Reassessing Speech Translation for Low-Resource Languages: Do LLMs Redefine the State-of-the-Art Against Cascaded Models?

Jonah Dauvet[1] **Min Ma**[2] **Jessica Ojo**[1] **David Ifeoluwa Adelani**[1,3]

[1]Mila - Quebec AI Institute, McGill University, [2]Google DeepMind, [3]Canada CIFAR AI Chair
jonah.dauvet@mail.mcgill.ca david.adelani@mila.quebec

Abstract

Automatic speech translation (AST) promotes seamless communication among speakers of different languages. While current state-of-the-art models excel with high-resource languages, their performance on low-resource languages (LRLs) is not well-established. We investigate this by evaluating state-of-the-art models on 10 LRLs with varying data amounts (10-30+ hours). Through six finetuning strategies and experimenting with three main AST paradigms, we observe that: (1) The latest Large Language Models (LLMs) may struggle with LRLs. (2) Comprehensive experiments suggest that for LRLs, more AST finetuning data is not always beneficial. (3) Our 2-Stage with ASR corrector finetuning recipe can substantially improve AST performance on LRLs, achieving up to a 5.8x BLEU score boost on translating related languages to English, while on par with the best monolingual finetuning in BLEU score when translating the target language to English. (4) We share our effective engineering practices, including how to effectively adapt AST models to unseen languages.

1 Introduction

Automatic speech translation directly converts speech from a source language into text or speech in a target language. The field has recently advanced at a rapid pace, driven by new paradigms like large-scale pre-training (Babu et al., 2021; Baevski et al., 2020; Conneau et al., 2020), large speech models, *e.g.* SeamlessM4T (Communication et al., 2023); Large Language Models (LLMs), *e.g.* ChatGPT (OpenAI, 2023); and speech-native audio LLMs, *e.g.* GPT-4o AUDIO (OpenAI, 2024), Gemini 2.0 Flash (Google, 2025), etc. Despite these progresses, many AST research centered on high-resource languages like English, French, German (Di Gangi et al., 2019; Bahar et al., 2019). Therefore, a most recent investigation of the novel modeling paradigms for the low-resource languages (LRLs) for AST is needed. AST for LRLs is constrained by scarce training data. Recent multilingual speech corpora like MuST-C (Di Gangi et al., 2019), CoVoST 2 (Wang et al., 2021), and FLEURS (Conneau et al., 2023) enable novel AST paradigms for these languages.

AST modeling paradigms fall into three categories: (1) **cascaded approaches** that apply automatic speech recognition (ASR) followed by machine translation (MT), (2) **multimodal MT approaches** like SeamlessM4T (Communication et al., 2023) that directly translate speech to text, and (3) multimodal **large language models** such as Gemini 2.0 Flash, which natively process text, speech and images, can perform direct speech-to-text translation (S2TT). Even for other LLMs which do not natively support audio input, mapping audio tokens to the token vocabulary (Wang et al., 2023; Ambilduke et al., 2025) can leverage MT capabilities, such as models like SALMONN (Tang et al., 2024), Qwen 2 Audio (Chu et al., 2024) and SPIRE (Ambilduke et al., 2025).

We investigate which approach works best for LRLs with small amounts of finetuning data. Specifically, *is the cascaded architecture superior with small data when compared to multimodal MT approaches like SeamlessM4T and Audio LLMs*? We experiment with ten LRLs from FLEURS (five Indic, five African) translating to English, choosing a translation direction $X \rightarrow$ *English* so that the multilingual capabilities of each method can be better assessed. We then compare performances across different AST paradigms against a high-resource language pair, *i.e.* *French* $\rightarrow$ *English*.

For cascaded approach, we proposed various finetuning strategies for all the three main modeling paradigms of AST. Through comprehensive

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 149–160

experiments across 11 languages, we show that the best AST approach depends on the resource-level of the languages: For languages with slightly better data availability, such as all five Indic languages and Swahili, prompting Gemini-2.0 Flash LLM works best. While for extremely low-resource languages, such as Hausa and Yorùbá, finetuning from large MT models or sequentially finetuning ASR and MT models can be more effective. To summarize, our contributions include:

- A comprehensive evaluation of AST for low-resource languages, establishing a **generalizable and highly effective blueprint**, comparing three modeling paradigms across 11 languages with various finetuning recipes.
- A simple yet effective "2-stage with ASR correction" strategy, that reduces WER by **54.2% relative on average** for African languages and yields a **5.8 times** increase in BLEU and a **2.6 times** increase in BLEU for African and Indic language groups, respectively, without additional data or model architectural changes.
- Our best recipe performs well on the target language while preserving **balanced AST performance** across languages, avoiding over-optimization for a single language. This offers practical guidelines for adapting multilingual AST models.
- Through comprehensive experiments, we share the finding that for low-resource languages, more AST finetuning is **not always beneficial**, providing a nuanced perspective on common practices.

We ensure **full reproducibility** by using only publicly available data and APIs, and open-sourcing our code and recipes.[1]

2 Related Work

The central challenge in AST is data scarcity (Xu et al., 2023) of high-quality paired (source speech, target text) data. Conventional AST thus uses cascaded approaches (Matusov, 2005) that first transcribe speech via ASR, then translate using MT. When it comes to the LRLs, the challenge of data scarcity is more severe. Multiple efforts address this challenge. Corpora like FLEURS and Common Voice (Ardila et al., 2019) enable AST for LRLs, while NaijaVoices (Emezue et al., 2025) and BhasaAnuvaad (Jain et al., 2024) contribute data for African and Indian languages, respectively, although wide gaps persist compared to high-resource languages.

Recent speech foundation models like Wav2Vec2 (Baevski et al., 2020) and multimodal LLMs (Google, 2025) have transformed AST: Bansal et al. (2018) and Stoian et al. (2020) demonstrated the benefit of pre-training AST models on high-resource ASR data to improve performance for low-resource language pairs.

Popular parameter-efficient finetuning methods such as LoRA (Hu et al., 2021; Liang et al., 2025), lightweight adapter (Le et al., 2021), always require changing the model architecture. In contrast to these studies, our research concentrates on the curriculum design of finetuning, to uncover hidden factors within simple full finetuning methods.

Kocmi et al. (2024) concluded that despite the rise of LLMs, AST still requires significant improvement, particularly in low-resource scenarios.

Multimodal benchmarks like SUPERB (Yang et al., 2021) cover many speech tasks but exclude AST, while mSTEB (Beyene et al., 2025) analyzes AST only at the language-family level. OWLS (Chen et al., 2025) demonstrates scaling benefits for low-resource performance, which our Whisper findings echo. We focus on broadly effective finetuning recipes and provide detailed analysis for low-resource African and Indic languages, underexplored in prior surveys.

