ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

Real-Time French Speech Translation into Four Congolese National Languages Using Whisper and Language-Specific LoRA Adapters for NLLB-200

Jean-Christophe B. Wizamo[1*]; Selain K. Kasereka[1,2,3*]

[1]Mathematics, Statistics and Computer Science Department, University of Kinshasa, Kinshasa, Democratic Republic of the Congo
[2]Department of Information Measurement Systems, Technical University of Sofia, Sofia, Bulgaria
[3]Artificial Intelligence, Big Data and Modeling Simulation Research Center (ABIL), Kinshasa, Democratic Republic of the Congo

Corresponding Authors: Jean-Christophe B. Wizamo*, Selain K. Kasereka*

Publication Date: 2026/11/13

Abstract: French is the sole official language of the Democratic Republic of the Congo, yet much of the population communicates daily in Lingala, Swahili, Tshiluba or Kikongo. This paper presents a cascaded system that transcribes spoken French with Whisper and translates it with NLLB-200-distilled-600M, specialized by one Low-Rank Adaptation (LoRA) adapter per target language on a shared frozen base model. A streaming front-end combines voice activity detection, overlapping windows and stable-prefix commitment, and the system is packaged as a Flutter application released on Google Play. The adapters were trained on 923,511 sentence pairs retained from 1.86 million after a cleaning pipeline that includes LASER alignment filtering. Against a quality threshold of 25 BLEU points fixed before experimentation, Kikongo (34.08), Lingala (31.50) and Swahili (28.33) pass, whereas Tshiluba reaches 21.49. Scores are unrelated to corpus size: Swahili, trained on eight times fewer pairs than Tshiluba, outperforms it by 6.8 points, and the retention rate under filtering, 94.8 percent for curated sources against 46.7 to 50.2 percent for mined bitext, points to alignment quality as the dominant factor. On three Android handsets, the released build, which serves inference remotely, meets a two-second budget only intermittently in live mode (1082 milliseconds at best, medians between 2.6 and 4.4 seconds) and uses 0.5 to 3.0 percent of the battery per ten minutes. The on-device chain, quantized to 8-bit integers and exported to ONNX, is specified with an estimated static footprint below one gigabyte but remains to be measured.

Keywords: *Low-Resource Machine Translation; Speech Translation; Whisper; NLLB-200; LoRA; Lingala; Swahili; Tshiluba; Kikongo; Edge AI.*

How to Cite: Jean-Christophe B. Wizamo; Selain K. Kasereka (2026) Real-Time French Speech Translation into Four Congolese National Languages Using Whisper and Language-Specific LoRA Adapters for NLLB-200. *International Journal of Innovative Science and Research Technology*, 11(11), 154-172. https://doi.org/10.38124/ijisrt/26nov245

I. INTRODUCTION

In an increasingly connected world, access to information, public services, healthcare and education still depends on the language in which these resources are delivered. The Democratic Republic of the Congo (DRC) illustrates this situation acutely: French is the sole official language, while the Constitution recognizes four national languages—Kikongo, Lingala, Swahili and Tshiluba [1]—which remain the preferred medium of everyday communication for a large part of the population. Because most digital and administrative content is produced in French, this asymmetry translates into an information gap that language technologies can help narrow [2], [3].

Foundation models have changed what is technically attainable for such languages. Whisper [4] provides robust multilingual speech recognition learned from 680,000 hours of weakly supervised audio, and NLLB-200 [5] translates among 200 languages, including the four DRC national languages. Applying these models to the Congolese context nevertheless meets two obstacles. The first is *data scarcity*: parallel corpora pairing French with Lingala, Tshiluba or Kikongo are small, noisy and dominated by a few domains [3], [6]. The second is *infrastructure*: connectivity is intermittent and mobile data are costly, so that a translation service usable in daily life must run on the entry-level smartphones that dominate the market, ideally without a permanent connection [7].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

These two obstacles interact. Larger models translate better but exceed the memory of a phone; smaller or compressed models fit but may translate worse; and a slow system breaks the rhythm of a conversation. We therefore frame the problem as a constrained trade-off. A speech signal A is transcribed by a recognition module, $T_{\text{fr}} = \mathcal{M}_{\text{ASR}}(A)$, and the transcript is translated by a translation module, $T_{\text{tgt}} = \mathcal{M}_{\text{NMT}}(T_{\text{fr}})$. The latency perceived by the user for one speech segment is

$$L_{\text{total}} = L_{\text{audio}} + L_{\text{inf}}(\mathcal{M}_{\text{ASR}}) + L_{\text{inf}}(\mathcal{M}_{\text{NMT}}) + L_{\text{post}}, \quad (1)$$

