InternVL2-5-78B Performance on MMSU Spoken Language Understanding with Acoustic Features

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does InternVL2-5-78B perform on spoken language understanding tasks that require integrating acoustic paralinguistic features (e.g., emotions, pitch) compared to text-only models on MMSU. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. Research question: How does InternVL2-5-78B perform on spoken language understanding tasks that require integrating acoustic paralinguistic features (e.g., emotions, pitch) compared to text-only models on MMSU benchmark sub-tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

9 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
MMSU encompasses a wider range of acoustic features spanning 47 distinct tasks.	×	0.09
MMSU is the first benchmark to systematically incorporate linguistically grounded phenomena into spoken language underst	✓	0.16
MMSU requires models to integrate paralinguistic, phonetic, and semantic information in tasks such as sarcasm detection	×	0.09
MMSU includes 47 distinct tasks covering various linguistic phenomena and acoustic features.	×	0.11
MMSU evaluates 22 models, including 12 Speech-LLMs and 10 Omni Large Language Models (OmniLLMs) with audio processing ca	×	0.09
Each instance in MMSU consists of an audio clip and a text prompt, with the model choosing one of four options (A–D).	×	0.03
Answer options in MMSU are randomly ordered and balanced across the dataset to avoid potential positional bias.	×	0.02
All models in MMSU are evaluated with the same optimized instruction-following prompts to ensure fairness and minimize p	×	0.05
The sentence 'It's nice to meet you' is a common greeting that typically ends with a neutral or slightly falling intonat	×	0.02
The first part 'It's nice to meet you,' is spoken in a neutral tone, which is characteristic of a greeting.	×	0.05
The second part, 'you,' is spoken with a rising intonation.	×	0.03

References

- http://arxiv.org/abs/2506.04779v3
- http://arxiv.org/abs/2507.18119v2
- http://arxiv.org/abs/2207.08179v1