Multilingual finetuning on models like Whisper (ASR) and SeamlessM4T (Communication et al., 2023) (AST) often degrades non-target languages, especially with monolingual finetuning. We propose an effective 2-stage finetuning curriculum that reduces this shift without architectural changes, much simpler than multi-stage methods proposed in Thillainathan et al. (2025). We also apply LLM correction to ASR components, previously used mainly in ASR systems (Ruder et al., 2023; Ma et al., 2025).

Trade-offs between cascaded and end-to-end systems remain debated, with methods lacking systematic evaluation across diverse LRLs. Our work aims to complement these efforts by providing a unified, cross-paradigm evaluation across LRLs, comparing data efficiency and generalization across cascaded, multimodal MT, and audio-LLM systems.

[1] https://github.com/McGill-NLP/ast-lrl-speech

3 Experimental Setup

3.1 Model Selection and Baselines

3.1.1 Cascaded approach

We employ OpenAI's WHISPER LARGE V3 1.5B given its robust zero-shot performance across 98 non-English languages from 680 K hours of weakly supervised ASR data and 125 K hours of speech-to-English translation pairs (Radford et al., 2022). For MT, we integrate Meta's NLLB-200 1.3B, trained on hundreds of billions of tokens spanning 200 languages (NLLB-Team, 2022). This setup strikes a balance between translation quality and computational efficiency. We evaluate WHISPER LARGE V3 on FLEURS test of 11 target languages to serve as a cascaded-approach baseline.

3.1.2 Multimodal machine translation

We evaluate Meta's SEAMLESSM4T LARGE 1.6B, pretrained on 4.1 M hours of speech and text data over 100 languages. It enables direct speech-to-text and speech-to-speech translation without separate ASR/MT modules (Communication et al., 2023), serving as our end-to-end baseline.

3.1.3 Audio LLMs

We benchmark two SOTA audio LLMs: OpenAI's GPT-4O AUDIO (GPT-4o backbone with audio pretraining), and Google's GEMINI 2.0 FLASH, a multimodal model that supports text and audio. Both reflect SOTA AST via their incorporation of leading-edge modeling and web-scale training data.

3.2 Data

Training and Evaluation data: We used the FLEURS dataset for the initial training data. FLEURS contains n-way parallel speech and text in 102 typologically and geographically diverse languages drawn from the FLoRes-101 benchmark (Goyal et al., 2021), with approximately 12 hours of high-quality, human-read speech per language. Since 80% of these are low-resource languages, FLEURS is well-suited for evaluating AST paradigms in such settings.

Data for ablation: For our ablation studies on African languages, we added 20 hours of validated speech from Mozilla Common Voice[2] (Swahili and Luganda) and the Naija Voice corpus (Lee et al., 2022) (Igbo, Hausa, and Yorùbá). Common Voice lacked sufficient validated data[3] for the Nigerian languages, whereas Naija Voice offers over 600 hours per language.

Model	Parameters	Used Capabilities	Unsupported Lang.
Whisper Large v3	1.5 B	ASR	Igbo, Luganda
NLLB-200 Large	1.3 B	MT	None
SeamlessM4T Large	1.6 B	Multimodal MT AST	Hausa
mT5-Base	580M	ASR correction (T2T)	Luganda
GPT-4o Audio	Unknown	End-to-End AST	Unknown
Gemini 2.0 Flash	Unknown	End-to-End AST	Unknown

Table 1: Model Information. Please refer to **Section** 3.1 for details.

3.3 Model finetuning Strategies

We detail our finetuning recipes for adapting the ASR model of the *cascaded approach* and for the general finetuning of multimodal MT.

3.3.1 ASR model finetuning

We *finetuned* on the FLEURS training data of each of 11 spoken languages for 10 epochs by updating all the parameters. To ensure consistent evaluation across all methods, the best model was selected after 10 epochs without using a validation set. We also note that as Igbo and Luganda are not included in Whisper's original language inventory, Whisper will reject any training examples tagged with an out-of-vocabulary language code. Therefore, we override the language identifier during finetuning by mapping languages to their closest relatives in the supported set based on phonology and lexical similarity. For instance, we map Igbo to Lingala and Luganda to Shona. Similar approach to finetune machine translation models for unseen languages has been mentioned in (Yang et al., 2021). We describe the different finetuning recipes below (all parameters were updated if not specified).

- Monolingual finetuning ("**Monolingual**" or "**S2**"): we independently finetuned ten separate WHISPER LARGE V3, where finetuning uses the entire FLEURS training data of the target language. The preprocessing pipeline and training hyperparameters are the same as the multilingual experiments.

- Multilingual finetuning (**S3**): we group our ten target languages into two regionally and typologically coherent subsets: "Indic" (Hindi, Punjabi, Tamil, Telugu, Malayalam), and

[2]https://huggingface.co/datasets/mozilla-foundation/common_voice_17_0

[3]Common Voice is volunteer-based, with recordings requiring validation for quality.

"African" (Swahili, Hausa, Yorùbá, Igbo, Luganda), We then finetuned two WHISPER LARGE V3 models on the combined data of all languages from each group, motivated by potential cross-language transfer (Conneau et al., 2020): *e.g.* African languages using a shared Latin script, while Indic languages use distinct writing systems but are similar in phonology.

- 2-stage FT (Multilingual + Monolingual, **S4**): to capture both cross-lingual transfer and language-specific specialization, we first conduct a multilingual finetuning with group data for 10 epochs, then continue finetuning on the target language only for 10 more epochs.

- ASR corrector (**S5** and **S6**): to explore how much text-only correction can reduce recognition errors beyond speech finetuning, we adopt the ASR correction strategy from XTREME-UP (Ruder et al., 2023), applying it to the optimal models finetuned by above recipes. We finetuned mT5-base (Xue et al., 2021) a Text-to-Text model for 20 epochs with earlystopping on ASR (finetuned WHISPER LARGE V3 **S3**) prediction-reference pairs from the FLEURS training set. This approach ensures no data leakage, as we leverage the same training data used in speech finetuning. Full training details are in Appendix A.

Once ASR transcribed input speech into text of source language, we used NLLB (NLLB-Team, 2022), an open-sourced large-scale machine translation model to translate text to the target language.

3.3.2 General MT finetuning

We finetuned SEAMLESSM4T LARGE model, which supports speech inputs, on Indic and African language groups separately, by updating all the parameters over 10 epochs. This method is a fully end-to-end approach of AST.

3.4 Evaluation metrics

We use BLEU to evaluate final performances of all machine translation systems. For cascaded systems, we also report ASR Word Error Rate (WER)[4].

