Does Regional Decoder Specialization Help Low-Resource ASR Based on the SLAM-ASR Framework?

Edwin Arkel Rios[1,2*], **Jocelyn Zaruma**[3,4], **Girish Ewoorkar**[5], **Sneh Sourabh**[6], **Julian Mack**[7], **Hung-Hui Juan**[2], **Stephen Huang**[2], **Bo-Cheng Lai**[1]
[1]National Yang Ming Chiao Tung University, Taiwan [2]Tranxform AI
[3]Institut national des langues et civilisations orientales, France
[4]Université Sorbonne Nouvelle, France [5]Cohere Labs Community
[6]Indian Institute of Technology Hyderabad, India [7]Cohere

Abstract

The SLAM-ASR recipe — a frozen speech encoder, a frozen large language model (LLM) as the decoder, and a small trainable projector between them — is a practical way to add the speech modality to LLMs. It is being explored to include low-resource languages in audio LLMs. However, the choice of decoder within it has not been studied under controlled conditions. We ask whether pairing each language with a *regionally specialized* decoder improves recognition, using the Tiny Aya family, whose four variants share one base model and one tokenizer and differ only in the regional weighting of their post-training data. Across eleven languages, the region-matched decoder is only 0.55 character error rate (CER) better on average than mismatched regional decoders (Wilcoxon signed-rank $p = 0.97$). The matched decoder shows no advantage on held-out test sets either, where the difference is -0.09 CER against a minimum detectable effect of 0.67. The advantage also does not grow with the amount of target-language data the decoder saw. For the median language, the matched variant saw only 1.3 percentage points more target-language data than the mismatched variants, and the largest difference is 5.7 points. Although this is up to 236 times more in relative terms, having more target-language data does not result in higher recognition accuracy. What does matter is training the projector: on low-resource languages, it improves CER over off-the-shelf models by up to 102.6, including a language supported by neither frozen component. For practitioners, a general-purpose multilingual decoder suffices, and the effort is better spent on in-language audio. We release all checkpoints and transcripts.[1]

Keywords: speech recognition, multilingual, low-resource languages, speech language model

[*]Work done during an internship at Tranxform AI.
[1]huggingface.co/ERISLab/ListAya

1 Introduction

Automatic speech recognition (ASR) models such as Whisper (Radford et al., 2023) and Omnilingual ASR (Team et al., 2025) have expanded multilingual speech recognition to a growing number of languages. Despite this progress, substantial work remains on the breadth and diversity of low-resource languages and dialects, which has motivated recent studies of the failure modes of ASR systems in low-resource settings (Imam et al., 2025b; Kyslyi et al., 2026; Kumar et al., 2026; B et al., 2026; Azis et al., 2026).

In parallel, large language models (LLMs) are being given a speech interface for low-resource languages. One approach is the SLAM-ASR architecture (Ma et al., 2024), which connects a pretrained speech encoder to an LLM through a trainable linear projector while keeping both pretrained models frozen. Follow-up work has studied the recipe's robustness (Kumar et al., 2025), its data requirements (Fong et al., 2025), encoder depth (Kolluri et al., 2026), language-family effects (Zhang et al., 2026), and prompt sensitivity (Burdisso et al., 2026). However, the effect of the decoder's *regional specialization* has not been studied. The Tiny Aya family (Salamanca et al., 2026) makes this experiment possible: its small multilingual LLMs share a common base model and tokenizer, and its variants differ only in the regional weighting of their post-training data.

We ask three questions. **RQ1:** Does pairing a language with its region-matched decoder improve recognition over mismatched regional decoders or a general-purpose one? **RQ2:** If there is an effect, does it scale with the amount of target-language data in the decoder's post-training mix, rather than with the region label? **RQ3:** Independent of the decoder choice, what advantages does training the projector bring over off-the-shelf models on low-

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 225–235
November 6-7, 2026 ©2026 Association for Computational Linguistics

resource languages?

Our contributions, over twelve languages:

- The first controlled test of decoder regional specialization in SLAM-ASR (RQ1). Matching a language to its regional decoder does not measurably improve recognition. The result holds on held-out test sets, where any effect is smaller than 0.67 CER — the smallest difference the design can detect — and is supported by a run-to-run noise floor measured from replicate runs.
- A measurement of target-language exposure across decoders (RQ2). Matched decoders saw up to 236 times the target-language data of mismatched ones, yet greater exposure does not correspond to higher recognition accuracy.
- A practical positive result (RQ3). Training the projector improves over off-the-shelf models by up to 102.6 CER precisely on low-resource languages, including one that neither the speech encoder nor the decoder supports.
- We release all checkpoints and evaluation transcripts, so that the comparison can be extended and the released models can serve as baselines for further work on low-resource ASR within the SLAM-ASR framework.

