Language Family Matters: Evaluating LLM-Based ASR Across Linguistic Boundaries

Yuchen Zhang[1,2] **and Ravi Shekhar**[1,2] **and Haralambos Mouratidis**[1,2]
[1]Institute for Analytics and Data Science, University of Essex
[2]School of Computer Science and Electronic Engineering, University of Essex
{yuchen.zhang,r.shekhar,h.mouratidis}@essex.ac.uk

Abstract

Large Language Model (LLM)-powered Automatic Speech Recognition (ASR) systems achieve strong performance with limited resources by linking a frozen speech encoder to a pretrained LLM via a lightweight connector. Prior work trains a separate connector per language, overlooking linguistic relatedness. We propose an efficient and novel connector-sharing strategy based on linguistic family membership, enabling one connector per family, and empirically validate its effectiveness across two multilingual LLMs and two real-world corpora spanning curated and crowd-sourced speech. Our results show that family-based connectors reduce parameter count while improving generalization across domains, offering a practical and scalable strategy for multilingual ASR deployment.

1 Introduction

Automatic speech recognition (ASR) has progressed rapidly with advances in model architectures and training methods. These advances have allowed ASR to evolve from single-language systems to models capable of covering several languages in a unified framework. Multilingual ASR systems, from early multilingual acoustic models (Schultz and Waibel, 2001; Kamper et al., 2021; Abate et al., 2020) to modern universal encoders such as wav2vec 2.0, HuBERT, XLS-R, and Whisper (Baevski et al., 2020; Hsu et al., 2021; Babu et al., 2022; Radford et al., 2023), demonstrate the potential of broad multilingual coverage but also face a continuing challenge in supporting a wide range of languages while preserving efficiency and accuracy.

One way forward is to combine speech encoders with large language models (LLMs) to build SpeechLLMs, which bring together the acoustic representation power of speech encoders and the reasoning and generative capabilities of LLMs (Xue et al., 2024; Verdini et al., 2024; Fan et al., 2025). Some approaches connect speech encoders and LLMs by training large end-to-end models, such as Qwen-Audio, Salmonn, and decoder-only pipelines (Chu et al., 2023; Tang et al., 2023; Wu et al., 2023). However, such end-to-end SpeechLLMs are typically parameter heavy and computationally expensive, making them costly to train and fine-tune across languages or domains. A more parameter-efficient alternative is the use of lightweight connectors or adapters, which map acoustic features from speech encoders into the text space of LLM decoders, while the encoder and decoder can be either frozen or trainable depending on task requirements (Ma et al., 2024, 2025; Kumar et al., 2025; Fong et al., 2025; Mundnich et al., 2025). Despite their progress, current studies on SpeechLLMs mainly focus on architectural design, raising the open question of how the level of data granularity affects multilingual ASR. To address this gap, we focus on two research questions:

RQ1: *Which level of data granularity, individual language or language family, is more effective for multilingual ASR*?

RQ2: *How well do connectors generalize across domains?*

This paper addresses these gaps with a large-scale study spanning ten language families comprising nearly forty languages, evaluated across two multilingual datasets and two LLM backbones. We systematically evaluate connectors trained at the family and language levels and further examine their robustness under cross-domain transfer. Our main contributions are: (1) We systematically compare language and family connectors for multilingual ASR, covering ten language families comprising nearly forty languages. (2) We conduct a detailed evaluation of cross-domain generalization, assessing how connectors trained on one corpus transfer to a different domain. (3) We provide empirical evidence that family-level connectors strike

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 487–499

the best balance between coverage and specificity, resulting in lower WERs and greater stability.

2 Method

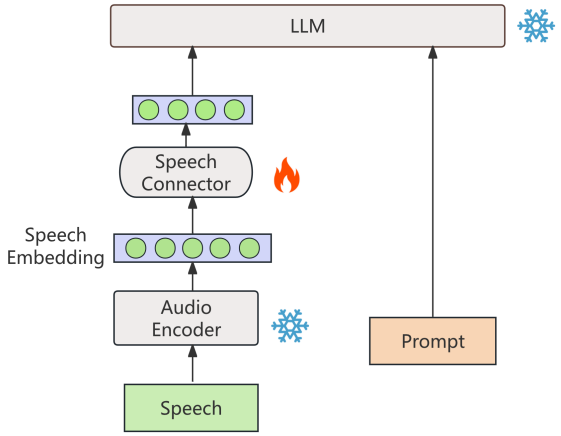

Figure 1: Overall framework for multilingual ASR.

We adopt an Encoder–Connector–Decoder architecture for the multilingual ASR, as shown in Figure 1. Given an input speech signal S, the audio encoder extracts a sequence of high-dimensional acoustic embeddings:

$$\mathbf{H}_{\text{speech}} = \mathcal{E}(S) \in \mathbb{R}^{B\times T\times E}, \quad (1)$$

where $\mathcal{E}$ denotes the audio encoder, and B, T, and E denote the batch size, the number of acoustic frames, and the encoder hidden dimension, respectively.

To align the encoder outputs with the embedding space of the LLM, the speech representations $\mathbf{H}_{\text{speech}}$ are first downsampled by a factor K, where every K consecutive frames are concatenated into a single vector. The stacked features are then transformed through two successive linear layers:

$$\mathbf{H}_{\text{proj}} = \mathcal{L}_2\big(\sigma_{\text{GELU}}(\mathcal{L}_1(\mathbf{H}_{\text{stacked}}))\big), \quad (2)$$

$$\mathbf{H}_{\text{stacked}} = \mathcal{D}_K(\mathbf{H}_{\text{speech}}) \in \mathbb{R}^{B\times \frac{T}{K}\times(E\cdot K)}, \quad (3)$$

where $\mathcal{D}_K(\cdot)$ denotes the downsampling operator, $\mathcal{L}_1$ and $\mathcal{L}_2$ denote a linear projector, and $\sigma_{\text{GELU}}(\cdot)$ is the Gaussian Error Linear Unit (GELU) activation function.

