



Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages

Seraphina Fong^{1,2}, Marco Matassoni², Alessio Brutti²

¹Department of Information Engineering and Computer Science, University of Trento, Italy

²Center for Augmented Intelligence, Fondazione Bruno Kessler, Italy

meiyueseraphina.fong@unitn.it, matasso@fbk.eu, brutti@fbk.eu

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 that leveraging mono- or multilingual projectors pretrained on high-resource languages reduces the impact of data scarcity, especially with small training sets. Using multilingual LLMs (EuroLLM, Salamandra) with whisper-large-v3-turbo, we evaluate performance on several public benchmarks, providing insights for future research on optimizing Speech LLMs for low-resource languages and multilinguality.

Index Terms: speech recognition, LLM, low-resource languages, multilinguality

1. Introduction

Low-resource settings, characterized by limited data or lack of manual annotations, remain a major challenge for speech technologies in production environments [1]. For example, current Automatic Speech Recognition (ASR) models are limited by the amount and diversity of training data required, supporting only about 100 of the more than 7,000 languages spoken globally [2, 3, 4]. As a result, the scarcity of high-quality training data continues to hinder advances in speech and language technologies for low-resource scenarios [3, 4, 5, 6]. Therefore, the need for technologies that support underrepresented languages is increasingly important, to not only preserve linguistic and cultural heritage, but to also ensure that a broader range of users worldwide have access to modern technologies.

Originally developed for text generation, large language models (LLM) have demonstrated increasing versatility in performing an expanding range of tasks [7, 8, 9, 10]. Of particular interest to the present work is how they have been extended to process speech and audio, enabling new possibilities for speech-related tasks. Recent LLM-based ASR approaches have implemented a trainable projector to integrate speech encoders with LLMs to improve ASR performance by leveraging the extensive knowledge of LLMs [11, 12, 13, 14, 15]. The projector's role is to align speech embeddings from the speech foundation encoder with the LLM's embedding space, enabling an LLM to handle speech input alongside textual prompts. The SLAM-ASR [16] approach demonstrated that implementing a lightweight layer as the projector between off-the-shelf speech

foundation encoders and LLMs could achieve state-of-the-art results on the English Librispeech dataset [17]. However, most emerging research in this domain predominantly focuses on high-resource languages (e.g., English, Mandarin Chinese), and therefore have large amounts of data available for training (e.g., 960 hours of Librispeech for SLAM-ASR [16, 18], 1000 hours of data per language in [15]). Other studies have investigated the impact of factors such as data characteristics or computational resources [18]. In particular, [19] demonstrates the crucial role of training data volumes for LLM-based audio-visual ASR. However, it remains unclear how well the LLM-based framework performs for low-resource settings, as data scarcity can impact all components of the pipeline. For example, the speech encoder may learn less robust representations, the LLM may generate less accurate transcripts due to limited exposure to the target language, and the projector may struggle to effectively align the different modality embeddings. Furthermore, in low-resource settings, there is often little to no data available not only for ASR, but also for other speech-related tasks such as spoken language understanding and spoken question answering. Given these challenges, it is essential to investigate the viability of Speech LLMs in low-resource scenarios. In this work, we use ASR as a case study. Therefore, our research questions (RQ) are as follows:

(RQ1) *How many hours of training data are needed to effectively train a linear projector?*

(RQ2) *Can linear projectors pretrained on a high-resource language be fine-tuned on another language to mitigate the impact of data scarcity in low-resource ASR?*

To address these questions, we progressively increase the amount of data used to train a linear projector within the SLAM-ASR framework, starting from 10 to 252 hours of the Common Voice (CV) Italian dataset [20], to determine the amount of data needed to match or exceed the performance of using a Whisper-only framework [21]. We subsequently explore the effects of leveraging a projector pretrained on English data (Librispeech 100 [17] and 100 hours of CV English) or 200 hours of CV Spanish data, fine-tuning them on varying amounts of Common Voice Italian data (10, 15, 100, and 200 hours) to assess its potential adaptability and performance in low-resource settings. We further extend this analysis to a real-world low-to-medium resource scenario with 10-15 hours of Galician data, where we also bootstrap a multilingual projector pretrained with a combination of English, Spanish, and Italian data. Consequently, we also investigate the impact of the selected LLM using two open-source multilingual LLMs (EuroLLM 1.7B [22] and Sala-

mandra 2B [23]), as well as reinforce previous findings [18] of SLAM-ASR performance in cross-domain evaluation settings.

