Fine-Tuning SLAM-ASR for Low-Resource Language Speech Recognition with High-Resource Alignment

Assignee Research

June 16, 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: What is the impact of fine-tuning SLAM-ASR on low-resource language speech recognition when the LLM is first aligned with high-resource language speech data vs. text data?.

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 7.5/10.

3 Results

15 papers retrieved. 12 claims extracted; 9 independently verified. Quality review score: 7.5/10.

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
Increasing the quantity of training data consistently improves the overall performance of the SLAM-ASR model, regardless	✓	0.24
EuroLLM 1.7B consistently outperforms Salamandra 2B in the SLAM-ASR framework.	×	0.14
The performance gap between Salamandra and EuroLLM tends to close as more data are available.	✓	0.15
The SLAM-ASR framework with EuroLLM 1.7B and 200 or 252 hours of training data obtains a WER of 6.4% and 6.1% respective	✓	0.31
The SLAM-ASR framework does not outperform a Whisper-large- or Whisper-large-v3-turbo-only set-up on Fleurs IT (WER = 4.	✓	0.27
With 200 hours of CV IT training data with EuroLLM 1.7B, the WER on CV IT is 6.4% but 13.2% on FL IT.	✓	0.26
LoRA fine-tuning of the LLM can improve the alignment between speech and text tokens.	✓	0.17
The SLAM-ASR framework struggles with generalizing across domains, as evidenced by a higher WER on out-of-domain data.	×	0.11
The research investigates the viability of Speech LLMs in low-resource scenarios using ASR as a case study.	×	0.13
The research questions (RQ1 and RQ2) focus on the training data requirements for effective linear projector training and	✓	0.17
The study uses the Common Voice (CV) Italian dataset, ranging from 10 to 252 hours, to determine the amount of data need	✓	0.27
The study explores the effects of leveraging a projector pretrained on English data (Librispeech 100 and 100 hours of CV	✓	0.29

References

- http://arxiv.org/abs/2508.05149v1
- http://arxiv.org/abs/2506.05671v2

- http://arxiv.org/abs/1805.07467v2