2 Related Work

Projector-based ASR. Ma et al. (2024) showed that a single trainable linear projector between a frozen speech encoder and a frozen LLM reaches competitive English recognition, and subsequent work has studied its robustness (Kumar et al., 2025), how much labeled audio it needs (Fong et al., 2025), how far the encoder can be pruned (Kolluri et al., 2026), how language-family structure affects it (Zhang et al., 2026), and its sensitivity to the training prompt (Burdisso et al., 2026). In these studies the decoder is either fixed or replaced by a model that differs in more than its training data.

Decoder selection. No prior work isolates the effect of the decoder's training data. Ma et al. (2024) ablate the decoder's scale (TinyLlama-1.1B against LLaMA-7B), base against instruction-tuned checkpoints, and family (Phi-2, LLaMA, Vicuna) on English. Fong et al. (2025) compare EuroLLM and Salamandra decoders in low-resource settings and find a large gap, 14.0 against 33.6 word error rate at ten hours of Italian, that narrows as the training data increases. These results show the decoder substantially affects recognition, but the models compared differ in scale, tokenizer and pretraining corpus at once. The Tiny Aya family isolates the remaining variable: its variants share a base model and a tokenizer and differ only in the regional weighting of their post-training mix.

Low-resource ASR. Surveys of African ASR describe a persistent shortage of transcribed audio and of evaluation resources (Imam et al., 2025a,b), recent work explores using ASR itself to bootstrap corpora for languages such as Fongbe and Hausa (Adjovi et al., 2026), and practitioner accounts report the same constraint (Klejch et al., 2025). Several of our study languages, including Marathi, Urdu, Hausa, Amharic and Kreol Seselwa, do not appear in the SLAM-ASR literature surveyed above, and Kreol Seselwa appears unstudied in this setting altogether. The checkpoints we release provide baselines for them. The projector-only recipe suits this setting because it requires no encoder training, no decoder training, and only a component small enough to train on one GPU.

3 Method

Our implementation follows the SLAM-ASR architecture (Ma et al., 2024), illustrated in Figure 1, which connects a pretrained speech encoder to a pretrained LLM serving as the decoder, mapping speech representations into the LLM's embedding space through a simple linear projector. During training, the encoder and decoder remain frozen; only the projector's parameters are updated, learning speech-text alignment through the ASR task. The decoder then autoregressively generates the transcription.

Decoder specialization. We use the Tiny Aya family (Salamanca et al., 2026) as the frozen decoder: four instruction-tuned variants of the same 3.35B-parameter multilingual base model, sharing one tokenizer, so that swapping one variant for another changes only the composition of its post-training data. Tiny Aya Global is post-trained across all regions, while each regional variant reweights the mix toward one cluster of regions: Earth toward Europe, West Asia and Africa; Fire toward South Asia; and Water toward Europe, West Asia and Asia-Pacific. We assign each study language the region of its cluster in the Tiny Aya re-

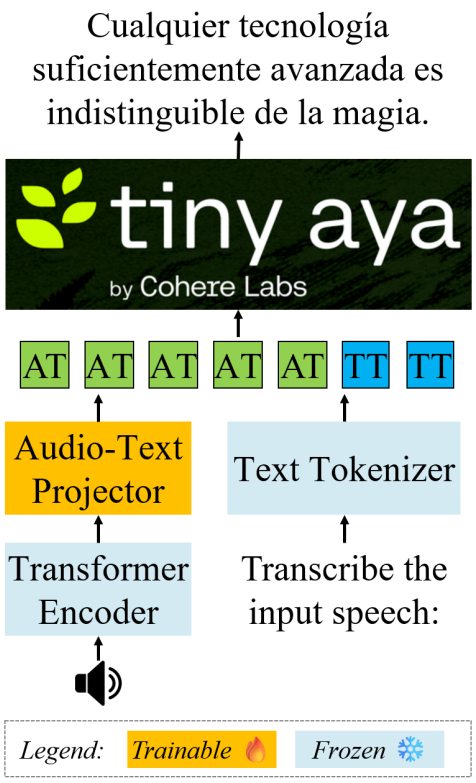

Figure 1: LisTAya (Listening Tiny Aya), our model family built on the SLAM-ASR architecture: a frozen Whisper-Medium encoder, a trainable audio–text projector, and a frozen Tiny Aya decoder that generates the transcription from the audio tokens (AT) and the prompt's text tokens (TT). Only the projector is trained.

port (Table 1). The variant of that region is the language's *matched* decoder, and the other two regional variants are its *mismatched* decoders. We call the share of a language in a variant's post-training data the variant's *exposure* to it. English is the one language whose matched variant has substantially less exposure than a mismatched one: the Fire mix contains 46.2% English against Water's 17.0% (Section 5.2). This design is what makes the comparison causal: across all experiments the encoder and projector recipe are identical, and only the decoder variant changes. We call the resulting individual models Earth, Fire, Water, and Global. We refer to this family of models as LisTAya (**Lis**tening **T**iny **A**ya).

