

Impact of LLM Selection in SLAM-ASR on Low- versus High-Resource Speech Recognition Performance on the MIRACL Benchmark

Assignee Research

June 22, 2026

Abstract

Large language models (LLMs) have demonstrated potential in handling spoken inputs for high-resource languages, reaching state-of-the-art performance in various tasks. However, their applicability is still less explored in low-resource settings. This work investigates the use of Speech LLMs for low-resource Automatic Speech Recognition using the SLAM-ASR framework, where a trainable lightweight projector connects a speech encoder and a LLM. Firstly, we assess training data volume requirements to match Whisper-only performance, re-emphasizing the challenges of limited data. Secondly, we show th

1 Introduction

This paper examines: Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages. Research question: How does the choice of LLM (e.g., Falcon, Llama, or Mistral) in the SLAM-ASR framework affect the downstream speech recognition performance in low-resource languages compared to high-resource languages, as measured by WER on the MIRACL benchmark?.

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 7.1/10.

3 Results

15 papers retrieved. 20 claims extracted; 15 independently verified. Quality review score: 7.1/10.

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
Librispeech contains 960 hours of data used for SLAM-ASR training.	×	0.09
Study [15] utilizes 1000 hours of data per language.	×	0.10
All models in the study are evaluated using the Word Error Rate (WER) metric.	✓	0.17
The linear projector training data from the CV IT dataset ranges from 10 hours to 252 hours.	✓	0.17
Increasing the quantity of training data consistently improves the overall performance of the model regardless of the LL	✓	0.27
EuroLLM 1.7B consistently outperformed Salamandra 2B in the experiments.	×	0.15
The performance gap between Salamandra and EuroLLM tends to close as more data becomes available.	✓	0.15
The SLAM-ASR configuration with EuroLLM 1.7B and 200 hours of training data achieves a WER of 6.4% on CV IT.	✓	0.15
The SLAM-ASR configuration with EuroLLM 1.7B and 252 hours of training data achieves a WER of 6.1% on CV IT.	×	0.13
Whisper-large-v3-turbo achieves a WER of 7.1% on CV IT.	✓	0.15
The SLAM-ASR framework with EuroLLM 1.7B and 200+ hours of data outperforms Whisper-large-v3-turbo on CV IT.	✓	0.21
Whisper-large achieves a WER of 4.7% on Fleurs IT.	×	0.12
Whisper-large-v3-turbo achieves a WER of 5.8% on Fleurs IT.	✓	0.16
The SLAM-ASR framework does not outperform Whisper-large or Whisper-large-v3-turbo on Fleurs IT.	✓	0.22
With 200 hours of CV IT training data and EuroLLM 1.7B, the WER on out-of-domain FL IT data is 13.2%.	✓	0.23
Additional LoRA fine-tuning experiments were conducted using 15 and 100 hours of training data.	✓	0.17
LoRA fine-tuning experiments were performed solely with EuroLLM 1.7B.	✓	0.19
The study uses the Common Voice (CV) Italian dataset for training data progression from 10 to 252 hours.	✓	0.16
The study explores leveraging a projector pre-trained on English data consisting of Librispeech 100 and 100 hours of CV E	✓	0.17
The study explores leveraging a projector pre-trained on 200 hours of CV Spanish data	✓	0.17

References

- <http://arxiv.org/abs/2501.17615v2>
- <http://arxiv.org/abs/2508.05149v1>
- <http://arxiv.org/abs/2304.00649v1>