The projected representations $\mathbf{H}_{\text{proj}}$ are then passed to the LLM decoder, which generates the transcript in an autoregressive manner. At each step t, the decoder predicts the next token y_t given the previous tokens $y_{1:t-1}$ and the projected speech features $\mathbf{H}_{\text{proj}}$. Decoding terminates once the end-of-sequence token is generated, at which point the complete transcription is obtained. It should be noted that throughout this process, the speech encoder and the LLM decoder remain frozen, while only the connector is trainable.

3 Experiments and Discussion

3.1 Datasets

We adopt two widely used multilingual speech corpora for our experiments, FLEURS (Conneau et al., 2023) and CommonVoice_22 (Ardila et al., 2020) [1]. To ensure diversity while maintaining balance across language families, we focus on seven major families, including Afro-Asiatic, Austronesian, Dravidian, Indo-European, Niger-Congo, Turkic, and Uralic, as shown in Table 1. Since each family contains a large number of languages, we sample up to five representatives per family. For Indo-European, which is especially diverse, we further group languages by branch. The final set of languages used in our study is summarized in Table 1, where we report the available training hours from both FLEURS and CommonVoice for each selected language. We cap the training data at 100 hours per language, as some CommonVoice languages contain extremely large training sets.

3.2 Experimental Setting

In this study, we employ Whisper-large-v3 (Radford et al., 2022) as the speech encoder and adopt Gemma-2-2b (Team et al., 2024), and Salamandra-2b (Gonzalez-Agirre et al., 2025) as LLM decoders [2]. Two types of connectors are evaluated: a language-specific connector LangConn, trained on data from a single language, and a family-level connector FamConn, trained on the combined data of all languages within a given family. In all of our experiments, the parameters of both the speech encoder and the LLM are frozen, and only the connector is trainable. Every connector was trained for 10 epochs with early stopping, using a batch size of 10, the AdamW optimizer, a learning rate of 1e-4, and a weight decay of 1e-6. During inference, beam search is used for decoding, with the beam size set to 2. The prompt is fixed for all experiments as "Transcribe the speech to text:". Training and evaluation are conducted on four NVIDIA A10

[1] For simplicity, we will use CommonVoice to refer to CommonVoice_22 in the rest of the paper.

[2] For simplicity, we will use Gemma and Salamandra to refer to Gemma-2-2b and Salamandra-2b in the rest of the paper.

Table 1: Selected languages and training set statistics in hours (FLEURS / CommonVoice).

Family	Languages (FLEURS / CommonVoice, hours)				
Afro-Asiatic	Amharic (11.1/0.9)	Arabic (6.1/32.4)	Hausa (13.6/2.3)	Hebrew (9.5/1.2)	Maltese (9.9/2.4)
Baltic	Latvian (6.5/23.3)	Lithuanian (9.8/11.8)			
Celtic	Irish (12.1/0.6)	Welsh (12.2/11.5)			
Dravidian	Malayalam (10.1/1.4)	Tamil (8.7/83.7)	Telugu (7.9/0.1)		
Germanic	Danish (7.5/4.2)	Dutch (7.7/54.2)	English (7.5/100)	German (9.0/100)	Swedish (8.4/9.1)
Indo-Iranian	Bengali (10.7/34.2)	Hindi (6.7/5.8)	Persian (12.1/31.6)	Punjabi (6.4/1.2)	Urdu (7.0/8.4)
Niger-Congo	Igbo (13.8/0.01)	Swahili (13.5/69.9)	Yoruba (10.0/2.4)		
Romance	French (10.3/100)	Galician (6.7/96.5)	Portuguese (10.2/26.3)	Romanian (10.1/5.7)	Spanish (8.8/100)
Slavic	Belarusian (9.5/100)	Polish (9.2/35.5)	Russian (8.1/38.1)	Serbian (10.7/1.8)	Slovenian (7.8/1.5)
Turkic	Azerbaijani (9.3/0.2)	Kazakh (11.8/0.8)	Kyrgyz (9.3/2.3)	Turkish (8.3/43.8)	

GPUs, and performance is assessed using the Word Error Rate (WER) metric.

3.3 Results and Discussion

RQ1: Impact of Data Granularity on Multilingual ASR

To investigate whether modeling linguistic structure at the family level benefits multilingual ASR, we compare the performance of LANGCONN and FAMCONN across both LLMs and datasets. Table 2 summarizes the WER at the family level, while Table 6 (in Appendix A) provides detailed language level WERs. It should be noted that for the FAMCONN, the family-level evaluation is performed directly on the merged test set containing all languages within a family. For the LANGCONN, predictions and references are obtained separately for each language, then concatenated before calculating family-level WER.

From the results, we can see that across all families and configurations, FAMCONN outperforms LANGCONN in almost all cases. For instance, on Salamandra with FLEURS, FAMCONN achieves 15.67% WER for Germanic and 11.15% for Romance, compared to 23.37% and 37.47% with LANGCONN. The gap widens on CommonVoice, with Germanic improving from 77.71% to 33.55%, and Romance from 62.18% to 25.49%. Gains are particularly pronounced in high-variance families such as Slavic, Baltic, and Romance, where shared morphological and phonological patterns may benefit from joint modeling. For instance, Belarusian and Latvian show WER reductions exceeding 70% on FLEURS, indicating strong transferability within families.