4 Experimental Setup

Implementation. Our SLAM-ASR implementation builds on the Qwen2-Audio implementation (Chu et al., 2024), from which we adopt the temporal downsampler, a stride-2 average-pooling layer that reduces the encoder output from 50 to 25 frames per second. We use only the encoder of Whisper-Medium (Radford et al., 2023), with 307M parameters, as the speech encoder, chosen for its accuracy–speed tradeoff (Ma et al., 2024). The four Tiny Aya decoders described in Section 3 each have 3.35B parameters. The encoder output dimension is 1024, and the decoder embedding dimension is 2048.

Datasets and languages. We train on WorldSpeech (Asonitis et al., 2026), a corpus of formal speech from parliamentary proceedings, international broadcasts and public-domain audiobooks. We exclude samples of 30 seconds or longer. We cover twelve languages spanning the Water, Fire, and Earth regions, plus Kreol Seselwa (referred to as Seselwa throughout). We call a language as spoken in one country, a WorldSpeech configuration such as en_ke, a *dialect* of that language. For each language, Table 1 gives the Tiny Aya region of its matched decoder variant and the dialect used for training. Seselwa is supported by neither the encoder nor any Tiny Aya decoder, making it a test case for what the projector alone can learn. To compare model behavior across amounts of training data, we group the languages into *resource tiers* by unique training-audio hours: Very Low (<50 h), Low (50–100 h), Mid (100–300 h), and High (>300 h). Hours are estimated from the training logs, except for Marathi, derived directly from the corpus duration column (216 h). We use the FLEURS (Conneau et al., 2023) validation splits as development sets for ten languages. Hausa trains on two dialects, Nigerian (ha_ng) and Chadian (ha_td). We use the ha_ng test split as its development set, because its FLEURS validation split is impractically large for repeated CER evaluation during training, and test on ha_td. Seselwa is not in FLEURS, so we use a held-out 2% split of its WorldSpeech data as the development set. For final evaluation, we use held-out FLEURS test splits for *out-of-domain* evaluation and WorldSpeech test splits for *in-domain* evaluation.

Training. Every run trains only the projector (2.1M parameters) with AdamW at learning rate 1.5×10^{-3}, batch size 8 with gradient accumulation 64 (effective batch 512), 1,000 steps, no weight decay. We use the prompt "Transcribe the input speech:". Transcriptions are decoded greedily (beam size 1, no sampling), with a maximum of

Language	Region	WS dialect	Tier	Train h
English	Water	en_us	High	≥2708
Spanish	Water	es_mx	High	969
French	Water	fr_ca	High	1028
Indonesian	Water	id_id	High	435
Hindi	Fire	hi_in	High	≥1939
Marathi	Fire	mr_in	Mid	216
Tamil	Fire	ta_in	Very Low	35
Urdu	Fire	ur_pk+ur_in	Low	85
Amharic	Earth	am_et	Very Low	38
Hausa	Earth	ha_ng+ha_td	Mid	126
Swahili	Earth	sw_ke+sw_tz	High	1050
Seselwa	–	crs_sc	High	1510

Table 1: Training configuration (WS = WorldSpeech; + joins two training sets): region assignment from the Tiny Aya report (– for Seselwa, which it does not cover), resource tier and unique training-audio hours per language. ≥ in Train h marks the audio seen in the 1,000 training steps, for languages whose data the run did not exhaust.

200 new tokens. During training, CER is evaluated on the development set every 10 steps. Checkpoints are saved every 50 steps and selected based on the lowest validation loss, then used for final evaluation. The experimental grid contains 12 languages $\times$ 4 decoders $=$ 48 language–decoder configurations. Amharic and Seselwa were trained twice with different seeds across all four decoders, giving eight *replicate pairs* from which we estimate the run-to-run noise floor directly. A run takes a median of 15.2 h on one NVIDIA H200 GPU.

Baselines and metrics. We compare our LisTAya models against three off-the-shelf models evaluated on the same test sets: Whisper-Medium (Radford et al., 2023), Voxtral-Mini-3B (Liu et al., 2025) and Qwen2-Audio-7B (Chu et al., 2024). CER is the primary metric throughout, and per-language error rates are never pooled across languages. A run's *best CER* is its lowest CER on the development set during training.

Statistical analysis. We compare the matched and mismatched decoders over the per-language CER differences with the Wilcoxon signed-rank test (Wilcoxon, 1945), a non-parametric paired test that does not assume the differences are normally distributed, and report the 95% bootstrap confidence interval of their mean. The *minimum detectable effect* is the smallest true difference that a paired test at the 5% significance level detects with 80% power, given how much the per-language differences vary. The run-to-run noise floor is the between-run standard deviation of the best CER over the eight replicate pairs. Relations between two quantities are measured with the Spearman rank correlation ρ and the Pearson correlation r. We use $p < 0.05$ as the threshold for statistical significance.

