high ALA alongside low GT accuracy, indicating systematic reliance on transcript-impli	×	0.09
The gap between ALA and GT accuracy is most pronounced in GPT-4o Audio, which matches yadv on 81.55% of examples while a	×	0.03

References

- http://arxiv.org/abs/2309.10783v1
- http://arxiv.org/abs/2605.27772v1
- http://arxiv.org/abs/2506.04779v3