Where L_{audio} covers signal pre-processing, $L_{\text{inf}}(\cdot)$ the inference time of each module and L_{post} post-processing and display. The system must jointly satisfy

$$\begin{cases} \text{RAM}(\mathcal{M}_{\text{ASR}}) + \text{RAM}(\mathcal{M}_{\text{NMT}}) \le M_{\text{lim}}, \\ \text{BLEU}(\mathcal{M}_{\text{NMT}}) \ge Q_{\text{min}}, \\ L_{\text{total}} \le \tau_{\text{rt}}, \end{cases} \quad (2)$$

Where M_{lim} is the memory available to the application, Q_{min} the minimum acceptable translation quality and $\tau_{\text{rt}} \approx 2000$ ms the largest delay compatible with a fluid exchange.

➢ *Identification of the Research Gaps*

Despite the growing interest in African-language processing, several gaps remain:

- Existing Congolese systems are pair-specific. The closest work handles French–Lingala in both directions, for text and speech [2], but no system covers French speech into all four national languages.
- Multilingual evaluations rarely examine the quality of the training bitext, although audits show that automatically mined corpora for low-resource languages frequently contain misaligned or wrong-language pairs [8]. The assumption that more parallel data yields better translations is seldom tested across languages within a single protocol.
- Studies on African-language translation report translation metrics but rarely system metrics—latency, memory, energy—on the entry-level devices that dominate the target market.
- Training-infrastructure constraints, such as an intermittent power supply, shape the experimental budget in many African laboratories, yet they are rarely documented, which makes results difficult to interpret.

➢ *Research Questions and Hypotheses*

This study addresses two research questions. RQ1 (adaptation): how can a cascade of Whisper and NLLB-200 be adapted to translate French into Lingala, Swahili, Tshiluba and Kikongo despite the scarcity of parallel data? RQ2 (embedded execution): which compression techniques allow the cascade to satisfy the memory and latency constraints of Equation (2) on a smartphone?

Two hypotheses, with thresholds fixed *before* any experiment, operationalize these questions:

- H1: freezing NLLB-200 and training only low-rank adapters $\Delta\theta$ with $|\Delta\theta| \ll |\theta_{\text{NMT}}|$ [9] yields BLEU $\ge Q_{\text{min}} = 25$ for each target language.
- H2: 8-bit post-training quantization [10], [11] executed with ONNX Runtime Mobile keeps the cascade below $M_{\text{lim}} \approx 1$ GB and $\tau_{\text{rt}} = 2000$ ms, with a relative degradation of WER (ASR) and BLEU (NMT) below $\epsilon = 0.05$.

The thresholds are maintained unchanged when confronted with the measurements: a threshold revised after the fact would turn a refutable hypothesis into an unfalsifiable statement.

➢ *Contributions of the Paper*

This paper makes the following contributions:

- A cascaded Whisper + NLLB-200 architecture in which four language-specific LoRA adapters share one frozen base model, together with a streaming front-end (voice activity detection with hysteresis, 3 s windows with a 1 s hop and stable-prefix commitment);
- A reproducible cleaning and partitioning pipeline for French-centric parallel corpora into the four national languages, with retention statistics per source family;
- An evaluation of the four adapters (BLEU, chrF, WER, token accuracy) showing that the ranking of scores is unrelated to corpus volume, and identifying the retention rate under alignment filtering as a measurable proxy of corpus quality;
- A Flutter application released on Google Play, whose latency and energy consumption are measured on three Android devices, with an explicit separation between what was measured (a build with remote inference) and what is specified (the on-device INT8 ONNX chain).

➢ *Paper Organization*

Section II reviews related work. Section III describes the architecture, the corpus, the adaptation and compression procedures, the mobile application and the evaluation protocol. Section IV reports the results, Section V discusses them against the hypotheses, and Section VI concludes and outlines future work.

II. RELATED WORK

➢ *Cascaded and End-to-End Speech Translation*

Speech translation (ST) systems are either *cascaded*, chaining automatic speech recognition (ASR) and machine translation (MT), or *end-to-end*, mapping source speech directly to target text [12], [13]. End-to-end models avoid error propagation between stages and can exploit prosody, but they require paired speech–translation data that do not exist for most African languages. Cascades remain competitive when strong components can be trained on separate resources, and they are easier to diagnose and update module by module [14]. Massively multilingual models such as SeamlessM4T [15] and MMS [16] extend speech technologies to hundreds of languages, but their size (billions of parameters) puts them out of reach of entry-level phones. Whisper [4] offers robust

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

recognition in about a hundred languages, but its built-in translation task only targets English, so French-to-Lingala translation still requires an MT stage.