4 Results & Analysis

4.1 ASR Performance

Table 2 presents an overview of ASR baseline created by WHISPER LARGE V3, with the finetuning recipes described in **Section 3.3.1**. We observed:

Monolingual finetuning is most efficient while 2-Stage better maintains generalization. Given the same finetuning amounts of speech data, solely finetuning on target languages significantly reduced average baseline WER from 88.39% to 45.90%. Multilingual finetuning (S3) also significantly reduced WER for the single target languages, though slightly worse than the monolingual ones. Interestingly, continuing finetuning from the multilingual model (S3) on individual target languages, without using any additional data, S4 not only recovered the performance on each language but also resulted in slightly better performance than monolingual finetuning (S2). This might be because the design of the 2-stage FT (S4) recipe allows the model to better learn from the common acoustic–phonetic and lexical properties shared by related languages.

Multilingual + Monolingual + Corrector is most effective. The system consistently performed best in 9 of the 10 low-resource languages. The strategy did not introduce any additional speech data, but leverage reference transcripts in a more effective way. Specifically, the corrector models learned from paired (ASR transcript, reference transcript) training data, leading to an average 15.2% relative reduction in WER compared to the Multilingual (S3) baseline and a significant 54.2% reduction relative to the initial Baseline (S1) models, all without increasing the footprint of the multilingual finetuned ASR models.

Zero-shot evaluation might be enough for ASR of high-resource languages. We selected French, a high-resource language, to evaluate the off-the-shelf models' performance and to understand the performance gap when compared against their performance on the low-resource languages we focus on in this paper. As shown in Table 2, by directly evaluating WHISPER LARGE V3 on French test data, the WER already achieved 12.73%; however, the simplest monolingual finetuning nearly doubled French WER to 24.72%. We hypothesize that, for languages with abundant data and well-optimized pre-training representations, aggressive monolingual adaptation can induce overfitting

[4]Adopted the implementation of https://huggingface.co/spaces/evaluate-metric/wer.

Language	Whisper ASR Baseline and Models finetuned from IT						$\Delta_{(S1,S6)}$	$\Delta_{(S3,S6)}$	FLEURS Training Hours
	Baseline (S1)	Mono. (S2)	Multi. (S3)	Multi. + Mono. (S4)	Multi. + ASR Corrector (S5)	S4 + ASR Corrector (S6)			
French	**12.73%**	24.72%	x	x	x	x	x	x	10.3
Hindi	46.67%	24.06%	25.00%	23.85%	23.12%	**21.63%**	-53.6%	-13.4%	6.6
Punjabi	84.46%	33.66%	33.91%	**32.68%**	40.70%	43.08%	-48.9%	+27.0%	6.3
Tamil	59.96%	45.33%	46.25%	44.40%	40.54%	**38.58%**	-35.6%	-16.5%	8.6
Telugu	78.12%	45.75%	46.03%	44.38%	39.46%	**37.63%**	-51.8%	-18.2%	7.9
Malayalam	138.91%	44.87%	46.20%	44.02%	43.45%	**39.81%**	-71.3%	-13.8%	10.0
Swahili	42.88%	33.11%	35.04%	33.26%	24.86%	**22.85%**	-46.7%	-34.8%	13.4
Hausa	112.78%	42.58%	49.27%	43.78%	40.07%	**34.47%**	-69.4%	-30.0%	13.6
Yorùbá	105.70%	68.67%	68.69%	66.36%	64.93%	**61.92%**	-41.4%	-9.8%	10.0
Igbo	106.56%*	59.26%	61.84%	56.98%	54.66%	**50.93%**	-52.2%	-17.6%	13.8
Luganda	107.90%*	61.68%	47.72%	60.46%	54.16%	**53.26%**	-50.6%	+11.6%	12.6
Average	88.39%	45.90%	47.72%	45.02%	42.59%	**40.45%**	-54.2%	-15.2%	10.3

[a] Starred (*) WERs indicate that the target languages were unseen by the model. **Bolded WERs** indicate the best score across different finetuning strategies and baseline.

Table 2: Overview of **WER**(↓) for Whisper Large ASR models using different finetuning strategies (denoted as S2 – S6). We show $\Delta_{(S1,S6)}$, the relative changes obtained by S6 using S1 as baseline. Similar notation for $\Delta_{(S3,S6)}$. please refer to **Section** 3.3.1 for definitions of finetuning strategies, and **Section** 4.1 for detailed analysis.

or catastrophic forgetting of general acoustic patterns. When comparing the output transcripts from both models, we observed peculiar word hallucinations in the monolingual model ($e.g.$ "Dans le climat chaud" was transcribed as "Dans le chumacho"). These phonetic hallucinations were similar to those seen in other languages, but unlike those instances, they were exacerbated rather than mitigated by monolingual finetuning. Such regression suggests more thoughts in the finetuning design to preserve the learned syntax while adapting large speech model to the target data domain.

Similar language serves as a good proxy when adapting to an unseen language. A key challenge in finetuning the Whisper ASR model for Igbo and Luganda was that they are not among the 98 languages Whisper supports. We notice that both the two *unseen* languages use Latin writing system, so we hypothesized that a similar language label could serve as a proxy. Specifically, we selected Lingala and Shona as the proxy language label for Igbo and Luganda respectively, considering their phonetic and regional similarities. Experimental results prove the method's effectiveness, with relative improvements of up to 52.2% for Igbo and 50.6% for Luganda achieved by the best finetuning recipe. This success suggests a strong potential to expand Whisper's coverage to 20+ additional low-resource languages beyond its current 98 non-English ones, with careful selection of proxy language: To verify the effect of proxy choices, we also conducted a comparative experiment by labeling Igbo as French: while both use Latin alphabet, they differ phonetically. The dramatic increase in WER indicates the importance of a proper proxy language.

4.2 Translation Quality

All three AST modeling paradigms, cascaded ASR+MT (with various finetuned ASR models), multimodal SeamlessM4T, and audio-centric LLMs (GPT-4o Audio and Gemini 2.0 Flash), have been evaluated in terms of BLEU in Table 4.

Gemini works best for Indic speech translation. For the five Indic languages (Hindi, Punjabi, Tamil, Telugu, Malayalam) and Swahili, Gemini 2.0 Flash achieves the highest BLEU in every case ($e.g.$ 35.38 on Hindi, 30.78 on Telugu, and 31.91 on Swahili), outperforming both GPT-4o Audio and all cascaded or multimodal MT baselines.