Overall, this work presents the following contributions to provide insights into the applicability of Speech LLMs in low-resource scenarios, using SLAM-ASR as a case-study framework:

- Achieving performance comparable to Whisper-only models requires 200 hours of training data, highlighting the persistent challenge of data scarcity in low-resource settings.
- Fine-tuning projectors pretrained on high-resource languages enhances cross-domain performance, especially when the fine-tuning dataset is limited (10–15 hours). Performance is further improved when leveraging projectors pretrained on multiple languages (i.e., a “multilingual” projector).
- The choice of multilingual LLM impacts performance, with EuroLLM 1.7B performing better than Salamandra 2B.

2. Methodology

We employ the popular SLAM-ASR framework and code-base for our experimental analysis [16], chosen for its open-source accessibility and strong performance on English benchmarks. As depicted in Figure 1, SLAM-ASR consists of a frozen speech foundation encoder, a trainable linear projector, and a frozen LLM. The input speech signal, its corresponding transcript, and the text prompt are denoted as X_s , X_T , and X_P , respectively. For each input sample X_s , speech embeddings H_s are first extracted from the speech signal X_s via the speech encoder where:

$$H_s = \text{Encoder}(X_s) \quad (1)$$

We consider the speech encoder from the *Whisper-large-v3-turbo* model¹[21], a state-of-the-art transformer-based system for multilingual speech recognition. The extracted speech embeddings are downsampled by a factor of $k = 5$ to address the length discrepancy between speech and text features. A linear projector then transforms the downsampled H_s into E_s to align with the LLM’s input embedding dimension. Following the original SLAM-ASR implementation, the linear projector consists of a single hidden layer followed by ReLU activation and a regression layer, denoted as:

$$E_s = \text{Linear}(\text{ReLU}(\text{Linear}(\text{Downsample}(H_s)))) \quad (2)$$

The transcript X_T and prompt X_P are tokenized by the LLM’s tokenizer to obtain the corresponding transcript embedding E_T and prompt embedding E_P . During both training and decoding, the embeddings are concatenated to form the final input embedding. Specifically, during training, E_s , E_P , and E_T are concatenated, while during decoding, only E_s and E_P are concatenated before being passed into the LLM to generate the predicted output transcript. We consider two open-source LLMs that are multilingual on European languages (EuroLLM 1.7B² and Salamandra 2B³).

We first use Italian, a well-resourced language, to simulate low-resource settings by artificially limiting the amount of data for training. This choice allows us to disentangle possible issues

¹<https://huggingface.co/openai/whisper-large-v3-turbo>

²<https://huggingface.co/utter-project/EuroLLM-1.7B>

³<https://huggingface.co/BSC-IT/salamandra-2b>

related to the quality of the LLM and the speech encoder. We then examine the potential of pretraining a projector on high-resource language data (Figure 2A), and then fine-tuning them to evaluate its applicability in low-resource settings (Figure 2B).

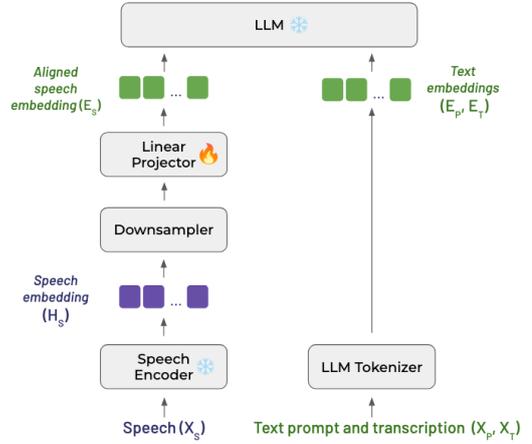


Figure 1: The SLAM-ASR framework [16] incorporates a trainable lightweight projector to align speech features H_s from a frozen speech encoder within the embedding space of a frozen LLM. The projector’s output features E_s are combined with the corresponding embeddings of a text prompt E_P (and, during training, of a transcription E_T) to generate the predicted transcription of the input speech signal.