For live use, the recognizer must process an unbounded stream. Simultaneous translation policies such as wait-*k* [17] trade latency for quality, and Macháček et al. [18] turned Whisper into a streaming recognizer by re-decoding a growing buffer and committing only the prefix on which consecutive hypotheses agree. Our front-end follows the same principle with fixed overlapping windows and a longest-common-subsequence alignment between successive hypotheses.

➢ *Multilingual Neural Machine Translation and African Languages*

Neural MT moved from recurrent encoder–decoders with attention [19], [20] to the Transformer [21], and then to single models covering many languages [22]–[24]. NLLB-200 [5] covers 200 languages, relies on large-scale bitext mining with multilingual sentence encoders [25], and is evaluated on FLORES-200 [26]. Its distilled 600M-parameter variant is the smallest member of the family and the natural candidate for constrained deployment.

For African languages, the Masakhane community demonstrated participatory MT research across the continent [3], and MAFAND-MT showed that a few thousand high-quality in-domain sentences can substantially improve pre-trained models on news translation [27]. Religious corpora such as JW300 [6] long supplied most of the parallel text for Tshiluba and Kikongo, at the cost of a strong domain bias. A recurring finding is that data quality matters as much as quantity: Kreutzer et al. [8] audited web-crawled multilingual corpora and found that, for many low-resource languages, a large fraction of the pairs were misaligned, in the wrong language or not linguistic content at all. Our results provide a controlled observation of this effect on four Congolese languages.

➢ *Parameter-Efficient Adaptation*

Full fine-tuning of a 600M-parameter model on small corpora is costly and prone to overfitting and catastrophic forgetting. Parameter-efficient methods insert small trainable modules into a frozen network: bottleneck adapters [28], language-pair adapters for NMT [29], and Low-Rank Adaptation (LoRA), which learns a low-rank update of selected weight matrices [9]. Because the base model is shared, one adapter per language can be swapped at run time, a property that suits mobile deployment where only one base model can reside in memory.

➢ *Efficient On-Device Inference*

Mobile edge computing moves computation close to the user to reduce latency, bandwidth and dependence on the network [7]. Quantization is the most effective compression technique for this purpose: mapping 32-bit floating-point weights and activations to 8-bit integers divides memory by four and enables integer arithmetic [10], [11], [30]. Transformers are sensitive to activation outliers, which motivates per-layer calibration and mixed precision [31], [32]. Distillation offers a complementary route [33], [34]. The ONNX exchange format and ONNX Runtime provide graph-level optimizations (operator fusion, constant folding) and mobile execution providers [35].

➢ *Lingala and Other Congolese Languages in NLP*

Work on Congolese languages has recently accelerated. Mandiya et al. [2] proposed a bidirectional French–Lingala pipeline combining LSTM, Transformer, BERT and Whisper on religious and medical corpora; BabaSpeech translates sign language into Lingala text and speech [36]; and deep learning has been applied to printed Lingala script recognition [37]. Table 1 positions the present work with respect to the closest systems.

Table 1 Positioning of this Work with Respect to Closely Related Systems

System	Task	Languages	Architecture	Metrics	Mobile/edge
Whisper [4]	ASR; speech-to-English ST	About 100	Transformer encoder–decoder, weak supervision	WER, BLEU	Server-oriented; small variants exist
NLLB-200 [5]	Text MT	200, incl. the four DRC languages	Dense and MoE Transformers	BLEU, chrF++	Distilled 600M variant
SeamlessM4T [15]	Speech and text MT	About 100	Unified multimodal model	BLEU, ASR-BLEU	No (billions of parameters)
MAFAND-MT [27]	News MT	16 African languages	Fine-tuned pre-trained models	BLEU, chrF	No
Streaming Whisper [18]	Real-time ASR	Whisper languages	Re-decoding with local agreement	WER, latency	GPU server
Mandiya et al. [2]	FR↔LN MT, text and speech	FR, LN	LSTM, Transformer, BERT, Whisper	BLEU, chrF, WER	Web interface
This work	FR speech → text in four languages, real time	FR → LN, SW, LU, KG	Whisper + NLLB-200 with four LoRA adapters; INT8 ONNX export	BLEU, chrF, WER, latency, energy	Flutter app (Google Play); on-device path specified

ST: speech translation; MT: machine translation; FR: French; LN: Lingala; SW: Swahili; LU: Tshiluba; KG: Kikongo.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

Compared with [2], of which the first author is a co-author, the present work keeps the cascaded principle but replaces task-specific models trained from scratch with a single pre-trained multilingual model, extends the target side from one to four languages through dedicated adapters, and adds the compression and mobile-integration layer that the earlier study did not address.