These benefits are robust across both Gemma and Salamandra, and across datasets. On CommonVoice, which is crowd-sourced and more acoustically diverse, the relative advantage of family-based connectors is even more pronounced, suggesting improved generalization under domain shift. For example, the Germanic family shows a reduction from 77.71% to 33.55% WER with Salamandra on CommonVoice. Additionally, while FAMCONN outperforms LANGCONN for both Gemma and Salamandra in most cases, the magnitude of improvement differs. Salamandra shows greater instability in language-specific transcription, including repetition and language drift. In these cases, family-level grouping provides substantial corrective effects, leading to larger relative gains for FAMCONN. Gemma exhibits more stable performance, resulting in smaller but more consistent gains across languages. These trends indicate that the benefit of family-level sharing interacts with the inherent stability of the underlying LLM backbone.

However, FAMCONN are not universally superior. In a minority of cases, such as the Afro-Asiatic and Dravidian families, FAMCONN underperforms. The families in which FAMCONN performs worse than LANGCONN tend to be those where historical family membership does not align well with the acoustic or phonological similarity required for stable connector sharing. For example, in the Afro-Asiatic family, Arabic, Hebrew, Amharic, and Maltese differ substantially in script and phonological structure (Fabri et al., 2014). Similarly, the Dravidian family is internally diverse, with clear cross-language variation in phonology (Kolipakam et al., 2018). These factors may limit the extent to which a single family-level connector can generalise across all languages in the group. By contrast, families such as Germanic, Romance, Slavic, and Baltic display more consistent improvements under FAMCONN, suggesting that family-level sharing is more effective when intra-family conditions are relatively coherent.

Overall, our results indicate that the benefits of family-level pooling depend on the interaction

Table 2: Family level WER↓ (%) across datasets and connector types. Δ = Fam – Lang. Best in **bold**.

Family	FLEURS						CommonVoice					
	Salamandra			Gemma			Salamandra			Gemma		
	Lang	Fam	Δ	Lang	Fam	Δ	Lang	Fam	Δ	Lang	Fam	Δ
Afro-Asiatic	**70.56**	92.69	+22.13	**44.77**	46.66	+1.89	**93.90**	109.56	+15.66	105.89	**91.60**	-14.29
Baltic	113.21	**39.64**	-73.57	31.23	**27.56**	-3.67	123.26	**91.75**	-31.51	61.63	**59.96**	-1.67
Celtic	116.78	**113.73**	-3.05	53.58	**53.11**	-0.47	211.35	**177.34**	-34.01	119.70	**92.65**	-27.05
Dravidian	**28.34**	38.60	+10.26	**27.22**	28.34	+1.12	125.18	**71.47**	-53.71	105.45	**85.59**	-19.86
Germanic	23.37	**15.67**	-7.70	17.26	**14.22**	-3.04	77.71	**33.55**	-44.16	44.86	**25.85**	-19.01
Indo-Iranian	49.20	**46.09**	-3.11	39.43	**29.60**	-9.83	73.39	**67.87**	-5.52	101.68	**97.14**	-4.54
Niger-Congo	53.59	**52.39**	-1.20	**46.25**	48.14	+1.89	278.72	**85.30**	-193.42	**82.43**	90.48	+8.05
Romance	37.47	**11.15**	-26.32	12.80	**10.58**	-2.22	62.18	**25.49**	-36.69	38.42	**29.53**	-8.89
Slavic	95.96	**24.61**	-71.35	25.07	**21.69**	-3.38	180.62	**44.69**	-135.93	47.54	**41.58**	-5.96
Turkic	60.09	**45.45**	-14.64	26.29	**24.55**	-1.74	82.02	**69.07**	-12.95	79.16	**71.59**	-7.57

Table 3: Family level cross-domain WER↓ (%) using Gemma. Δ = Fam – Lang. Best in **bold**.

Family	CV → FL			FL → CV		
	Lang	Fam	Δ	Lang	Fam	Δ
Afro-Asiatic	154.16	**150.60**	-3.56	79.64	**67.89**	-11.75
Baltic	**111.54**	122.68	+11.14	73.84	**49.66**	-24.18
Celtic	179.45	**169.31**	-10.14	**104.15**	105.71	+1.56
Dravidian	**128.39**	150.83	+22.44	**75.81**	76.23	+0.42
Germanic	124.01	**56.03**	-67.98	26.41	**23.66**	-2.75
Indo-Iranian	144.19	**132.79**	-11.40	90.92	**84.98**	-5.94
Niger-Congo	**127.46**	132.63	+5.17	**69.03**	89.46	+20.43
Romance	100.25	**72.91**	-27.34	27.51	**26.23**	-1.28
Slavic	110.59	**102.64**	-7.95	103.42	**65.81**	-37.61
Turkic	**118.94**	130.80	+11.86	47.76	**46.03**	-1.73

between linguistic similarity and dataset quality, and that genealogical relatedness alone does not guarantee improved performance. These observations also suggest that heterogeneous families may benefit from finer-grained alternatives such as subfamily connectors or lightweight language-specific adapters.

RQ2: Cross-Domain Generalization ability of Connectors

Table 3 presents the family-level WERs across language families when training and evaluation occur on mismatched speech domains, i.e., connectors were trained on FLEURS but tested on CommonVoice, and vice versa. The results compare FAMCONN and LANGCONN for both directions of transfer. Detailed language-specific WERs are summarized in Table 7 available in Appendix A.