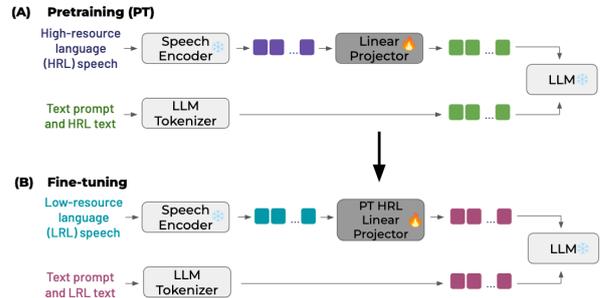


Figure 2: The pretraining and finetuning framework. A lightweight projector is first pretrained (PT) on a high-resource language (HRL; Panel A) and then finetuned on low-resource language data (LRL; Panel B).

3. Experimental Setup

3.1. Datasets

Our experiments are conducted using the Librispeech (LS) [17], Common Voice 20.0 (CV) [20], and Fleurs (FL) datasets [24]. For LS, we use the *train-clean-100* hours, *dev-other*, and *test-clean* subsets. For CV and FL, we use the English (EN), Italian (IT), Spanish (ES.419), and Galician (GL) data. With the Common Voice datasets, we create training data subsets totaling between 10 and 252 hours of data for IT, 100 hours for EN, 200 hours for ES, and 10 to 15 hours for GL, randomly selecting signals less than 20 seconds (14 for GL) due to computational limitations.

3.2. Training Configuration

All models were trained using the code-base provided by the official SLAM-ASR Github repository⁴. For all experiments, we utilize *Whisper-large-v3-turbo* as the speech encoder and a linear projector. We use two multilingual LLMs (EuroLLM 1.7B and Salamandra 2B) to compare the resulting performance as the choice of LLM has been found to impact it [18].

Our training setup follows the original SLAM-ASR implementation [16] in several aspects. We use the AdamW optimizer (learning rate = $1e-4$) and no weight decay for the trainable parameters. A learning rate scheduler with a 1,000-step warmup followed by a fixed maximum learning rate throughout training is applied. Training is set for a maximum of 100,000 training steps, which is stopped early if the validation loss does not decrease. We also use beam search decoding with *beam size* = 4. In a few experiments, we finetune the query and value projection layers of EuroLLM 1.7B with low-rank matrices using the repository’s default parameters ($r = 8$, $\alpha = 32$, drop-out = 0.05) while training the projector. This added 1.38M tunable parameters alongside the 17.31M parameter linear projector.

Due to the availability of computational resources, we differ in some areas: training is conducted on 100 hours of LS for three epochs and on other language datasets for six epochs. The batch size was set to four. All models were trained with one NVIDIA Ada Lovelace L40S GPU. Additionally, the context length of the model was adjusted accordingly based on the LLM used (i.e., 4096 for EuroLLM 1.7B and 8192 for Salamandra 2B).

For our monolingual projectors, we provided a language-specific prompt to the LLM, e.g., “*Transcribe [LANGUAGE] speech to text*”. For instance, the prompt used for Italian data was “*Transcribe Italian speech to text*”. Preliminary experiments showed that a language-specific prompt, compared to the general one utilized in the original SLAM-ASR implementation (“*Transcribe speech to text*”), affected the model’s performance. For instance, with English, the language-specific prompt reduced the WER on LS Test Clean from 5.9% to 4.6%.

4. Results

This section is organized according to the research questions posed in Section 1. All models are evaluated using the Word Error Rate (WER) metric to measure the similarity between the ground-truth transcripts versus a given model’s predicted transcripts.

4.1. RQ1: Training Data Requirements for Effective Linear Projector Training

Table 1 includes the results of the linear projector’s performance as the amount of training data from the CV IT dataset increases from 10 hours to 252 hours. The results demonstrate that increasing the quantity of training data consistently improves the overall performance of the model, regardless of the LLM used within the SLAM-ASR framework. However, using EuroLLM 1.7B consistently outperformed Salamandra 2B, highlighting the significant impact of LLM selection on downstream ASR performance. Note that the performance gap between Salamandra and EuroLLM tends to close as more data are available. This could be related to the larger context required by Salamandra. When comparing the performance of the SLAM-

⁴https://github.com/X-LANCE/SLAM-LLM/tree/main/examples/asr_librispeech

ASR framework to standalone whisper-only models, the configuration with EuroLLM 1.7B and 200 or 252 hours of training data obtains a WER of 6.4% and 6.1% respectively on CV IT, outperforming *Whisper-large-v3-turbo* (7.1%). This suggests that the SLAM-ASR framework can provide slightly better performance than whisper-only models, but requires a significant amount of training data to achieve these gains, in spite of the extremely lightweight projector being trained. Moreover, it does not outperform a *Whisper-large-* or *Whisper-large-v3-turbo*-only set-up on Fleurs IT (WER = 4.7% and 5.8%, respectively).