Cascaded ASR+MT models and expert MT models seem more effective to finetune for under-represented languages. For lower-resource African languages (Hausa, Yorùbá, Igbo, Luganda), the best results are obtained by finetuned Whisper variants + NLLB and SeamlessM4T, rather than audio LLMs: Whisper Multi. + Mono. + ASR Corrector reaches 13.93 on Igbo and 20.05 on Hausa, and SeamlessM4T Multilingual peaks at 18.92 on Luganda – each exceeding Gemini 2.0 Flash's corresponding 2.19, 16.29, and 11.93. When averaging across all languages except French, the cascaded Whisper Monolingual (21.26), Whisper Multilingual + ASR Corrector (21.82), and SeamlessM4T Multilingual (21.28) nearly match Gemini 2.0 Flash's 22.09, while Whisper Multilingual + Monolingual + ASR Corrector ($i.e.$ T6), actually outperforms Gemini with 22.24 BLEU, indicating targeted finetuning on low-resource corpora can rival SOTA audio LLMs in AST performance.

Zero-shot evaluation might be enough for the translation of high-resource languages. As stated

Source Language X	Monolingual		Multilingual + Monolingual		Monolingual		Multilingual + Monolingual	
	WER(X)	Average WER(Others)	WER(X)	Average WER(Others)	BLEU(X)	Average BLEU(Others)	BLEU(X)	Average BLEU(Others)
Hindi	24.06%	56.07%	23.85%	22.76%	31.18	14.81	30.90	23.86
Punjabi	33.66%	80.50%	32.68%	34.22%	26.59	3.42	26.68	19.50
Tamil	45.33%	74.32%	44.40%	40.25%	22.65	4.69	22.78	16.85
Telugu	45.75%	87.83%	44.38%	43.70%	25.12	2.96	25.15	18.32
Malayalam	44.87%	98.13%	44.02%	41.39%	27.07	1.77	27.68	20.66
Indic Group	38.73%	79.37%	37.87% (-2%)	36.46% (-54%)	26.52	5.53	26.64 (+0.5%)	19.84 (+259%)
Swahili	33.11%	76.52%	33.26%	30.33%	27.55	4.58	27.70	20.80
Hausa	42.58%	87.31%	43.78%	44.74%	18.45	0.61	18.34	12.22
Yorùbá	68.67%	80.75%	66.36%	62.80%	11.14	0.88	11.04	6.53
Igbo	59.26%	83.92%	56.98%	56.30%	11.46	1.03	11.60	7.05
Luganda	61.68%	117.59%	60.46%	53.78%	11.36	1.05	11.56	8.69
African Group	53.06%	89.22%	52.17% (-2%)	49.59% (-44%)	15.99	1.63	16.05 (+0.4%)	11.06 (+579%)

Table 3: A comparison of Monolingual and Multilingual+Monolingual models. The table displays **WER** (↓) and **BLEU** scores (↑) for various Indic and African languages. Highlighted cells show the performance for the grouped languages. “Others" refers to the other languages in the same group except target language X.

Language	ASR (Whisper) + MT (NLLB)						Multimodal Speech Translation		Audio LLMs	
	Cascaded Baseline (T1)	Cascaded Mono. (T2)	Cascaded Multi. (T3)	Cascaded Multi. + Mono. (T4)	Cascaded Multi. + ASR Corrector (T5)	T4 + ASR Corrector (T6)	SeamlessM4T Baseline	SeamlessM4T Multi.	GPT-4o audio	Gemini 2.0 Flash
French	**38.30**	31.49	x	x	x	x	33.77	x	37.49	36.16
Hindi	27.79	31.18	30.85	30.90	31.08	31.48	24.62	28.71	29.28	**35.38**
Punjabi	13.87	26.59	26.65	26.68	25.38	24.62	28.71	28.10	19.15	**29.58**
Tamil	19.53	22.65	22.48	22.78	23.10	23.43	19.93	21.87	15.17	**25.14**
Telugu	17.51	25.12	25.26	25.15	27.37	27.53	23.27	24.93	19.83	**30.78**
Malayalam	1.32	27.07	27.05	27.68	27.79	28.45	21.20	25.95	23.55	**30.31**
Swahili	25.01	27.55	27.18	27.70	28.40	28.38	14.81	31.22	19.37	**31.91**
Hausa	3.36	18.45	15.90	18.34	18.04	**20.05**	1.01*	6.07	1.07	16.29
Yorùbá	2.62	11.14	10.66	11.04	10.62	11.18	12.64	**15.36**	2.31	7.35
Igbo	1.80*	11.46	9.86	11.60	13.23	**13.93**	0.19	11.65	1.63	2.19
Luganda	4.07*	11.36	11.34	11.56	13.14	13.35	5.95	**18.92**	4.95	11.93
Average	11.69	21.26	20.72	21.34	21.82	**22.24**	15.23	21.28	13.63	22.09

[a] Starred (*) BLEUs indicate that the target languages were unseen by the model. **Bolded BLEUs** indicate the best score across different finetuning strategies and baseline.

Table 4: Overview of **BLEU** scores (↑) achieved by SOTA models with different finetuning strategies. please refer to **Section** 3.3 for definitions of finetuning strategies, and **Section** 4.1 for detailed analysis.

before, French is an exception: the Whisper-Large-v3 baseline attains the highest BLEU of 38.30, surpassing GPT-4o Audio (37.49) and Gemini 2.0 Flash (36.16). This underscores the robustness of Whisper’s original capacity on high-resource languages – further finetuning may introduce degradation in such well-represented language settings.

4.3 Generalization vs. Specialization

A typical challenge for finetuned multilingual models is balancing **specialization** and **generalization**. While finetuning solely on a target language might yield the lowest ASR WER and the highest BLEU score for that language, severe performance degradation in other languages must be avoided. This consideration is also critical for practical applications. When serving a speech translation model for Hindi to English, users in the same region might not always speak Hindi but may use other local languages such as Punjabi. Even predominantly Hindi speakers might code-switch between Hindi and other local languages – this is a significant concern in the engineering and application of speech translation models. Therefore, we measured the ASR and MT performances not only on target languages but also on their average performance across other languages within the same geographical region (dubbed “Average Other"). As shown in Table 3, for monolingual finetuned ASR models, even if their WER for a single target language is slightly lower than that of multilingual + monolingual finetuned models (e.g. 23.85% WER $vs.$ 24.06% WER on Hindi, obtained by the two finetuned models, respectively), the monolingual model clearly shifts too heavily toward Hindi. This specialization causes finetuned model to fail to perform well on other Indic languages, as indicated by the 56.07% average WER on other languages. In contrast, the first stage of multilingual finetuning allows the final finetuned models to maintain their performance on the other Indic languages, with a 22.76% average WER, which is a **59%** relative improvement over their monolingual finetuned counterparts. We found similar patterns in terms

of BLEU scores among the MT models. The necessity of a two-stage finetuning approach is thus highlighted by two significant benefits: it maintains ASR and MT performance on related languages and offers potential gains from sharing common cross-lingual features.

