CodeGen-2.7B Benchmark Performance Across Reasoning and Language Tasks

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of CodeGen-2.7B on reasoning mathematics coding and language understanding tasks. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. 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: What are the benchmark performance scores of CodeGen-2.7B on reasoning mathematics coding and language understanding tasks.

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 3.0/10.

3 Results

4 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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 47 distinct tasks.	×	0.10
MMSU covers the capability types of Perception, Reasoning, and Prosody.	×	0.06
MMSU covers the linguistic phenomena of Intonation, Phonetics, Rhetoric, Syntactics, Non-Verbal, and Disfluency.	×	0.10
AudioBench covers 8 tasks and only includes the Perception capability type.	×	0.01
SD-Eval covers 4 tasks and only includes the Perception capability type.	×	0.01
SpokenWOZ covers 8 tasks and only includes the Reasoning capability type.	×	0.03
ADU-Bench covers 20 tasks and includes Reasoning and Prosody capability types.	×	0.03
VoxDilogue covers 12 tasks and includes Perception, Reasoning, Prosody, and Non-Verbal phenomena.	×	0.04
MMAU covers 27 tasks and includes Perception, Reasoning, and Prosody capability types.	×	0.04
VoiceBench covers 7 tasks and does not cover any of the listed capability types or linguistic phenomena in the table.	×	0.04
AIR-Bench covers 23 tasks and includes Perception and Reasoning capability types.	×	0.02
The evaluation conducted on MMSU included 22 models.	×	0.04
12 of the evaluated models are Speech-LLMs, including BLSP, LTU, SALMONN, and Qwen-Audio-Chat.	×	0.05
10 of the evaluated models are Omni Large Language Models (OmniLLMs), including GPT-4o-Audio and Gemini-1.5-Pro.	×	0.08
In the MMSU evaluation strategy, each instance consists of an audio clip and a text prompt where the model chooses one o	×	0.03
Answer options in the MMSU evaluation are randomly ordered and balanced across the dataset to avoid positional bias.	×	0.03
In the provided Table (p9) example, the ground truth answer for the intonation question is '(A) Failing Intonation'.	×	0.04
In the provided Table (p9) example, the model Qwen2.5-Omni-7B predicted '(D) Fall-Rise Intonation'.	×	0.02

References

- http://arxiv.org/abs/2506.04779v3
- http://arxiv.org/abs/2407.04973v1
- http://arxiv.org/abs/2207.08179v1