Performance is notably better on in-domain (CV IT) data compared to out-of-domain (FL IT) data across all configurations. For example, 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. This gap demonstrates the challenge that the SLAM-ASR framework has when generalizing across domains, a finding aligned with existing research [18].

Following [18]’s suggestion that LoRA fine-tuning of the LLM could improve the alignment between speech and text tokens, additional experiments with 15 and 100 hours of training data were conducted. These experiments were performed solely with EuroLLM 1.7B, given its better performance over Salamandra 2B. Our findings confirmed that LoRA can be beneficial. However, we observed only a marginal difference in in-domain performance: a reduction of 0.3% WER with 15 hours of data and 0.5% with 100 hours of data. Note, however, that LoRA improves out-of-domain performance on FL IT, reducing WER by 1.2% (with 15 hours of data) and 2.5% (with 100 hours of data). Given the marginal changes in performance, we do not employ LoRA in subsequent experiments.

Table 1: Results obtained over the CommonVoice (CV) Italian (IT) test set and the Fleurs IT test set. We progressively increase the amount of data used to train a linear projector within the SLAM-ASR framework, starting from 10 to 252 hours of the CV IT training set. Bolded values with an asterisk (*) indicate results surpassing a setup consisting only of *Whisper-large-v3-turbo* (CV IT WER = 7.1%). For reference, *Whisper-large* has a CV IT WER = 6.5% and a Fleurs IT WER = 4.7%, while *Whisper-large-v3-turbo* has a Fleurs IT WER = 5.8%.

Train	hours	EuroLLM 1.7B		Salamandra 2B	
		CV IT	Fleurs IT	CV IT	Fleurs IT
Fleurs	9	-	18.0	-	22.5
CV	10	14.0	35.2	33.6	111.4
CV	15	13.9	34.0	26.0	146.3
+LoRA		13.6	32.8	-	-
CV	20	11.4	31.6	19.3	68.3
CV	50	9.2	20.6	12.2	56.6
CV	100	7.6	16.3	10.4	80.6
+LoRA		7.1	13.8	-	-
CV	200	6.4*	13.2	7.7	65.7
CV	252	6.1*	12.3	-	-

4.2. RQ2: Bootstrapping pretrained linear projectors towards low-resource languages

Results in Table 1 show that at least 100-200 hours of labeled data are needed to achieve state-of-the-art ASR performance. This amount may be prohibitive in several languages. Therefore, we investigate a transfer learning approach by pretraining the projector on a different language. Given its superior per-

Table 2: Results on Common Voice Italian (CV IT) and Fleurs (FL) IT test sets after finetuning a projector pretrained on English LibriSpeech 100 (LS100 EN→), 100 hours of Common Voice (CV) English data (CV100 EN→), or 200 hours of Common Voice Spanish data (CV200 ES→) using 10, 15, 100, or 200 hours of CV IT training data. Performance is compared to a projector trained from scratch on CV IT (Scratch).

Training data (hours)	Scratch		LS100 EN→		CV100 EN→		CV200 ES→	
	CV IT	FL IT	CV IT	FL IT	CV IT	FL IT	CV IT	FL IT
CV (10)	14.0	35.2	12.0	16.1	9.8	17.1	8.6	20.3
CV (15)	13.9	34.0	12.8	15.7	10.0	16.9	8.8	16.1
CV (100)	7.6	16.3	8.1	12.6	7.7	16.4	7.3	17.6
CV (200)	6.4	13.2	6.6	11.6	6.7	15.8	6.4	17.3

Table 3: Results on the Common Voice (CV) and Fleurs (FL) Galician (GL) test sets after finetuning a projector pretrained on 200 hours of CV Spanish (CV200 ES→), Italian (CV200 IT→), or a combination of the two former datasets with Librispeech 100 (MULTI→), using 10–15 hours (h) of CV GL training data. Performance is compared to a projector trained from scratch on CV GL (Scratch). For reference, standalone Whisper-large-v3-turbo has a CV GL WER = 16.2% and FL GL WER = 14.8%.