When training on CommonVoice and evaluating on FLEURS, FAMCONN outperforms LANGCONN in more cases. For example, in the Germanic family, the WER drops from 124.01% using LANGCONN to 56.03% using FAMCONN, representing a substantial reduction in error. Similar trends are observed in the Slavic and Romance families, where FAMCONN achieves better generalization. However, we also observe that for some families, such as Dravidian, LANGCONN surpasses FAMCONN over 20%. Overall, under this setting, FAMCONN can yield large gains in certain families, while LANGCONN remains competitive and occasionally better for others. In the reverse setting, where training is conducted on FLEURS and evaluation takes place on CommonVoice, the benefits of FAMCONN are clearer. FAMCONN is better across seven out of ten families. While the improvements are generally smaller than those observed in the Germanic family in the CV-to-FL setting, we still observe notable gains in some families. For instance, in the Slavic family, the WER is reduced from 103.43% to 64.81%, yielding a 37.61% absolute improvement. Moreover, in this setting, when FAMCONN underperforms LANGCONN, the performance gap is typically modest. Despite some family-specific exceptions and varying effect sizes across transfer directions, FAMCONN generally outperforms LANGCONN. This pattern suggests that family-level representations capture broader phonological and prosodic regularities that support cross-domain transfer.

At the language level, the above pattern becomes even more apparent (see Table 7 in Appendix A). A majority of languages benefit significantly from FAMCONN, with large reductions in WER. For instance, Serbian exhibits a large absolute improvement in both transfer directions, indicating that family-level information provides strong inductive bias for cross-domain generalization. Similarly, Slovenian consistently shows lower error rates under FAMCONN, suggesting that shared phonological and morphological characteristics within the Slavic family are effectively exploited by family-level connectors. In contrast, some languages show a strong preference for LANGCONN. In these cases,

FAMCONN leads to noticeable performance degradation, with WER increases exceeding 20%. This behavior can be attributed to the high internal diversity within their respective language families. For instance, for Telugu in the Dravidian family and Punjabi in the Indo-Iranian family, the substantial phonetic and prosodic differences across family members limits the usefulness of family-level representations, making language-specific connectors more effective. (Jairam et al., 2024).

In summary, our cross-domain evaluation demonstrates that FAMCONN provides a robust inductive bias for transfer across mismatched speech domains. Specifically, FAMCONN outperforms LANGCONN for most language families and individual languages, particularly those with strong intra-family phonological similarity (e.g., Slavic and Germanic). We also observe that the effectiveness of FAMCONN is limited when the target domain introduces greater acoustic and lexical variability. This highlights the importance of accounting for both source- and target-domain characteristics when designing multilingual generalization strategies.

Table 4: Family level WER↓ (%) using Gemma across LANGCONN, FAMCONN and UNICONN. Best in **bold**.

Family	FLEURS		
	Lang	Fam	Uni
Afro-Asiatic	**44.77**	46.66	54.97
Baltic	31.23	**27.56**	29.86
Celtic	53.58	**53.11**	63.30
Dravidian	**27.22**	28.34	39.36
Germanic	17.26	**14.22**	15.37
Indo-Iranian	39.43	**29.60**	50.38
Niger-Congo	**46.25**	48.14	60.96
Romance	12.80	**10.58**	11.55
Slavic	25.07	**21.69**	22.68
Turkic	26.29	**24.55**	29.43

Impact of training data volume

In our experiments, LANGCONN is trained on data from a single language, whereas FAMCONN is trained on pooled data from multiple languages within the same family. To validate that the improvements brought by FAMCONN are tied to linguistic relatedness rather than simply to a larger training data volume, we further train a universal connector, UNICONN, on all FLEURS languages pooled together using Gemma, as a complement to LANGCONN and FAMCONN.

Table 4 shows the family-level WER across these three connectors, while Table 8 (in Appendix A) provides detailed language level WERs. The results show that, across all families, FAMCONN consistently achieves lower WER than UNICONN, even though UNICONN is trained with the largest volume of data. At the per-language level, UNICONN generally falls between LANGCONN and FAMCONN and rarely surpasses the family connector. These results suggest that linguistic relatedness, rather than data volume, is the key factor in the gains observed with FAMCONN.

Extreme WERs

We also conduct an analysis of cases with extreme WERs. After manually reviewing the predicted transcripts, we observe a clear distinction between low WER and high WER cases. The extreme WERs mainly arise from repetition and overlong outputs. For instance, for LANGCONN on Polish in FLEURS with Salamandra, 94.08% of predictions contain at least one word or phrase repeated three or more times consecutively, and 45% of predictions have output lengths exceeding 1.25 times the reference length. For more detailed analysis results, please see Table 9 in Appendix A.

4 Conclusion

This study investigated grouping strategies for multilingual ASR, comparing family-level and language-specific connectors across two LLMs, two real-world multilingual datasets, and ten language families covering nearly forty languages. The findings demonstrate that family-level grouping consistently achieves lower and more stable WERs than language-specific alternatives. Although language-specific connectors occasionally deliver marginal improvements for individual languages, these gains are outweighed by frequent and sometimes catastrophic failures. Cross-domain testing further highlights that family-level sharing is particularly beneficial for multilingual ASR, where transfer across related languages enhances robustness. Overall, family-level grouping emerges as the more effective and reliable solution for multilingual ASR, striking a balance between transferability and specialization while avoiding the instability inherent in language-specific approaches.

5 Limitations and Ethical Statement

Limitations: Our work investigates speech connectors with two LLM backbones, Gemma-2_2b and Salamandra_2b, evaluated on the FLEURS and CommonVoice datasets covering nearly forty languages across ten language families. We have carefully limited our analysis to a maximum of five languages per family to ensure a manageable yet diverse experimental setup, enabling systematic comparisons without compromising generalizability. However, we acknowledge that while this choice allows us to control the grouping size and analyse a broad set of languages, it does not capture other potentially relevant dimensions such as script, morphological type, or subfamily structure. Future work could therefore consider typology-based or script-based groupings, or branch-level subdivisions within large families, to better understand how different forms of cross-lingual similarity influence connector design.