5 Results

5.1 The matched decoder shows no advantage on the development set

To answer RQ1, Table 2 reports, for each language, the best CER of the region-matched decoder against the mean of the two mismatched regional decoders and against Global. Across the eleven languages covered by a Tiny Aya region, the matched decoder is 0.55 CER better on average, with a 95% bootstrap confidence interval of $[-3.89, 1.99]$. A Wilcoxon signed-rank test over the eleven per-language differences finds no significant difference ($p = 0.97$). The matched decoder has the lower CER in six of eleven languages. Against Global, the general-purpose variant, it is 0.50 CER better on average ($p = 0.97$). Figure 2 shows the CER learning curves, evaluated every 10 training steps. Within each language the four variants give nearly the same curve, while error rates span two orders of magnitude across languages.

We compare this result with two reference values. The minimum detectable effect is 4.90 CER, and the between-run standard deviation is 0.89. The observed 0.55 CER advantage is below both.

5.2 Differences in target-language exposure

RQ2 asks whether the effect scales with exposure. Because no advantage was found, we also check how large the difference in exposure between decoders was, using the Tiny Aya report's own composition tables (Table 3). In absolute terms the difference is small: the matched variant saw a median of 1.3 percentage points more of the target language than the mismatched variants did. In relative terms it is not: the matched variant saw 236 times more Amharic, Hausa or Swahili than the mismatched ones, and 204 times more Tamil, with six of eleven languages above a factor of twenty. The report rounds shares to one decimal place, so for four languages the mismatched share is below 0.05% and is recovered from the printed per-region subtotals. The ten African languages sum to 0.1% of the Fire mix (0.010% apiece). South Asia sums

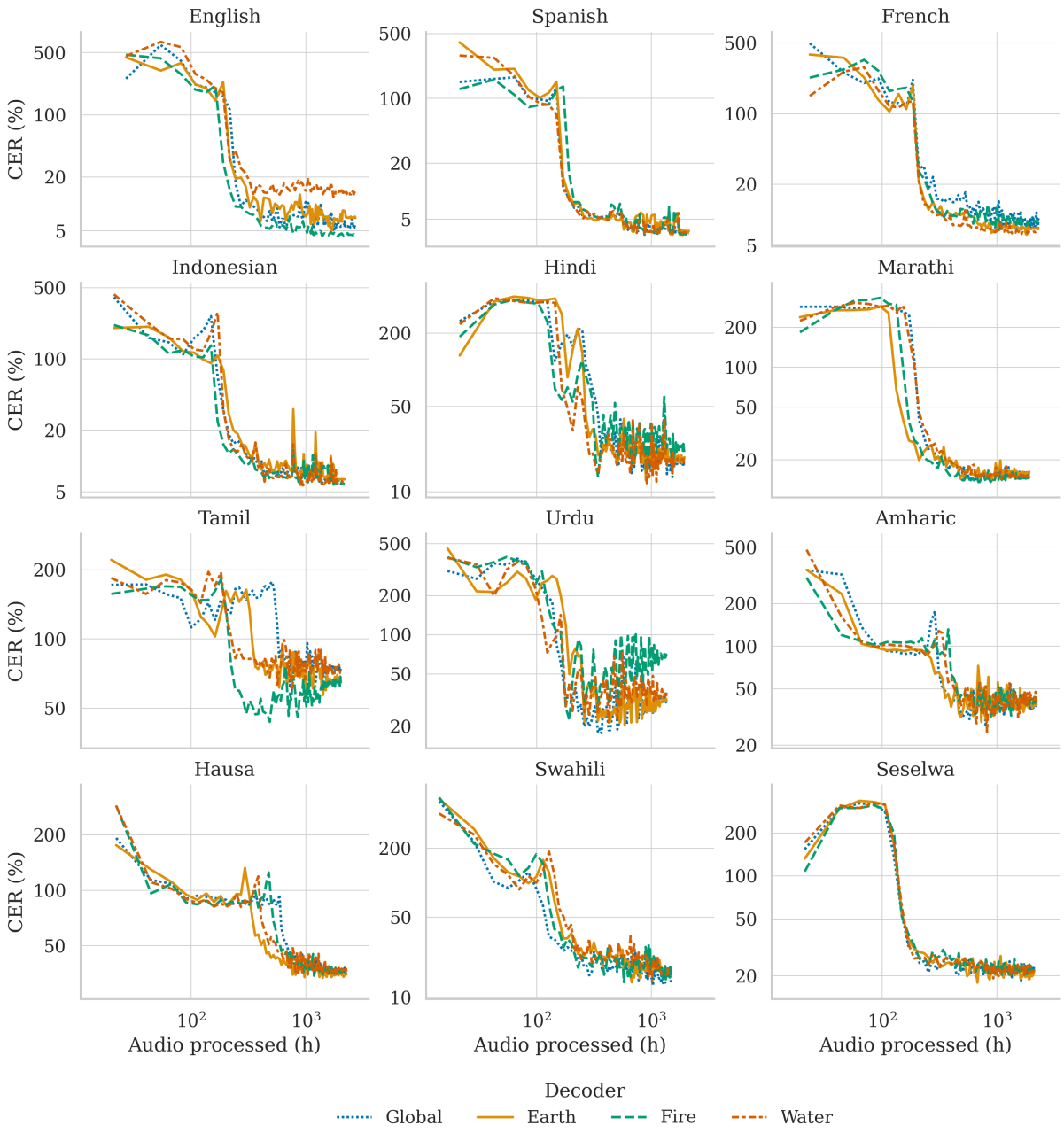

Figure 2: CER on the development set against processed audio, both on logarithmic axes: one panel per language, one line per decoder variant, distinguished by color and line style.