4.4 Effect of finetuning Data Volume

Acquiring finetuning speech data for extremely LRLs is highly challenging. Therefore, we conducted an ablation study to investigate the minimum hours of speech required to develop a speech translation model with acceptable performance. We use all five African languages as examples, and present ablation studies in terms of both WER and BLEU, across different finetuning data amount: 0, 1, 2.5, 5, 10, and 20 hours per language, in Figure 1 and 2 for ASR and MT component of cascaded system respectively.

Zero-shot evaluation is a better choice when finetuning data is too limited. While the initial one hour of fine-tuning on Common Voice or Naija Voice indeed yields a marked degradation in ASR quality and downstream translation – evidenced by WER jumps (Hausa 42.5 % → 54.4%, Yorùbá 68.6% → 70.5%) and BLEU drops (Hausa 18.45 → 16.87, Yorùbá 11.14 → 10.26) – subsequent training yields recovery and improvement: at 2.5 h, WER for all five languages recedes toward or below baseline (Igbo 59.2% → 55.7%) and BLEU surpasses the baseline model (Igbo 11.46 → 12.38).

Gains are most pronounced between 2.5 – 5 h, as BLEU increases by up to +1.30 points (Yorùbá 11.14 → 12.44), while WER reduces by up to -8.4% (Hausa 54.4% → 46.0%). Between 5–10 h, improvements continue but at a reduced rate (e.g. Swahili BLEU plateaus at 28.20, Luganda WER only marginally improves from 59.5% to 59.0%), indicating that the model rapidly ingests new acoustic-textual patterns within the first 10 h. Beyond 10 h, additional data yields diminishing, or even slightly negative returns (Hausa BLEU 19.13 → 19.01; Luganda BLEU 12.42 → 12.13), suggesting an inflection point where the domain shift of the supplemental corpus begins to outweigh its benefit. Nonetheless, we observe that on average, the addition of new unseen data to the monolingual model matches best scores shown in Tables 2 and 4.

Especially for the ablation on MT, results showed a "U-shaped" curve, suggesting initial overfitting to new data followed by swift adaptation. We identified an optimal fine-tuning window of 2.5–10 h for maximizing ASR robustness and translation fidelity in African low-resource languages.

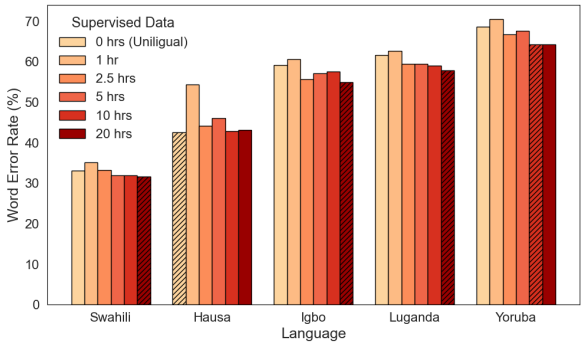

Figure 1: Sample efficiency measured by ASR **WER** (%) scores (↓) with varying amounts of finetuning hours; dashed bars indicate the best system for each language. Please refer to **Section 4.4** for details.

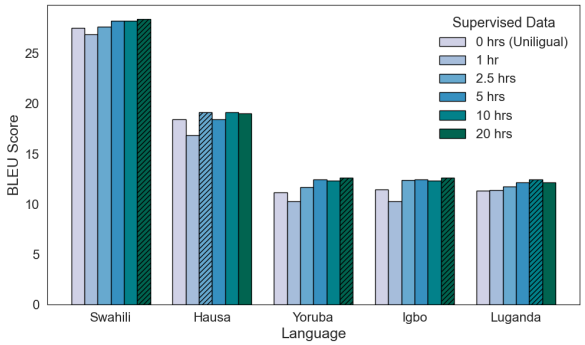

Figure 2: Sample efficiency measured by MT **BLEU** scores (↑) with varying amounts of finetuning hours; dashed bars indicate the best system for each language. Please refer to **Section 4.4** for details.

4.5 Beyond BLEU: Part-of-speech Tag Steering Analysis

To gain insights beyond a single BLEU score, we analyzed part-of-speech (POS)–specific translation errors for our baseline cascaded model (T1) and the cascaded architecture with ASR correction (T6), across five African languages. POS tagging was performed using `spaCy`[5]'s large English statistical model, which produced Universal Dependencies tags for each token.

Following the methodology of (Popović and Ney, 2007), we computed POS-specific WER, which reflects sequence-level accuracy and highlights error patterns across linguistic categories. Our analysis (Tables 5–6) shows that T1 exhibits high WER for NOUN, PUNCT, and DET categories, especially

[5]spaCy is a library for NLP in Python and Cython.

POS Tag	Swahili	Hausa	Igbo	Luganda	Yorùbá	Avg
ADJ	5.31%	10.29%	10.39%	7.79%	9.68%	8.69%
ADP	6.55%	**16.32%**	13.65%	11.56%	**16.41%**	12.89%
ADV	2.49%	8.21%	6.99%	5.55%	5.30%	5.71%
AUX	3.34%	9.22%	9.13%	6.30%	14.23%	8.44%
CCONJ	1.70%	5.92%	5.91%	3.15%	5.44%	4.42%
DET	5.91%	**17.57%**	**24.76%**	14.91%	**16.72%**	**15.97%**
NOUN	13.04%	**31.54%**	**34.15%**	**25.97%**	**31.96%**	**27.33%**
NUM	1.20%	1.86%	2.10%	1.56%	1.55%	1.65%
PART	1.50%	3.12%	3.01%	2.87%	7.97%	3.69%
PRON	2.68%	10.05%	10.71%	5.23%	**19.50%**	9.63%
PROPN	3.91%	14.01%	9.04%	8.57%	10.02%	9.11%
PUNCT	4.72%	**23.92%**	**28.41%**	**15.90%**	**21.68%**	**18.93%**
SCONJ	0.77%	1.88%	1.68%	1.68%	2.95%	1.79%
VERB	6.78%	13.36%	12.44%	10.27%	**20.84%**	12.74%
Macro Avg	5.04%	12.60%	12.80%	9.17%	13.47%	10.62%
Weighted Avg	5.99%	16.86%	17.27%	12.14%	18.44%	14.14%

Table 5: **WER** (↓) over English POS tags of translation by Whisper Baseline (T1) for all five African languages.