Ethical Statement: In this work, we have used publicly available models, architectures, and datasets and have not collected any sensitive/private data. The ultimate goal of our study is to contribute to analyzing the effect of language family on LLM-based ASR to improve multi-lingual ASR that is robust to domain shift. research

6 Acknowledgments

This work was supported by the ELOQUENCE project (grant number 101070558) funded by the UKRI and the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

References

Solomon Teferra Abate, Martha Yifiru Tachbelie, and Tanja Schultz. 2020. Multilingual acoustic and language modeling for ethio-semitic languages. In *Interspeech*, pages 1047–1051.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Ray Fabri, Michael Gasser, Nizar Habash, George Kiraz, and Shuly Wintner. 2014. Linguistic introduction: The orthography, morphology and syntax of semitic languages. In *Natural language processing of semitic languages*, pages 3–41. Springer.

Ruchao Fan, Bo Ren, Yuxuan Hu, Rui Zhao, Shujie Liu, and Jinyu Li. 2025. Alignformer: Modality matching can achieve better zero-shot instruction-following speech-llm. *IEEE Journal of Selected Topics in Signal Processing*.

Seraphina Fong, Marco Matassoni, and Alessio Brutti. 2025. Speech llms in low-resource scenarios: Data volume requirements and the impact of pretraining on high-resource languages. *arXiv preprint arXiv:2508.05149*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, and 1 others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

R Jairam, G Jyothish, and B Premjith. 2024. A few-shot multi-accented speech classification for indian languages using transformers and llm's fine-tuning approaches. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 1–9.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. 2021. Improved acoustic word embeddings for zero-resource languages using multilingual transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1107–1118.

Vishnupriya Kolipakam, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science*, 5(3):171504.

Shashi Kumar, Iuliia Thorbecke, Sergio Burdisso, Esaú Villatoro-Tello, Manjunath KE, Kadri Hacioğlu, Pradeep Rangappa, Petr Motlicek, Aravind Ganapathiraju, and Andreas Stolcke. 2025. Performance evaluation of slam-asr: The good, the bad, the ugly, and the way forward. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and 1 others. 2024. An embarrassingly simple approach for LLM with strong ASR capacity. *arXiv preprint arXiv:2402.08846*.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and 1 others. 2025. Speech recognition meets large language model: Benchmarking, models, and exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24840–24848.

Karel Mundnich, Xing Niu, Prashant Mathur, Srikanth Ronanki, Brady Houston, Veera Raghavendra Elluru, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Anshu Bhatia, and 1 others. 2025. Zero-resource speech translation and recognition with llms. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Tanja Schultz and Alex Waibel. 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1-2):31–51.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sébastien Bratières, and 1 others. 2024. How to connect speech foundation models and large language models? what matters and what does not. *arXiv preprint arXiv:2409.17044*.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and 1 others. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Hongfei Xue, Wei Ren, Xuelong Geng, Kun Wei, Longhao Li, Qijie Shao, Linju Yang, Kai Diao, and Lei Xie. 2024. Ideal-llm: Integrating dual encoders and language-adapted llm for multilingual speech-to-text. *arXiv preprint arXiv:2409.11214*.

A Appendix

In this appendix, we present detailed statistics of the training data and language-level experimental results.

Table 5 reports the total duration (in hours) of the training, validation, and test splits for each language in the FLEURS and CommonVoice datasets used in our experiments. It should be noted that for CommonVoice, the training data are capped at 100 hours per language.

Table 6 presents the full language-level WER results for RQ1, comparing family-level and language-specific connectors across both FLEURS and CommonVoice with two different LLM decoders.

Table 7 reports WER results for cross-domain evaluation using Gemma, where models are trained on one dataset and tested on the other.

Table 8 provides detailed language-level WER results using Gemma with three different types of connectors, LANGCONN, FAMCONN and UNICONN, to analyse impact of the training data volume.

Finally, to analyse error patterns in extreme cases, we select three low-WER and three high-WER examples and compute the repetition rate and overlong rate from the predicted transcriptions. Table 9 reports these metrics for the selected cases.

Table 5: Full dataset statistics in hours (FLEURS and CommonVoice). Train/validation/test splits are reported; CommonVoice values are capped at 100 hours.