to 0.3% of Water against 0.1% each for Hindi and Marathi, leaving 0.017% apiece for the remaining six.

Neither measure of exposure is significantly correlated with the matched decoder's advantage, the Δ of Table 2. Across the eleven languages, the difference in percentage points gives Spearman $\rho = -0.29$ ($p = 0.38$) and Pearson $r = -0.53$ ($p = 0.09$). The log exposure ratio, which varies by a factor of roughly 400 across these languages, gives $\rho = -0.39$ ($p = 0.24$) and $r = -0.51$ ($p = 0.11$). With the advantage measured on the in-domain test instead, ρ with the log ratio is -0.04.

English is a Water language, but the Fire mix contains 46.2% English against Water's 17.0%, so English is the one language whose matched decoder saw substantially less of it. It is also the language where the matched decoder is worse than the mismatched decoders, which saw more English, by the largest margin (+7.39 CER).

Language	Earth	Fire	Water	Global	Δ
English	4.98	4.35	**12.05**	4.96	7.39
Spanish	3.37	3.32	**3.41**	3.40	0.07
French	6.37	7.21	**6.36**	6.61	-0.43
Indonesian	6.42	5.85	**5.74**	6.05	-0.40
Hindi	12.36	**13.11**	11.74	13.18	1.06
Marathi	13.81	**13.57**	14.21	14.14	-0.44
Tamil	56.76	**43.63**	59.89	61.21	-14.70
Urdu	20.28	**22.24**	21.33	17.64	1.44
Amharic	**29.25**	31.01	24.85	27.00	1.32
Hausa	**33.43**	34.37	34.86	35.03	-1.19
Swahili	**13.53**	13.59	13.76	12.60	-0.15

Table 2: Best CER per language and decoder variant, defined as the lowest CER on the development set during training. The matched variant is in bold. Δ is the matched CER minus the mean of the other two regional variants. Negative values favor the matched decoder. Languages are grouped by their Tiny Aya region, as in Table 1.

5.3 The matched decoder shows no advantage on held-out test sets either

Section 5.1 compares the matched and mismatched decoders on the development set, the data each checkpoint was chosen on. Here we repeat that comparison on the held-out test sets. On FLEURS test the matched decoder is 0.70 CER better on average ($p = 0.83$); on the in-domain WorldSpeech test it is 1.71 CER worse ($p = 0.70$). Urdu's in-domain test is the one test set where every decoder has a very high CER (Table 5). The lowest value among the matched decoder, the mean of the two mismatched decoders and Global is 75.3 CER, against 36.6 for the next-worst language. Excluding Urdu gives -0.02 CER over ten languages, with a minimum detectable effect of 0.44. Restricting instead to the seven languages whose exposure differs most between decoders gives -0.09 CER ($p = 0.69$) on the in-domain test, against a minimum detectable effect of **0.67 CER**. The other four, English, Spanish, French and Urdu, appear in every regional mix, with exposure ratios of 0.53–2.72 against 20 or more for the seven.

The sign of the mean difference changes across evaluation sets (-0.55 on the development set, -0.70 on FLEURS test, $+1.71$ on the in-domain test), which is consistent with noise. The two largest per-language differences are specific to FLEURS rather than to the decoder. For Tamil the matched decoder is 14.70 CER better on the development set and 19.58 better on FLEURS test, but on its own in-domain test the matched-minus-mismatched difference is $+0.02$. For English that difference is $+7.39$ on the development set and $+0.19$ on the in-domain test.

5.4 Overfitting decreases with the amount of training audio

The runs do not differ measurably by decoder variant, but they differ by the amount of in-language audio. Table 4 groups the runs by resource tier. We measure overfitting as the rise in evaluation loss after its own minimum, without reference to the training distribution. The rise falls monotonically from 0.211 in the very-low tier to 0.003 in the high tier, and the lowest-resource cells reach their best evaluation loss less than a third of the way into training. Low-resource cells need early stopping; high-resource cells barely overfit at all.