POS Tag	Swahili	Hausa	Igbo	Luganda	Yorùbá	Avg
ADJ	4.64%	5.84%	6.97%	6.39%	6.54%	6.08%
ADP	5.64%	7.48%	8.76%	7.45%	7.79%	7.42%
ADV	2.36%	2.77%	3.42%	3.26%	3.11%	2.98%
AUX	2.77%	3.93%	4.25%	3.64%	4.03%	3.72%
CCONJ	1.72%	2.65%	2.61%	2.23%	2.37%	2.32%
DET	4.97%	8.58%	8.48%	7.37%	7.61%	7.40%
NOUN	12.38%	**16.51%**	**18.53%**	**16.74%**	**16.91%**	**16.21%**
NUM	0.86%	1.35%	1.45%	1.21%	1.08%	1.19%
PART	1.40%	1.50%	1.97%	1.75%	1.66%	1.66%
PRON	2.47%	3.30%	3.57%	3.40%	3.39%	3.23%
PROPN	3.75%	5.41%	6.53%	5.54%	5.26%	5.30%
PUNCT	5.38%	7.54%	8.21%	8.69%	7.09%	7.38%
SCONJ	0.69%	0.81%	1.17%	1.02%	1.03%	0.94%
VERB	5.90%	7.86%	8.78%	8.06%	8.69%	7.86%
Macro Avg	4.18%	5.79%	6.47%	5.81%	6.05%	5.66%
Weighted Avg	5.49%	7.56%	8.48%	7.68%	7.66%	7.37%

Table 6: **WER** (↓) over English POS tags of translation by our best recipe (T6) for all five African languages.

for Yorùbá, the lowest-BLEU language. This indicates frequent issues with content words, determiners, and punctuation, limiting translation quality.

Setting a threshold of 15% for POS-wise WER, then as highlighted in Table 5, the most common errors were made over NOUN, PUNCT, and DET classes, indicating the deficiencies of Whisper model, on the African language group. For Yorùbá, the language with the lowest BLEU score, high WERs are observed across multiple POS classes. This unveils underlying error patterns and suggests that these specific word types require focused attention to improve translation performance.

Comparing Table 6 to Table 5, we observed a large reduction in errors for PUNCT and DET, along with a smaller, yet significant, reduction for NOUN. These substantial improvements across all five languages—particularly in Yorùbá, Hausa, Igbo, and Luganda—further demonstrate the effectiveness of the best T6 recipe. We also conducted more detailed analysis of position-independent error, inflectional error and missing words, details are in Appendix B.

4.6 Summary of Trends

Across our experiments, three consistent patterns were observed. First, in the cascaded method, finetuning from SOTA ASR model Whisper on even modest amounts of in-domain data produces substantial WER reductions for low-resource languages (Table 2). The Multi. + Mono. + ASR Corrected variant yielded the best WER for 9 of 10 lanaguages, as it leverages extended exposure and cross-lingual transfer. Only French deviates from this trend, underscoring the risk of overfitting when pretraining already provides ample coverage. Second, in multimodal machine translation quality (Table 4), a complementary pattern appears: audio-LLMs like Gemini 2.0 Flash can translate well in Indic languages → English and Swahili → English, achieving BLEU gains of 4–7 points over cascaded baselines, whereas finetuned translation expert models (either built multimodaly or ASR+MT cascadedly) excel on low-resource African languages, often exceeding Gemini's scores by 2–10 BLEU points. Third, our ablation on finetuning volume (Figs. 1–2) reveals a pronounced "U-shaped" curve: an initial performance dip at 1 h, rapid recovery and peak gains between 2.5–10 h, and plateau or slight regression beyond 10 h. This identifies an optimal finetuning window for balancing adaptation speed against domain shift.

Together, these trends suggest a best recipe for speech-translation in low-resource contexts: (1) apply multilingual finetuning followed by targeted monolingual finetuning, with Corrector to minimize WER and maximize the final translation performances on related languages; (2) reserve audio-LLMs for languages with ample training data, while relying on cascaded or multimodal MT systems for under-represented tongues; (3) allocate finetuning budgets within the identified "sweet spot" of 2.5–10 h to maximize returns without incurring diminishing gains.

5 Conclusions

Our systematic comparison of cascaded ASR+MT, multimodal speech translation, and audio-centric LLMs across 11 diverse languages yields several important insights: (1) Our 2-stage FT strategy can improve translation performances on target language, and offer the additional performance benefit on regional related languages for both ASR and

MT, with a up to 5.8x boost in BLEU on them than monolingual FT. This approach is particularly effective for meeting the demands of practical, real-world scenarios. (2) Our 2-stage FT + ASR Corrector recipe can further improve WER across 9 of 10 languages, and carry on the additional gains to ultimate MT task. (3) While SOTA audio-LLMs excel on higher-resource languages, our evaluations unveil that they may struggle on truly low-resource languages such as African ones. Finetuned Whisper variants and SeamlessM4T can match or exceed audio-LLM performance by up to 10 BLEU, suggesting the most reliable choices for AST of under-represented spoken languages. (4) Our ablation study reveals that not always "the more finetuning data, the better" in low-resource ASR. Future work should focus on expanding high-quality parallel speech–text resources and developing regularized, domain-aware adaptation techniques to ensure robust translation across the full spectrum of the world's languages.

6 Acknowledgment

This research was supported by the Google grant via Mila. David Adelani acknowledges the funding of the Natural Sciences and Engineering Research Council of Canada (NSERC)—Discovery Grants Program, IVADO and the Canada First Research Excellence Fund. We would also like to thank Google Cloud for the GCP credits Award through the Gemma 2 Academic Program for providing API credits.

7 Limitations

This study provides valuable insights into speech-to-text translation for low-resource languages, but its scope is bounded by several factors. There is bias introduced the selection of low-resource languages, $e.g.$ we experimented with clean speech rather than noisy speech to initialize the comparative studies. Future work with diverse, in-the-wild data is crucial for robust systems. Secondly, while we selected 10 typologically diverse African and Indic languages to evaluate low-resource performance, our findings may not extend to all such languages, especially those with different linguistic features or data availability. Thirdly, we focused on selected architectures (Whisper+NLLB, SeamlessM4T, GPT-4o Audio, Gemini 2.0 Flash). While proprietary APIs offered state-of-the-art insights, their closed nature and cost limited extensive testing. Open models were finetuned within practical compute budgets, constraining exploration of larger variants and complex adaptation. These choices, driven by resource constraints, introduce selection bias in model coverage and task prioritization.

References

Kshitij Ambilduke, Ben Peters, Sonal Sannigrahi, Anil Keshwani, Tsz Kin Lam, Bruno Martins, Marcely Zanon Boito, and André FT Martins. 2025. From tower to spire: Adding the speech modality to a text-only llm. *arXiv preprint arXiv:2503.10620*.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

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.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799. IEEE.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. msteb: Massively multilingual evaluation of llms on speech and text tasks. *arXiv preprint arXiv:2506.08400*.

William Chen, Jinchuan Tian, Yifan Peng, Brian Yan, Chao-Han Huck Yang, and Shinji Watanabe. 2025. Owls: Scaling laws for multilingual speech recognition and translation models. *arXiv preprint arXiv:2502.10373*.