Family (Subgroup)	Language	FLEURS			CommonVoice		
		Train	Val	Test	Train	Val	Test
Afro-Asiatic	Amharic	11.1	0.65	1.61	0.9	0.39	0.44
	Arabic	6.1	0.87	1.3	32.4	12.99	12.66
	Hausa	13.6	1.53	3.34	2.3	0.74	0.97
	Hebrew	9.5	0.82	2.05	1.2	0.88	0.58
	Maltese	9.9	1.49	3.55	2.4	2.1	2.35
Dravidian	Malayalam	10.1	1.68	3.91	1.39	1.02	1.03
	Tamil	8.7	1.25	2.13	83.7	19.3	19.4
	Telugu	7.9	0.89	1.45	0.08	0.08	0.08
Germanic (Indo-Euro)	Danish	7.5	1.17	2.94	4.2	3.42	3.56
	Dutch	7.7	0.45	0.97	54.2	16.0	16.4
	English	7.5	1.05	1.77	100	27.3	27.1
	German	9.0	1.26	3.15	100	27.5	27.6
	Swedish	8.4	0.98	2.33	9.1	6.1	6.9
Romance (Indo-Euro)	French	10.3	0.8	2.0	100	26.1	26.3
	Galician	6.7	1.05	2.6	96.5	18.9	19.8
	Portuguese	10.2	1.29	3.24	26.3	12.0	12.9
	Romanian	10.1	1.08	2.53	5.7	4.2	4.6
	Spanish	8.8	1.35	3.09	100	26.8	27.0
Slavic (Indo-Euro)	Belarusian	9.5	1.65	4.03	100	25.2	25.6
	Polish	9.2	0.84	2.05	35.5	14.2	14.1
	Russian	8.1	1.08	2.5	38.1	15.5	16.0
	Serbian	10.7	0.83	2.12	1.8	1.6	1.9
	Slovenian	7.8	0.89	2.27	1.5	1.5	1.7
Indo-Iranian (Indo-Euro)	Bengali	10.7	1.45	3.43	34.2	16.4	16.7
	Hindi	6.7	0.71	1.34	5.8	3.6	4.7
	Persian	12.1	1.53	3.7	31.6	12.6	14.7
	Punjabi	6.4	0.75	1.84	1.2	0.49	0.74
	Urdu	7.0	0.76	0.81	8.4	6.3	6.5
Baltic (Indo-Euro)	Latvian	6.5	1.12	2.84	23.3	12.3	12.3
	Lithuanian	9.8	1.17	2.97	11.8	7.1	7.8
Celtic (Indo-Euro)	Irish	12.1	1.49	3.46	0.6	0.58	0.64
	Welsh	12.2	1.8	4.28	11.5	8.3	8.3
Niger-Congo	Igbo	13.8	1.9	4.81	0.01	0.0	0.01
	Swahili	13.5	0.8	1.93	69.9	19.2	19.0
	Yoruba	10.0	1.71	3.77	2.4	1.35	1.91
Turkic	Azerbaijani	9.3	1.35	3.24	0.23	0.1	0.16
	Kazakh	11.8	1.53	3.83	0.81	0.67	0.75
	Kyrgyz	9.3	1.34	3.25	2.3	2.1	2.2
	Turkish	8.3	1.12	2.61	43.8	12.1	14.1

Table 6: Detailed WER↓ (%) per language on FLEURS and CommonVoice datasets with both LLMs. Δ_f = Fam – Lang. Best in **bold**