Train(h)	Scratch		CV200 ES→		CV200 IT→		MULTI→	
	CV	FL	CV	FL	CV	FL	CV	FL
CV (10)	18.6	43.0	13.9	27.1	15.1	26.8	13.3	19.4
CV (15)	17.5	42.4	12.4	22.0	13.5	22.1	12.4	15.8

formance, we only consider EuroLLM 1.7B in the next experiments. Table 2 compares performance across various bootstrapping setups using models pretrained on high-resource datasets (100–200 hours) and evaluated on Italian datasets. Models pretrained on English (LS100 EN→, CV100 EN→) and Spanish (CV200 ES→) fairly consistently outperformed the baseline “Scratch” model trained solely on CV Italian.

In low-resource settings, such as when only 10 or 15 hours data are available, leveraging pretrained models appears to provide substantial gains both within in- and out-of-domain contexts. When fine-tuned on smaller datasets (e.g., CV IT with 10 or 15 hours of training data), the pretrained projectors demonstrate a significant improvement in WER over a model trained from scratch. For example, on CV IT with 10 hours of training data, the WER drops from 14.0% (Scratch) to 9.8% (CV100 EN→) and 8.6% (CV200 ES→). Similarly, on FL IT with 10 hours of training data, the WER decreases from 35.2% (Scratch) to 17.1% (CV100 EN→) and 20.3% (CV200 ES→). These findings suggest that pretraining may help generalize the model. However, as the amount of fine-tuning data increases to 100–200 hours, the advantage of fine-tuning a pretrained model diminishes with smaller performance gains. When fine-tuning using 200 hours of CV IT, the WER of the Scratch model is 6.4%, which aligns with the performance of LS100 EN→ at 6.6%.

It is also worth noting that performance appears to be dependent on the language, semantic domain, and acoustic properties of the data used to pretrain the projector. In the case of CV IT, starting from a projector pretrained on CV gives better results. Moreover, using Spanish data is better than employing a projector trained on English as the acoustic similarity with Italian is higher for the former. Interestingly, we observe that bootstrapping the projector from an English pretrained model (LS100/CV100 EN→) has more performance gains in the out-of-domain context compared to Spanish. Future work could consider further exploring linguistic factors, multilingual pre-

training, or other methods to enhance cross-lingual transfer, especially between closely related languages to further optimize outcomes.

4.3. Galician: a low-resource case study

Building on the findings in subsection 4.2, we extend our bootstrapping analysis to Galician (GL) to assess the applicability of bootstrapping with an actual low resource language. Table 3 includes the results of fine-tuning a projector pretrained either on 200 hours of Common Voice Spanish (CV200 ES→) or Italian data (CV200 IT→), or on multiple language datasets (MULTI→; LibriSpeech 100, CV200 ES, and CV200 IT), compared to a projector trained from scratch on 10 or 15 hours of CV GL (Scratch). These findings align with our analysis in subsection 4.2, demonstrating that transfer learning from a related high-resource language enhances performance over training from scratch, even with limited Galician data. Moreover, we find that a multilingual projector (MULTI→) can further improve performance and generalization capabilities. With only 10 hours of GL finetuning data, a fine-tuned multilingual projector achieves a WER of 13.3% and 19.4% on CV GL and FL GL, respectively, versus 18.6% and 43% when trained solely from Scratch. While benefits are still observed at 15 hours of data (WER 12.4% CV GL and 15.8% FL GL), they are less pronounced.

5. Conclusion

This study aimed to investigate the potential of SLAM-ASR, a recent Speech LLM-based framework, in low-resource scenarios by assessing the data volume requirements and the impact of high-resource pretraining and fine-tuning on data scarcity. Using *Whisper-large-v3-turbo* as the speech encoder, a linear projector, and a multilingual LLM, results indicate that at least 100–200 hours of training data are needed for the SLAM-ASR framework to match the performance of Whisper-only models, reinforcing the difficulty of training in low-resource settings. Findings however showed that pretraining on high-resource languages can significantly improve cross-domain performance, especially with smaller fine-tuning datasets (10–15 hours).

Overall, Speech LLMs, in the context of the SLAM-ASR framework, show potential in low-resource scenarios through pretraining and fine-tuning strategies. Although its reliance on larger amounts of training data and challenges with cross-domain performance present areas for further development, these findings highlight avenues for future research towards adapting and optimizing the framework for low-resource languages.

