GSA-1.7B vs. Text-Only LLMs on MMSU Under Paralinguistic Cue Removal

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does GSA-1.7B's accuracy on the MMSU benchmark compare to text-only LLMs of similar parameter counts when paralinguistic cues are removed. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Do Audio LLMs Listen or Read? Analyzing and Mitigating Paralinguistic Failures with VoxParadox. Research question: How does GSA-1.7B's accuracy on the MMSU benchmark compare to text-only LLMs of similar parameter counts when paralinguistic cues are removed?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

4 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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) achieves an average score of 30.85 on the VoxParadox benchmark across paralinguistic tasks.	×	0.09
Audio Flamingo 2 (AF2) achieves a score of 99.00 on the Gender prediction task.	×	0.02
Audio Flamingo 2 (AF2) achieves a score of 0.00 on the Emotion recognition task.	×	0.03
Audio Flamingo 3 (AF3) achieves an average score of 17.40 on the VoxParadox benchmark.	×	0.09
Qwen2-Audio-7B-Instruct achieves a score of 0.50 on the Speaker identity recognition task.	×	0.01
SALMONN-7B achieves a score of 0.00 on Emotion recognition, Pitch comparison, Volume comparison, Range comparison, Inton	×	0.04
Kimi-Audio-7B-Instruct achieves a score of 79.00 on the Emotion recognition task.	×	0.02
Step-Audio-R1 achieves a score of 93.00 on the Total speaker count task.	×	0.01
GPT-4o Audio achieves an average score of 8.60 on the VoxParadox benchmark.	×	0.03
Gemini 2.5 Flash achieves a score of 92.50 on the Total speaker count task.	×	0.00
GPT-4o Audio matches adversarial labels (yadv) on 81.55% of examples while achieving only 8.60% ground truth (GT) accura	×	0.04
Most evaluated Audio LLMs exhibit high Adversarial Label Accuracy (ALA) alongside low Ground Truth (GT) accuracy.	×	0.08
The performance gap between adversarial label matching and ground truth accuracy indicates a systematic reliance on tran	×	0.11
Qwen3-Omni achieves a score of 54.00 on the Total speaker count task.	×	0.01
MiMo-Audio-7B-Instruct achieves a score of 50.00 on the Gender prediction task.	×	0.01

References

- http://arxiv.org/abs/2506.04779v3

- http://arxiv.org/abs/2509.16589v2
- http://arxiv.org/abs/2605.27772v1