Family	Language	FLEURS				CommonVoice			
		Salamandra_2b		Gemma-2_2b		Salamandra_2b		Gemma-2_2b	
		Lang	Fam (Δ_f)	Lang	Fam (Δ_f)	Lang	Fam (Δ_f)	Lang	Fam (Δ_f)
Afro-Asiatic	Amharic	**97.45**	98.74 (+1.29)	**52.29**	52.89 (+0.60)	**111.53**	130.68 (+19.15)	**137.32**	150.75 (+13.43)
	Arabic	**38.95**	53.66 (+14.71)	**29.83**	30.17 (+0.34)	**50.69**	57.65 (+6.96)	**60.81**	64.34 (+3.53)
	Hausa	**55.72**	101.22 (+45.50)	**51.32**	57.48 (+6.16)	**97.57**	110.61 (+13.04)	108.88	**98.80** (-10.08)
	Hebrew	**54.58**	71.25 (+16.67)	**42.90**	46.59 (+3.69)	**52.77**	53.22 (+0.45)	77.62	**65.35** (-12.27)
	Maltese	**107.49**	111.56 (+4.07)	**40.37**	42.41 (+2.04)	**131.25**	169.01 (+37.76)	116.29	**92.25** (-24.04)
Baltic	Latvian	115.47	**36.30** (-79.17)	28.27	**24.63** (-3.64)	137.97	**77.39** (-60.58)	56.64	**45.98** (-10.66)
	Lithuanian	110.76	**42.29** (-68.47)	34.56	**30.18** (-4.38)	110.80	**103.10** (-7.70)	**59.01**	70.52 (+11.51)
Celtic	Irish	122.39	**117.09** (-5.30)	71.54	**70.99** (-0.55)	302.18	**219.56** (-82.62)	121.59	**107.64** (-13.95)
	Welsh	115.08	**110.97** (-4.11)	38.74	**38.51** (-0.23)	**125.94**	159.53 (+33.59)	119.93	**84.89** (-35.04)
Dravidian	Malayalam	**23.65**	44.79 (+21.14)	**21.85**	22.95 (+1.10)	112.78	**84.91** (-27.87)	**71.72**	96.59 (+24.87)
	Tamil	33.93	**33.63** (-0.30)	32.24	**31.92** (-0.32)	124.42	**54.11** (-70.31)	49.67	**68.50** (-18.83)
	Telugu	35.45	**34.60** (-0.85)	**34.34**	35.16 (+0.82)	140.21	**111.64** (-28.57)	285.65	**176.63** (-109.02)
Germanic	Danish	32.12	**20.51** (-11.61)	24.33	**19.24** (-5.09)	69.38	**43.28** (-26.10)	44.62	**33.42** (-11.20)
	Dutch	15.76	**13.84** (-1.92)	15.33	**12.62** (-2.71)	104.41	**30.24** (-74.17)	59.33	**21.20** (-38.13)
	English	13.06	**9.37** (-3.69)	8.59	**6.79** (-1.80)	53.96	**32.58** (-21.38)	37.12	**23.18** (-13.94)
	German	22.56	**13.62** (-8.94)	17.95	**13.34** (-4.61)	97.67	**27.62** (-70.05)	37.12	**24.75** (-12.37)
	Swedish	23.11	**18.75** (-4.36)	20.61	**16.56** (-4.05)	46.58	**33.93** (-12.65)	39.05	**27.61** (-11.44)
Indo-Iranian	Bengali	**41.47**	52.17 (+10.7)	23.95	**23.64** (-0.31)	**62.68**	101.19 (+38.51)	**101.20**	101.65 (+0.45)
	Hindi	47.53	**36.92** (-10.61)	21.45	**19.57** (-1.88)	64.67	**35.35** (-29.32)	72.20	**42.99** (-29.21)
	Persian	**37.56**	46.01 (+8.45)	**25.12**	30.88 (+5.76)	**66.85**	78.97 (+12.12)	**138.66**	154.45 (+15.79)
	Punjabi	46.85	**45.67** (-1.18)	26.71	**25.34** (-1.37)	83.76	**64.62** (-19.14)	115.23	**48.21** (-67.02)
	Urdu	81.18	**54.46** (-26.72)	85.45	**33.71** (-51.74)	71.21	**49.51** (-21.70)	69.62	**64.44** (-5.18)
Niger-Congo	Igbo	61.72	**60.55** (-1.17)	**55.25**	56.69 (+1.44)	**100.00**	120.13 (+20.13)	**100.00**	108.51 (+8.51)
	Swahili	**28.87**	31.63 (+2.76)	**24.91**	28.53 (+3.62)	482.60	**81.12** (-401.48)	82.16	**78.86** (-3.30)
	Yoruba	52.24	**52.07** (-0.17)	**45.51**	46.63 (+1.12)	**88.30**	90.50 (+2.20)	**75.90**	100.84 (+24.94)
Romance	French	13.90	**11.03** (-2.87)	13.29	**10.80** (-2.49)	52.25	**18.96** (-33.29)	35.88	**26.09** (-9.79)
	Galician	102.36	**14.23** (-88.13)	16.74	**14.58** (-2.16)	61.56	**32.67** (-28.89)	**43.03**	43.70 (+0.67)
	Portuguese	10.27	**7.43** (-2.84)	8.74	**7.73** (-1.01)	25.05	**12.29** (-12.76)	18.31	**14.09** (-4.22)
	Romanian	47.85	**18.03** (-29.82)	17.02	**14.89** (-2.13)	93.31	**26.79** (-66.52)	**30.04**	29.20 (-0.84)
	Spanish	8.34	**6.66** (-1.68)	8.86	**5.50** (-3.36)	49.59	**26.03** (-23.56)	**29.64**	30.10 (+0.46)
Slavic	Belarusian	100.82	**30.13** (-70.69)	25.51	**24.13** (-1.38)	183.00	**62.35** (-120.65)	60.21	**51.05** (-9.16)
	Polish	119.71	**16.71** (-103.00)	17.86	**16.35** (-1.51)	376.28	**33.73** (-342.55)	42.05	**32.48** (-9.57)
	Russian	108.17	**12.46** (-95.71)	14.62	**13.11** (-1.51)	158.95	**42.50** (-116.45)	**42.77**	43.88 (+1.11)
	Serbian	33.99	**30.58** (-3.41)	30.31	**27.91** (-2.40)	63.18	**33.15** (-30.03)	34.50	**31.69** (-2.81)
	Slovenian	109.02	**30.49** (-78.53)	30.50	**25.83** (-4.67)	92.37	**49.02** (-43.35)	45.13	**47.60** (+2.47)
Turkic	Azerbaijani	104.03	**47.22** (-56.81)	27.51	**26.23** (-1.28)	145.09	**79.00** (-66.09)	143.74	**85.73** (-58.01)
	Turkish	**21.55**	38.98 (+17.43)	20.21	**18.51** (-1.70)	62.75	**55.35** (-7.40)	68.74	**57.49** (-11.25)
	Kazakh	66.30	**39.26** (-27.04)	**21.12**	21.16 (+0.04)	**62.01**	69.48 (+7.47)	**52.10**	64.78 (+12.68)
	Kyrgyz	**36.82**	53.47 (+16.65)	30.82	**30.33** (-0.49)	**68.69**	87.49 (+18.80)	91.40	**88.64** (-2.76)

Table 7: Cross-domain WER↓ (%) per language using Gemma. Δ_f = Fam – Lang. Best in **bold**