5.5 Advantages of projector training over off-the-shelf models

For RQ3, Table 5 reports, for each held-out test set, the lowest CER among the three off-the-shelf models and the lowest CER among the language's four trained LisTAya models, one per decoder variant. Training lowers CER most where the best off-the-shelf model has a high CER. The change in CER from training is negatively correlated with the best baseline's CER (Spearman $\rho = -0.42$, $p = 0.005$ over all 43 test sets, the 23 of Table 5 and the 20 dialects of Table 6). Splitting at the median baseline CER of 15.8, the trained model has the lower CER on 12 of the 22 test sets at or above it and on 3 of the 21 below. Below the median the changes

Language	Earth	Fire	Water	Global	Ratio	Δ CER
English	18.1	46.2	**17.0**	13.8	0.5	7.39
Spanish	2.2	0.2	**2.1**	1.7	1.8	0.07
French	2.3	0.2	**2.2**	1.8	1.8	-0.43
Indonesian	0.1	0.1	**2.0**	1.6	20.0	-0.40
Hindi	0.1	**5.8**	0.1	1.7	58.0	1.06
Marathi	0.1	**3.7**	0.1	1.1	37.0	-0.44
Tamil†	0.0167	**3.4**	0.0167	1.0	204.0	-14.70
Urdu	1.3	**3.4**	1.2	1.0	2.7	1.44
Amharic†	**1.3**	0.01	0.001	1.0	236.4	1.32
Hausa†	**1.3**	0.01	0.001	1.0	236.4	-1.19
Swahili†	**1.3**	0.01	0.001	1.0	236.4	-0.15

Table 3: Post-training share of the target language (%) per decoder variant, from the Tiny Aya report. The region-matched variant is in bold. Ratio is the matched share over the mean of the other two regional variants, and Δ CER the Δ of Table 2. † marks a language with a share below the report's 0.05% reporting threshold, recovered from its region subtotal.

Tier	Languages	Loss rise	Fraction to best
Very Low	2	0.211	0.282
Low	1	0.188	0.292
Mid	2	0.041	0.627
High	7	0.003	0.816

Table 4: Overfitting by resource tier. Loss rise: how much evaluation loss increases after its own minimum (larger = more overfitting). Fraction to best: how far into training the minimum occurs. Each language enters as the median over its runs; Loss rise is the median and Fraction to best the mean over the tier's languages.

are small (median +3.7 CER). Above it the reductions are large: Amharic improves by 102.6 CER in-domain and 78.7 on FLEURS, Seselwa by 68.9, Hausa by 45.4, Tanzanian Swahili by 32.0.

Seselwa is the clearest single case, since neither frozen component supports it. Of the three off-the-shelf models, two fail to produce output on it at all, and the third reaches 85.18 CER; a trained projector reaches 16.28. On the held-out test the four decoder variants are within 0.50 CER of one another (16.28–16.78), so the choice of regional variant makes as little difference for Seselwa as for the other languages. The projector learns a usable alignment for a language absent from both pretraining corpora.

5.6 The trained models do not outperform the baselines on held-out dialects

WorldSpeech provides several dialects per language, and we trained on only a subset of them, allowing each trained checkpoint to be evaluated on dialects absent from its training data (Table 6). On the eighteen held-out dialects, the trained model has a lower CER than the best off-the-shelf model on none, and its CER is higher by a median of 4.1. On the dialects it trained on, it has the lower CER on 9 of 12. Of the eighteen held-out dialects, seventeen belong to English, Spanish and French, high-resource languages where even the trained dialect gained at most 2.7 CER over the baseline, so there was little advantage to transfer. Whether projector gains transfer across dialects of a low-resource language, where the gains are tens of CER, is an open question this data cannot answer.

6 Discussion

Implications for practitioners. When giving an LLM a speech interface for a low-resource language, the choice of decoder variant makes little difference. Choosing the regionally matched Tiny Aya decoder over a mismatched one changes CER by less than the difference between two runs of the same configuration. The same absence of an advantage holds on held-out data, where the minimum detectable effect is eleven times smaller than on the development set (0.44 against 4.90 CER). What does change CER is whether a projector is trained at all, and how much in-language audio is available to train it.

What the negative result does and does not show. Our results are limited to differences within this

Language	Test set	Domain	Base.	Trained (LisTAya)	Δ
English	FLEURS	OOD	2.1	5.5	3.4
	WS	ID	10.0	9.4	-0.6
Spanish	FLEURS	OOD	1.3	3.1	1.8
	WS	ID	8.5	8.9	0.4
French	FLEURS	OOD	2.1	7.5	5.4
	WS	ID	10.9	8.2	-2.7
Indonesian	FLEURS	OOD	3.3	5.9	2.6
	WS	ID	13.9	10.9	-3.0
Hindi	FLEURS	OOD	5.7	13.7	8.0
	WS	ID	24.5	25.8	1.3
Marathi	FLEURS	OOD	19.5	14.4	-5.1
	WS	ID	18.6	10.3	-8.3
Tamil	FLEURS	OOD	16.2	48.3	32.1
	WS	ID	18.4	15.4	-3.0
Urdu	FLEURS	OOD	11.7	20.0	8.3
	WS	ID	46.0	74.1	28.1
Amharic	FLEURS	OOD	111.6	32.9	-78.7
	WS	ID	133.1	30.5	-102.6
Hausa	FLEURS	OOD	55.9	39.1	-16.8
	WS	ID	82.0	36.6	-45.4
Swahili	FLEURS	OOD	26.3	15.1	-11.2
	WS	ID	18.6	14.5	-4.1
Seselwa	WS	ID	85.2	16.3	-68.9