6. Acknowledgements

This paper was partially funded by the European Union's Horizon 2020 project ELOQUENCE (grant 101070558).

7. References

- [1] L.-M. Lam-Yee-Mui, W. B. Kheder, V. B. Le, C. Barras, and J.-L. Gauvain, "Multilingual Models with Language Embeddings for Low-resource Speech Recognition," *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:264886709>
- [2] K. Fatehi, M. T. Torres, and A. Kucukylmaz, "An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments," *Speech Communication*, pp. 103–151, 2024.
- [3] Y. Liu, X. Yang, and D. Qu, "Exploration of Whisper fine-tuning strategies for low-resource ASR," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, p. 29, 2024.
- [4] M.-H. Chiang, C.-H. Lai, and H.-S. Chiu, "WhisperHakka: A Hybrid Architecture Speech Recognition System for Low-Resource Taiwanese Hakka," in *35th Conference on Computational Linguistics and Speech Processing*, 2023, pp. 390–396.
- [5] V. Roger, J. Farinas, and J. Pinquier, "Deep neural networks for automatic speech processing: a survey from large corpora to limited data," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, p. 19, 2022.
- [6] Y.-F. Cheng, L.-W. Chen, H.-S. Lee, and H.-M. Wang, "Exploring the impact of data quantity on asr in extremely low-resource languages," *ArXiv*, vol. abs/2409.08872, 2024.
- [7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.
- [8] J. Wu, W. Gan, Z. Chen, S. Wan, and S. Y. Philip, "Multimodal large language models: A survey," in *IEEE International Conference on Big Data*. IEEE, 2023, pp. 2247–2256.
- [9] Y. Bai, J. Chen, J. Chen, W. Chen, Z. Chen, C. Ding, L. Dong, Q. Dong, Y. Du, K. Gao *et al.*, "Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition," *arXiv preprint arXiv:2407.04675*, 2024.
- [10] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli *et al.*, "Prompting large language models with speech recognition abilities," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 351–13 355.
- [11] A. Mittal, D. Prabhu, S. Sarawagi, and P. Jyothi, "Salsa: Speedy asr-llm synchronous aggregation," *arXiv preprint arXiv:2408.16542*, 2024.
- [12] K. Mundnich, X. Niu, P. Mathur, S. Ronanki, B. Houston, V. R. Elluru, N. Das, Z. Hou, G. Huybrechts, A. Bhatia *et al.*, "Zero-resource speech translation and recognition with LLMs," *arXiv preprint arXiv:2412.18566*, 2024.
- [13] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, M. Zejun, and C. Zhang, "SALMONN: Towards Generic Hearing Abilities for Large Language Models," in *The Twelfth International Conference on Learning Representations*, 2024.
- [14] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.
- [15] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu *et al.*, "On decoder-only architecture for speech-to-text and large language model integration," in *IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE, 2023, pp. 1–8.
- [16] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang *et al.*, "An Embarrassingly Simple Approach for LLM with Strong ASR Capacity," *arXiv preprint arXiv:2402.08846*, 2024.
- [17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing*. IEEE, 2015, pp. 5206–5210.
- [18] S. Kumar, I. Thorbecke, S. Burdisso, E. Villatoro-Tello, K. Hacıoğlu, P. Rangappa, P. Motlicek, A. Ganapathiraju, A. Stolcke *et al.*, "Performance evaluation of SLAM-ASR: The Good, the Bad, the Ugly, and the Way Forward," *arXiv preprint arXiv:2411.03866*, 2024.
- [19] U. Cappellazzo, M. Kim, H. Chen, P. Ma, S. Petridis, D. Falavigna, A. Brutti, and M. Pantic, "Large Language Models Are Strong Audio-Visual Speech Recognition Learners," 2024.
- [20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.
- [21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *40th International Conference on Machine Learning*, ser. ICML'23, 2023.
- [22] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. P. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. C. de Souza, A. Birch, and A. Martins, "Euroollm: Multilingual language models for europe," *ArXiv*, vol. abs/2409.16235, 2024.
- [23] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, and M. Villegas, "Salamandra technical report," 2025.
- [24] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech," in *IEEE Spoken Language Technology Workshop*, 2023, pp. 798–805.