III. MATERIALS AND METHODS

➢ *Motivation for a Cascaded, Adapter-Based Design*

Three considerations motivate a cascade of pre-trained models specialized by small adapters, rather than an end-to-end model or one fully fine-tuned model per language.

- *Data availability*. No corpus pairs French speech with text in the four target languages, which rules out end-to-end speech translation training. A cascade can instead exploit French speech recognition, which is well covered by Whisper, and French-centric text bitext, which exists for each target language.
- *Memory*. Four fully fine-tuned copies of a 600M-parameter model would require about 10 GB in single precision. With LoRA, a single frozen base is kept in memory and only a small adapter is swapped when the user changes the target language.
- *Modularity*. Recognition and translation can be evaluated, replaced or compressed independently, which eases error attribution (transcription versus translation) and maintenance in a resource-constrained deployment context.

➢ *System Architecture*

The system is a three-stage cascade: acquisition and segmentation of the audio stream, speech recognition, and adapted neural translation (Fig. 1).

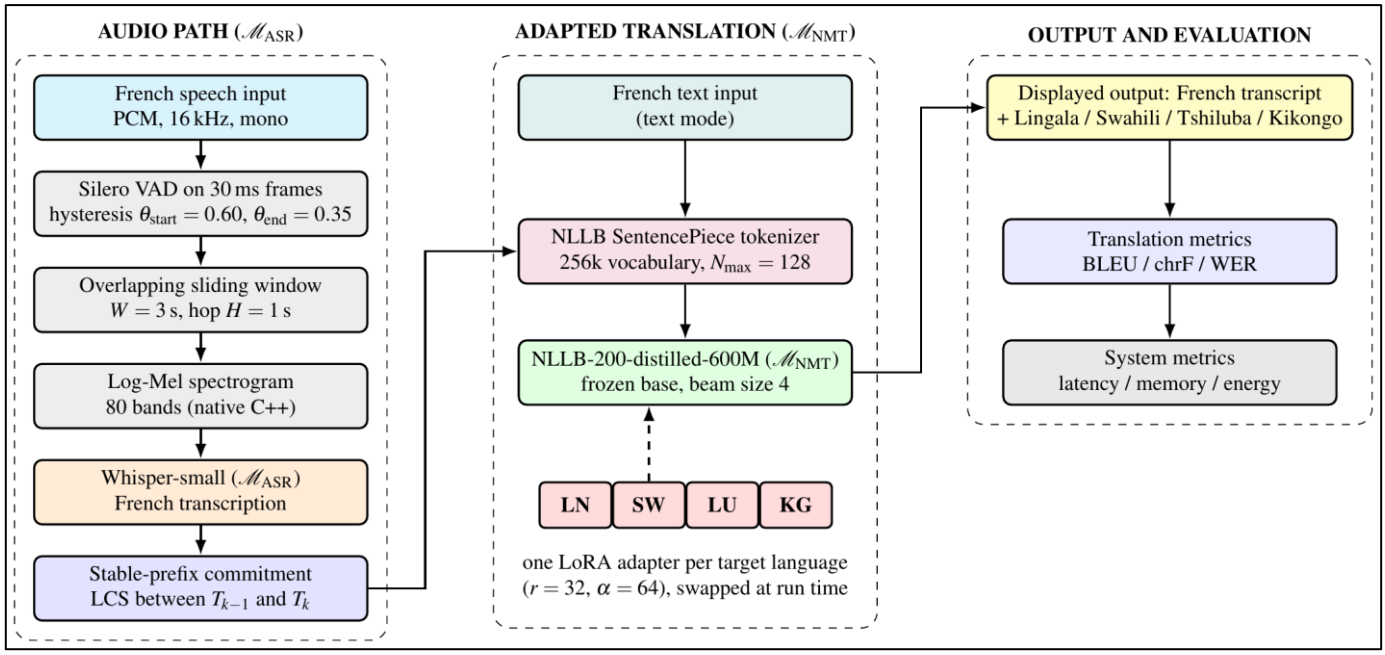


Fig 1 Architecture of the proposed system. The audio path segments the French speech stream and transcribes it with Whisper; the committed transcript (or a typed French text) is translated by a frozen NLLB-200 model specialized by the LoRA adapter of the selected target language.