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Seamless Communication et al. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

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.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics.

C. Emezue, T. N. Community, B. Awobade, A. Owodunni, H. Emezue, G. M. T. Emezue, others, and C. Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. *arXiv preprint arXiv:2505.20564*.

Google. 2025. Gemini 2.0 flash.

N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzman, and A. Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*, 10.

Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 13 indian languages. *arXiv preprint arXiv:2411.04699*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. *arXiv preprint arXiv:2106.01463*.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pretrained multilingual sequence-to-sequence models: A hope for low-resource language translation? *arXiv preprint arXiv:2203.08850*.

Xiao Liang, Yen-Min Jasmina Khaw, Soung-Yue Liew, Tien-Ping Tan, and Donghong Qin. 2025. Towards low-resource languages machine translation: A language-specific fine-tuning with lora for specialized large language models. *IEEE Access*.

Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill. 2025. Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*.

Evgeny Matusov. 2005. On the integration of speech recognition and statistical machine translation.

The NLLB-Team. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2023. Chatgpt (mar 14 version) [large language model].

OpenAI. 2024. Hello gpt-4o. OpenAI Blog.

Maja Popović and Hermann Ney. 2007. Word error rates: Decomposition over pos classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT '07)*, pages 48–55, Prague.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

S. Ruder, J. H. Clark, A. Gutkin, M. Kale, M. Ma, M. Nicosia, others, and P. Talukdar. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, et al. 2024. Salmonn: Towards generic hearing abilities for large language models. In *Proc. ICLR 2024*.

Sarubi Thillainathan, Songchen Yuan, En-Shiun Annie Lee, Sanath Jayasena, and Surangika Ranathunga. 2025. Beyond vanilla fine-tuning: Leveraging multistage, multilingual, and domain-specific methods for low-resource machine translation. *arXiv preprint arXiv:2503.22582*.

C. Wang, A. Wu, J. Gu, and J. Pino. 2021. Covost 2 and massively multilingual speech translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, et al. 2023. Slm: Bridge the thin gap between speech and text foundation models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. *arXiv preprint arXiv:2306.11646*.

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, others, and C. Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I.J. Lai, K. Lakhotia, Y.Y. Lin, A.T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee. 2021. Superb: Speech processing universal performance benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

A ASR Corrector Training Details

Goal. We train a text-to-text ASR corrector to reduce recognition errors made by the ASR model. The corrector is a language-specific mT5-Base model that maps noisy ASR hypotheses to corrected transcripts.

Data and pairing. For each language, we take predictions from the finetuned WHISPER LARGE V3 (**S3**, see §3.3.1) on the FLEURS *training* split and pair them with their gold references to form (hypothesis, reference) examples. The FLEURS *dev* split is used only for early stopping and hyperparameter selection. This ensures no data leakage: the corrector never sees dev/test references during training.

Model and objective. We finetune mT5-Base for up to **20 epochs** with early stopping on the dev set. The model is trained as a standard seq2seq text editor: input is the ASR hypothesis; target is the reference transcript.

Compute. All runs use **2× A100L GPUs**, **6 CPUs**, and **32 GB** RAM.

Setting	Value
Base model	mT5-base (Text-to-Text)
Task framing	ASR post-correction (seq2seq)
Max src / tgt length	200
Epochs	20 (early stopping on dev loss)
Batch size (per device)	8
Decoding	Beam search, `num_beams=10`
Model selection	`metric_for_best_model=loss`
Eval / Save strategy	`epoch`
Optimizer / LR / Scheduler	HF defaults (not overridden)

Table 7: Hyperparameters for the mT5-base ASR corrector (Hausa).

Outputs. At inference, the corrector takes WHISPER LARGE V3 outputs and returns corrected text. Training and decoding hyperparameters are summarized in Table 7.

B More Detailed POS-specific Metrics

In addition to WER, we compute the F-Based Position-independent Error Rate (FPER) (Popović and Ney, 2007), which disregards word order and instead captures errors in the distribution of POS classes. FPER is defined as:

$$\mathrm{FPER}(p) = \frac{1}{N^*_{\mathrm{ref}} + N_{\mathrm{hyp}}} \cdot \sum_{k=1}^{K} \left(n\left(p, \mathrm{rerr}_k\right) + n\left(p, \mathrm{herr}_k\right) \right) \quad (1)$$

where p is a POS class, N^*_{ref} and N_{hyp} are the reference and hypothesis token counts (excluding punctuation), and $n(\cdot)$ counts errors of class p in reference (rerr) or hypothesis (herr) for each sentence k. The metric gives the proportion of position-independent errors for p over the corpus. WER and FPER together capture complementary aspects of translation quality: WER is sensitive to word order and thus reflects overall sequence-level accuracy, while FPER disregards position and focuses on the distribution of POS-specific errors. Using both allows us to assess not only how closely a translation matches the reference in form, but also which linguistic categories contribute most to the errors, providing a more targeted diagnostic of system performance.

The POS-specific FPER results (Tables 8–9) complement WER by highlighting position-independent mismatches. T6 cuts errors sharply for PUNCT and DET, indicating fewer spurious or missing tokens regardless of order. Reductions for AUX and PROPN further suggest stronger preservation of grammatical auxiliaries and named enti-

POS Tag	Swahili	Hausa	Igbo	Luganda	Yorùbá	Avg
ADJ	4.11%	4.97%	5.09%	4.94%	4.41%	4.70%
ADP	4.79%	7.62%	6.27%	6.47%	7.23%	6.48%
ADV	1.66%	3.67%	2.96%	2.94%	2.24%	2.69%
AUX	2.48%	4.97%	4.85%	4.31%	6.51%	4.62%
CCONJ	1.22%	2.62%	2.82%	1.88%	2.41%	2.19%
DET	4.07%	8.03%	**11.16%**	8.84%	7.22%	7.86%
NOUN	**10.62%**	**15.38%**	**16.49%**	**15.89%**	**14.35%**	**14.55%**
NUM	0.92%	0.84%	1.01%	0.99%	0.69%	0.89%
PART	1.11%	1.73%	1.53%	1.95%	3.32%	1.93%
PRON	2.18%	5.24%	5.03%	3.58%	8.65%	4.94%
PROPN	2.75%	7.44%	4.37%	4.48%	4.73%	4.75%
PUNCT	4.20%	**11.45%**	**13.11%**	9.53%	**10.04%**	9.67%
SCONJ	0.64%	0.91%	0.78%	1.00%	1.39%	0.94%
VERB	5.23%	6.59%	5.97%	6.31%	9.20%	6.66%
Macro Avg	3.87%	5.87%	5.74%	5.45%	6.23%	5.43%
Weighted Avg	4.60%	8.21%	8.16%	7.32%	8.25%	7.31%

Table 8: **FPER** (↓) over English POS tags of translation by Whisper Baseline (T1) for all five African languages.