Table 5: Per test set, the lowest CER among the three off-the-shelf models (Base.) and among the language's four trained LisTAya models, one per decoder variant. WS is the WorldSpeech test split, in-domain (ID), and FLEURS the FLEURS test split, out-of-domain (OOD). Δ is Trained minus Base.; negative (teal) favors training.

Language	Dialect	Base.	Trained (LisTAya)	Δ
English	en_au	9.5	14.9	5.4
	en_jm	7.6	9.5	1.9
	en_ke	19.5	23.8	4.3
	en_nz	8.5	17.3	8.8
	en_pk	78.5	422.9	344.4
	en_sl	14.1	20.3	6.2
	en_zm	15.8	21.4	5.6
Spanish	es_ar	20.0	22.0	2.0
	es_cl	19.6	23.1	3.5
	es_co	11.8	15.2	3.4
	es_es	7.4	13.0	5.6
	es_pe	6.6	10.4	3.8
	es_pr	9.5	12.2	2.7
	es_py	8.7	12.4	3.7
	es_uy	20.7	23.5	2.8
French	fr_cd	7.2	10.9	3.7
	fr_ci	7.8	13.2	5.4
Tamil	ta_lk	64.3	69.1	4.8
Urdu	ur_in†	20.2	17.4	-2.8
Swahili	sw_tz†	45.5	13.5	-32.0

Table 6: Dialects held out from training. Per dialect, the lowest CER among the three off-the-shelf models (Base.) and among the parent language's four trained LisTAya models, one per decoder variant. Δ is Trained minus Base.; negative (teal) favors training. † marks the two dialects included in the training data and not held out.

model family. These differences are small in absolute terms, with a median of 1.3 percentage points of post-training data, but large in relative terms, reaching up to 236 times more target-language data. The matched decoder's advantage is not significantly correlated with either the absolute or relative difference. This indicates that, when choosing among the released decoder variants of one family, performance does not depend on the regional label. The main exception in our experiments is English, where the matched decoder saw substantially less target-language data and the mismatched decoders achieve lower CER. This one case suggests that specialization may matter when the data difference is tens of percentage points rather than only one. More generally, decoder choice may still matter across model families. Fong et al. (2025), for example, report large gaps between EuroLLM and Salamandra.

The decoder's contribution is not primarily language-specific. Seselwa is supported by neither frozen component, and the projector still reaches error rates comparable to the low-resource end of the supported languages, with the four variants within 0.50 CER of one another. The decoder's contribution is therefore not primarily language-specific knowledge of the target language, which is consistent with the small differences between regional variants in the other languages.

7 Conclusion

We asked whether pairing a language with its regionally specialized decoder improves speech recognition in the SLAM-ASR framework. Across twelve languages and four Tiny Aya decoders, it does not. The matched decoder is 0.55 CER better on average than the mismatched decoders ($p = 0.97$) and shows no advantage on held-out test sets either. Giving a decoder up to 236 times more target-language data does not improve its recognition of that language. What does matter is training the projector. On low-resource languages, the trained projector improves over off-the-shelf models by up to 102.6 CER, including on Seselwa, a language neither frozen component supports. For practitioners, a general-purpose multilingual decoder is enough, and the effort is better spent on

collecting in-language audio. All LisTAya checkpoints and evaluation transcripts are released, so the comparison can be extended to other decoder families, where the differences between models are larger. A SLAM-ASR model, unlike a traditional ASR system, can be further instruction-tuned for tasks such as summarization, question answering or sentiment analysis. The checkpoints can therefore also be used as initial models for such tasks in these low-resource languages.

Limitations

Our grid covers 48 language–decoder configurations. Forty are trained once with a single learning rate and seed and have no direct uncertainty estimate. The eight configurations for Amharic and Seselwa are trained twice with different seeds, giving eight replicate pairs from which we estimate the between-run standard deviation. Of the twelve languages, ten are trained on WorldSpeech and use FLEURS development sets for checkpoint selection. This introduces a domain shift into model selection for these languages. We also report in-domain WorldSpeech test results, but the effect of this shift cannot be fully isolated. We train on WorldSpeech's human transcripts, which may contain errors. Due to rounding in the Tiny Aya report, which lists some exposure shares as 0.0, we estimate the shares behind four exposure ratios from the reported region subtotals. These estimates are constrained by the printed subtotals, and our conclusions hold across the full range they allow. Finally, our work only studies whether a specialized multilingual decoder benefits ASR within the SLAM-ASR architecture. It does not address whether this specialization benefits downstream spoken-language understanding tasks, where the decoder's multilingual knowledge would be more directly exercised.