Let $X = (x_1, \ldots, x_T)$ denote the log-Mel frames of a speech segment, with $x_t \in \mathbb{R}^{80}$. Whisper generates the French token sequence $Y_{fr} = (y_{fr}^1, \ldots, y_{fr}^U)$ autoregressively:

$$P(\mathbf{Y}_{fr} \mid \mathbf{X}) = \prod_{u=1}^{U} P\left(y_{fr}^u \mid y_{fr}^{<u}, \mathcal{E}_{ASR}(\mathbf{X})\right), \quad (3)$$

Where $\mathcal{E}_{ASR}$ is the acoustic encoder (two strided one-dimensional convolutions followed by Transformer blocks) and the decoder is prompted with the French language token and the *transcribe* task token. The committed French text is then tokenized and translated by NLLB-200:

$$P\left(\mathbf{Y}_{tgt} \mid \mathbf{Y}_{fr}, \ell\right) = \prod_{v=1}^{V} P\left(y_{tgt}^v \mid y_{tgt}^{<v}, \mathcal{E}_{NMT}(\mathbf{Y}_{fr}), \ell; \boldsymbol{\theta}_{NMT} + \Delta\boldsymbol{\theta}_\ell\right), \quad (4)$$

Where $\ell \in$ {lin_Latn,swh_Latn,lua_Latn,kon_Latn} is the target language, enforced as the first decoder token, and $\Delta\boldsymbol{\theta}_\ell$ the corresponding LoRA adapter. The main risk of a cascade is error propagation: a misrecognized word perturbs the source representation $\mathcal{E}_{NMT}(\mathbf{Y}_{fr})$. Beam-search decoding and the stable-prefix mechanism described below are meant to limit the transmission of unstable hypotheses to the translation stage.

➢ *Streaming Front-End*

Submitting long segments to Whisper increases latency and processor load, while a fixed segmentation may cut words and create recognition errors that propagate to translation. The front-end therefore combines voice activity detection (VAD) with overlapping windows.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

- *Voice Activity Detection*

The raw microphone stream (mono, 16 kHz, 16-bit PCM) is split into 30 ms frames $\mathbf{f}_k$ (480 samples), for which the lightweight Silero VAD [38] estimates a speech probability $p_k \in [0,1]$. A hysteresis avoids rapid toggling:

$$s_k = \begin{cases} \text{speech}, & p_k > \theta_{\text{start}}, \\ \text{silence}, & p_k < \theta_{\text{end}}, \\ s_{k-1}, & \text{otherwise}, \end{cases} \qquad \theta_{\text{start}} = 0.60, \quad \theta_{\text{end}} = 0.35. \tag{5}$$

Only speech frames enter the buffer, which saves energy during silences and suppresses noise-only segments that are known to trigger Whisper hallucinations; a silence longer than $\tau_{\text{silence}} = 500$ ms closes the current utterance.

- *Overlapping Sliding Window*

During continuous speech, the buffer is processed by a window of maximum length $W_{\max} = 3$ s advanced by a hop $H = 1$ s, i.e., with an overlap $O = W_{\max} - H = 2$ s:

$$\mathbf{B}_t = \Big[\underbrace{\mathbf{x}_{t-W_{\max}+1}, \dots, \mathbf{x}_{t-H}}_{\text{context}}, \; \underbrace{\mathbf{x}_{t-H+1}, \dots, \mathbf{x}_t}_{\text{new audio}} \Big]. \tag{6}$$

The overlap preserves the acoustic context at window boundaries and stabilizes recognition.

- *Stable-Prefix Commitment*

Because consecutive windows overlap, consecutive hypotheses T_{k-1} and T_k share words. Translating every hypothesis would duplicate text and propagate unstable endings. Following the local-agreement principle [18], the words of T_{k-1} are aligned with those of T_k by a longest common subsequence (LCS); the committed increment Δ_k consists of the words of T_{k-1} that follow the text already committed and are confirmed by T_k (Fig. 2). The pipeline proceeds as follows:

- ✓ 30 ms frames are written to a circular buffer only when the VAD reports speech;
- ✓ every second, a 3 s window is extracted, converted into an 80-band log-Mel spectrogram and transcribed by Whisper into a partial hypothesis T_k;
- ✓ T_k is aligned with T_{k-1} by LCS to obtain the confirmed increment Δ_k;
- ✓ Δ_k is sent to NLLB-200 for translation, and the oldest $H = 1$ s of audio is discarded from the buffer.