Family	Language	CV → FL		FL → CV	
		Lang	Fam (Δ_f)	Lang	Fam (Δ_f)
Afro-Asiatic	Amharic	**119.70**	154.38 (+34.68)	**91.23**	95.59 (+4.36)
	Arabic	93.60	**78.01** (-15.59)	116.22	**76.51** (-39.71)
	Hausa	202.45	**180.67** (-21.78)	**65.79**	80.19 (+14.40)
	Hebrew	**92.47**	101.35 (+8.88)	**44.53**	50.93 (+6.40)
	Maltese	194.60	**150.86** (-43.74)	97.87	**73.74** (-24.13)
Baltic	Latvian	**127.89**	128.01 (+0.12)	91.65	**54.89** (-36.76)
	Lithuanian	**98.04**	122.45 (+24.41)	42.49	**40.31** (-2.18)
Celtic	Irish	210.63	**172.54** (-38.09)	**109.72**	131.11 (+21.39)
	Welsh	**120.70**	155.43 (+34.73)	102.27	**81.35** (-20.92)
Dravidian	Malayalam	**104.26**	113.49 (+9.23)	**50.56**	74.78 (+24.22)
	Tamil	**89.59**	117.97 (+28.38)	120.75	**67.94** (-52.81)
	Telugu	**120.07**	164.27 (+44.20)	**73.37**	111.97 (+38.60)
Germanic	Danish	121.21	**62.57** (-58.64)	34.67	**24.63** (-10.04)
	Dutch	89.28	**48.38** (-40.90)	21.12	**19.64** (-1.48)
	English	65.07	**28.39** (-36.68)	23.59	**17.30** (-6.29)
	German	108.71	**61.84** (-46.87)	23.57	**22.18** (-1.39)
	Swedish	212.82	**60.42** (-152.40)	34.00	**28.07** (-5.93)
Indo-Iranian	Bengali	119.81	**103.87** (-15.94)	**40.16**	41.89 (+1.73)
	Hindi	110.46	**88.96** (-21.50)	**49.34**	59.05 (+9.71)
	Persian	214.23	**186.80** (-27.43)	142.85	**119.56** (-23.29)
	Punjabi	**78.34**	139.35 (+61.01)	**37.81**	56.84 (+19.03)
	Urdu	94.79	**89.74** (-5.05)	152.09	**134.70** (-17.39)
Niger-Congo	Igbo	**110.32**	125.20 (+14.88)	**75.97**	97.39 (+21.42)
	Swahili	199.47	**187.92** (-11.55)	**56.27**	94.87 (+38.60)
	Yoruba	**92.67**	95.23 (+2.56)	**67.33**	80.43 (+13.10)
Romance	French	106.50	**58.13** (-48.37)	41.50	**35.73** (-5.77)
	Galician	86.53	**81.75** (-4.78)	**26.57**	29.00 (+2.43)
	Portuguese	82.82	**56.48** (-26.34)	15.89	**12.25** (-3.64)
	Romanian	120.29	**76.23** (-44.06)	29.77	**29.34** (-0.43)
	Spanish	79.45	**44.99** (-34.46)	22.05	**17.81** (-4.24)
Slavic	Belarusian	**97.15**	108.45 (+11.30)	**35.48**	36.42 (+0.94)
	Polish	105.60	**63.44** (-42.16)	29.51	**25.71** (-3.80)
	Russian	**90.24**	98.83 (+8.59)	24.40	**22.61** (-1.79)
	Serbian	149.45	**116.11** (-33.34)	278.90	**117.43** (-161.47)
	Slovenian	113.94	**80.06** (-33.88)	79.51	**51.00** (-28.51)
Turkic	Azerbaijani	**108.60**	118.26 (+9.66)	47.11	**43.61** (-3.50)
	Turkish	131.39	**113.38** (-18.01)	40.78	**37.12** (-3.66)
	Kazakh	**108.74**	146.46 (+37.72)	65.02	**61.34** (-3.68)
	Kyrgyz	**106.19**	109.15 (+2.96)	**64.54**	66.95 (+2.41)

Table 8: Detailed WER↓ (%) per language using Gemma across LANGCONN, FAMCONN and UNICONN. Best in **bold**.

Family	Language	FLEURS		
		Lang	**Fam**	**Uni**
Afro-Asiatic	Amharic	**52.29**	52.89	62.33
	Arabic	29.83	30.17	**25.58**
	Hausa	**51.32**	57.48	78.79
	Hebrew	**42.90**	46.59	53.02
	Maltese	**40.37**	42.41	51.15
Baltic	Latvian	28.27	**24.63**	25.22
	Lithuanian	34.56	**30.18**	34.16
Celtic	Irish	71.54	**70.99**	84.97
	Welsh	38.74	**38.51**	45.53
Dravidian	Malayalam	**21.85**	22.95	33.57
	Tamil	32.24	**31.92**	40.70
	Telugu	**34.34**	35.16	50.77
Germanic	Danish	24.33	**19.24**	21.45
	Dutch	15.33	**12.62**	13.34
	English	8.59	**6.79**	6.96
	German	17.95	**13.34**	13.51
	Swedish	20.61	**16.56**	18.83
Indo-Iranian	Bengali	23.95	**23.64**	32.05
	Hindi	21.4	**19.57**	24.99
	Persian	**25.12**	30.88	27.63
	Punjabi	26.71	**25.34**	31.68
	Urdu	85.45	**33.71**	110.27
Niger-Congo	Igbo	**55.25**	56.69	80.69
	Swahili	**24.91**	28.53	31.14
	Yoruba	**45.51**	46.63	55.14
Romance	French	13.29	**10.80**	11.55
	Galician	16.74	**14.58**	17.60
	Portuguese	8.74	7.73	**7.39**
	Romanian	17.02	**14.89**	15.64
	Spanish	8.86	**5.50**	6.25
Slavic	Belarusian	**25.51**	24.13	29.03
	Polish	**17.86**	16.35	17.69
	Russian	14.62	**13.11**	13.39
	Serbian	30.31	**27.91**	30.61
	Slovenian	30.50	**25.83**	28.51
Turkic	Azerbaijani	27.51	**26.23**	31.99
	Turkish	20.21	**18.51**	18.85
	Kazakh	**21.12**	21.16	26.23
	Kyrgyz	30.82	**30.33**	38.27

Table 9: Repeat rate and overlong rate of low-WER and high-WER cases. Repeat Rate:the proportion of predictions in which any word or phrase appears three or more times consecutively. Overlong Rate:the proportion of predictions where the output length exceeds 1.25 times the label length.

Language	**Setting**	**WER**	**Repeat Rate**	**Overlong Rate**
English	Fam_FL_Gemma	6.79	0.01	0.02
Portuguese	Lang_CV_Salamandra	15.72	0.01	0.01
Hindi	Fam_CV_Salamandra	35.35	0.02	0.06
Polish	Lang_FL_Salamandra	119.71	0.94	0.45
Welsh	Fam_CV_Gemma	159.53	0.61	0.35
Swahili	Lang_CV_Salamandra	482.6	0.87	0.64