Acknowledgments

This work was supported by Tranxform AI and NSTC under Grant 114-2221-E-A49-186.

References

Mahounan Pericles Adjovi, Victor Olufemi, Roald Eiselen, and Prasenjit Mitra. 2026. From Speech to Text Corpora: Evaluating ASR-Based Data Acquisition for Low-Resource Fongbe and Hausa. ArXiv:2606.22274 [cs.CL].

Antonis Asonitis, Luca A. Lanzendörfer, Frédéric Berdoz, and Roger Wattenhofer. 2026. WorldSpeech: A Multilingual Speech Corpus from Around the World. ArXiv:2605.09167 [cs.CL].

Huzain Azis, Muh Fatwah Fajriansyah M, Herdianti Darwis, Purnawansyah, Tasrif Hasanuddin, and Sugiarti. 2026. Development of a Low-Resource Automatic Speech Recognition System for the Makassar Dialect. In *2026 20th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–8.

Bharathi B, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Saranya S, and Suhasini S. 2026. Findings in Tamil Dialect Speech Recognition and Classification. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–79, Underline (Virtual). Association for Computational Linguistics.

Sergio Burdisso, Esaú Villatoro-Tello, Shashi Kumar, Srikanth Madikeri, Andrés Carofilis, Pradeep Rangappa, Manjunath K E, Kadri Hacioglu, Petr Motlicek, and Andreas Stolcke. 2026. Reducing Prompt Sensitivity in LLM-Based Speech Recognition Through Learnable Projection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16372–16376.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio Technical Report. ArXiv:2407.10759 [eess].

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.

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. In *Interspeech 2025*, pages 2003–2007.

Sukairaj Hafiz Imam, Tadesse Destaw Belay, Kedir Yassin Husse, Ibrahim Said Ahmad, Idris Abdulmumin, Hadiza Ali Umar, Muhammad Yahuza Bello, Joyce Nakatumba-Nabende, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. 2025a. Automatic Speech Recognition (ASR) for African Low-Resource Languages: A Systematic Literature Review. ArXiv:2510.01145 [cs.CL].

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025b. Automatic Speech Recognition

for African Low-Resource Languages: Challenges and Future Directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, Vienna, Austria. Association for Computational Linguistics.

Ondřej Klejch, William Lamb, and Peter Bell. 2025. A Practitioner's Guide to Building ASR Models for Low-Resource Languages: A Case Study on Scottish Gaelic. In *Proc. Interspeech 2025*, pages 728–732.

Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis, and Ravi Shekhar. 2026. On the Role of Encoder Depth: Pruning Whisper and LoRA Fine-Tuning in SLAM-ASR. In *Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) @ LREC 2026*, pages 183–193.

Chowdam Venkata Kumar, Kumud Tripathi, and Pankaj Wasnik. 2026. Overcoming Decoder Inconsistencies in Whisper for Dravidian and Low-Resource Languages. ArXiv:2606.09535 [cs.CL].

Shashi Kumar, Iuliia Thorbecke, Sergio Burdisso, Esaú Villatoro-Tello, Manjunath K E, 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.

Roman Kyslyi, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko, and Zakhar Guzii. 2026. Building ASR Resources for the Hutsul Dialect of Ukrainian. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 186–195, Rabat, Morocco. Association for Computational Linguistics.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devendra Singh Chaplot, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gabrielle Berrada, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jason Rute, Jean-Hadrien Chabran, Jessica Chudnovsky, Joachim Studnia, Joep Barmentlo, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Karmesh Yadav, Kartik Khandelwal, Kush Jain, Lélio Renard Lavaud, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Matthieu Dinot, Maxime Darrin, Maximilian Augustin, Mickaël Seznec, Neha Gupta, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Rémi Delacourt, Romain Sauvestre, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Shashwat Dalal, Siddharth Gandhi, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Tom Bewley, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, Yihan Wan, and Yunhao Tang. 2025. Voxtral. ArXiv:2507.13264 [cs].

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. ArXiv:2402.08846 [cs].

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 28492–28518, Honolulu, Hawaii, USA. JMLR.org.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, Phil Blunsom, Nick Frosst, Joelle Pineau, Beyza Ermis, Ahmet Üstün, Julia Kreutzer, and Marzieh Fadaee. 2026. Tiny Aya: Bridging Scale and Multilingual Depth. ArXiv:2603.11510 [cs.CL].

Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages. ArXiv:2511.09690 [cs].

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Yuchen Zhang, Ravi Shekhar, and Haralambos Mouratidis. 2026. Language Family Matters: Evaluating SpeechLLMs Across Linguistic Boundaries. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 487–499,

Rabat, Morocco. Association for Computational Linguistics.