POS Tag	Swahili	Hausa	Igbo	Luganda	Yorùbá	Avg
ADJ	3.80%	4.59%	5.41%	5.16%	5.25%	4.84%
ADP	4.23%	5.06%	6.23%	5.76%	5.90%	5.44%
ADV	1.71%	1.98%	2.28%	2.33%	2.22%	2.10%
AUX	2.27%	2.87%	3.19%	3.01%	3.37%	2.94%
CCONJ	1.22%	1.71%	1.77%	1.64%	1.70%	1.61%
DET	3.49%	5.46%	5.85%	5.30%	5.54%	5.13%
NOUN	**10.17%**	**13.31%**	**15.33%**	**14.36%**	**14.37%**	**13.51%**
NUM	0.61%	1.01%	1.10%	1.00%	0.89%	0.92%
PART	1.04%	1.06%	1.42%	1.44%	1.41%	1.27%
PRON	1.90%	2.39%	2.70%	2.59%	2.91%	2.50%
PROPN	2.53%	3.48%	3.83%	3.68%	3.59%	3.42%
PUNCT	4.11%	4.44%	4.62%	5.53%	4.98%	4.74%
SCONJ	0.56%	0.64%	0.90%	0.83%	0.85%	0.76%
VERB	4.89%	6.03%	6.74%	7.09%	7.08%	6.37%
Macro Avg	2.99%	3.98%	4.54%	4.40%	4.55%	4.09%
Weighted Avg	4.26%	5.41%	6.14%	5.98%	6.01%	5.56%

Table 9: **FPER** (↓) over English POS tags of translation by our best recipe (T6) for all five African languages.

ties. Even NOUN exhibits modest improvements, consistent with its WER gains. Together, WER and FPER reveal that T6 improves both ordering accuracy and lexical coverage.

Beyond WER and FPER, Popović and Ney (2007) introduced two additional complementary diagnostics: Inflectional POS Error Rates (IFPER) and Missing Words Distribution.

IFPER evaluates morphological competence by identifying cases where a system produces the correct lemma but with the wrong inflection. As shown in Tables 10 and 11, this analysis highlights the POS categories most prone to inflectional errors, thus uncovering weaknesses not visible in WER/FPER alone.

Missing words analysis distinguishes between truly omitted words and those simply reordered. Results in Tables 12 and 13 indicate which grammatical categories are systematically underproduced. These findings can directly inform targeted improvements in model design, such as handling of phrase coverage and language modeling.

POS Tag	Swahili	Hausa	Igbo	Luganda	Yoruba	Average
ADJ	0.47%	0.14%	0.13%	0.21%	0.13%	0.22%
ADP	0.10%	0.07%	0.05%	0.10%	0.11%	0.09%
ADV	0.15%	0.40%	0.03%	0.04%	0.05%	0.13%
AUX	0.72%	2.21%	2.92%	2.04%	2.10%	2.01%
CCONJ	0.06%	0.02%	0.02%	0.03%	0.04%	0.03%
DET	0.13%	0.08%	0.05%	0.08%	0.12%	0.09%
NOUN	2.67%	1.02%	0.84%	2.25%	1.70%	1.70%
NUM	0.10%	0.03%	0.14%	0.06%	0.14%	0.09%
PART	0.17%	0.06%	0.07%	0.08%	0.12%	0.10%
PRON	0.28%	0.17%	0.08%	0.17%	0.30%	0.20%
PROPN	0.73%	0.94%	0.51%	0.67%	0.57%	0.68%
SCONJ	0.00%	0.01%	0.01%	0.00%	0.01%	0.01%
VERB	0.93%	0.44%	0.31%	0.54%	0.42%	0.53%

Table 10: IFPER (↓) over English POS tags of translation by T1 for all five African languages.

POS Tag	Swahili	Hausa	Igbo	Luganda	Yorùbá	Average
ADJ	0.50%	0.45%	0.46%	0.43%	0.42%	0.45%
ADP	0.12%	0.14%	0.12%	0.14%	0.15%	0.13%
ADV	0.25%	0.24%	0.16%	0.12%	0.12%	0.18%
AUX	0.74%	1.00%	1.25%	1.01%	1.32%	1.06%
CCONJ	0.08%	0.08%	0.06%	0.07%	0.06%	0.07%
DET	0.13%	0.19%	0.12%	0.13%	0.08%	0.13%
NOUN	3.57%	3.20%	3.08%	2.84%	2.65%	3.07%
NUM	0.24%	0.29%	0.24%	0.19%	0.23%	0.24%
PART	0.19%	0.09%	0.12%	0.11%	0.15%	0.13%
PRON	0.29%	0.23%	0.27%	0.13%	0.26%	0.24%
PROPN	0.82%	0.85%	0.83%	0.80%	0.63%	0.79%
SCONJ	0.01%	0.01%	0.02%	0.01%	0.02%	0.01%
VERB	1.01%	1.10%	1.00%	1.10%	1.04%	1.05%

Table 11: IFPER (↓) over English POS tags of translation by T6 for all five African languages.

POS Tag	Swahili	Hausa	Igbo	Luganda	Yorùbá	Total
ADJ	122	204	283	198	210	1017
ADP	163	233	357	264	226	1243
ADV	70	87	132	128	108	525
AUX	91	133	207	124	140	695
CCONJ	59	65	80	58	82	344
DET	146	241	283	234	272	1176
NOUN	302	449	642	483	503	2379
NUM	22	33	56	29	54	194
PART	39	48	93	59	56	295
PRON	93	119	207	111	123	653
PROPN	63	172	266	166	182	849
PUNCT	101	169	215	143	169	797
SCONJ	22	40	74	36	47	219
VERB	166	218	332	250	238	1204

Table 12: Missing word counts by POS tag for English POS tagging across the five African languages for T1 translations.

POS Tag	Swahili	Hausa	Igbo	Luganda	Yorùbá	Total
ADJ	118	184	237	295	308	1142
ADP	178	239	336	437	358	1548
ADV	66	91	137	186	159	639
AUX	74	93	145	173	203	688
CCONJ	77	85	121	131	137	551
DET	161	224	327	417	344	1473
NOUN	320	443	677	795	701	2936
NUM	20	47	62	47	52	228
PART	43	44	63	94	79	323
PRON	73	115	164	183	165	700
PROPN	69	173	172	185	209	808
PUNCT	109	170	231	253	266	1029
SCONJ	20	20	37	45	26	148
VERB	149	216	327	337	356	1385

Table 13: Missing word counts by POS tag for English POS tagging across the five African languages for T6 translations.