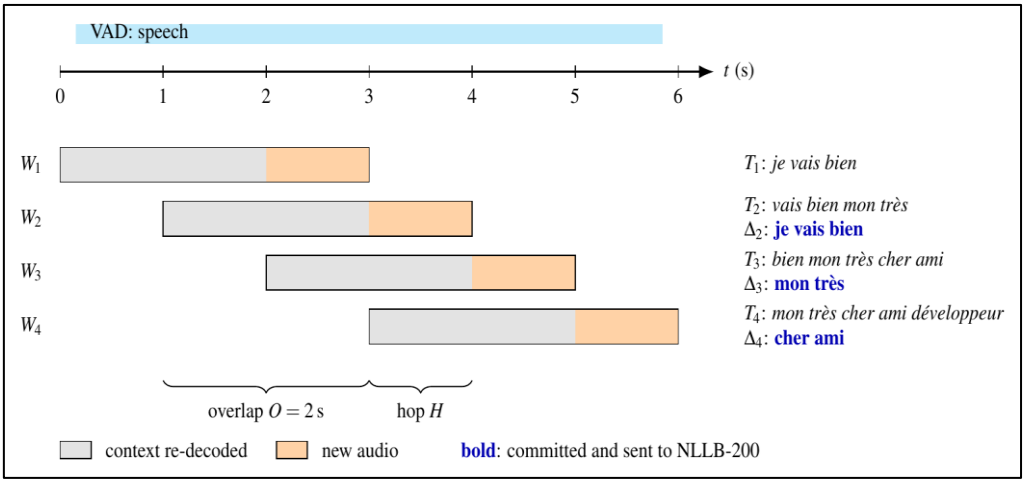


Fig 2 Streaming segmentation and stable-prefix commitment, illustrated on the French utterance *je vais bien mon très cher ami développeur*. Each 3 s window W_k is re-decoded every second; only the words confirmed by two consecutive hypotheses (Δ_k) are translated.

➢ *Parallel Corpus Construction and Cleaning*

No single parallel corpus covers French and the four national languages. We therefore aggregated sources of different nature (Table 2). Three pairs rely on the NLLB bitext [5], obtained by automatic alignment of comparable documents at web scale. The Swahili side aggregates curated sources: manually verified sentences from Tatoeba [39], translated news from GlobalVoices, TED talk transcripts, the reviewed health-domain corpus TICO-19 [40] and Wikimedia, all distributed through OPUS [41]. This asymmetry of provenance—automatic alignment on one side, human curation on the other—is a variable that pair counts alone do not capture.

Table 2 Origin of the Parallel Corpora by Translation Direction

Pair	Main source	Raw pairs
fr–ln (Lingala)	NLLB bitext (automatically aligned)	713,599
fr–lu (Tshiluba)	NLLB bitext, direction fr–lua	592,536
fr–kg (Kikongo)	NLLB bitext (automatically aligned)	515,359
fr–sw (Swahili)	Tatoeba (fr–swc), GlobalVoices, TED, TICO-19, Wikimedia	35,648
Total		1,857,142

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

The cleaning script applies, in order: Unicode and whitespace normalization; removal of HTML tags and URLs; removal of empty, too short or too long sentences; rejection of pairs whose length ratio exceeds 3; rejection of pairs whose source and target are identical; deduplication; and filtering by LASER alignment score [25] with a threshold of 1.05. Each pair is then split 80/10/10 into training, validation and test sets with a fixed random seed (42), which makes the partition reproducible (Table 3, Fig. 3).

Table 3 Effect of the Cleaning Pipeline and Resulting Partitions (seed 42)

Pair	Raw	Kept	Retention	Train	Valid.	Test
fr–ln	713,599	357,866	50.2%	286,292	35,786	35,788
fr–lu	592,536	276,579	46.7%	221,263	27,657	27,659
fr–kg	515,359	255,279	49.5%	204,223	25,527	25,529
fr–sw	35,648	33,787	**94.8%**	27,029	3378	3380
Total	1,857,142	923,511	49.7%	738,807	92,348	92,356

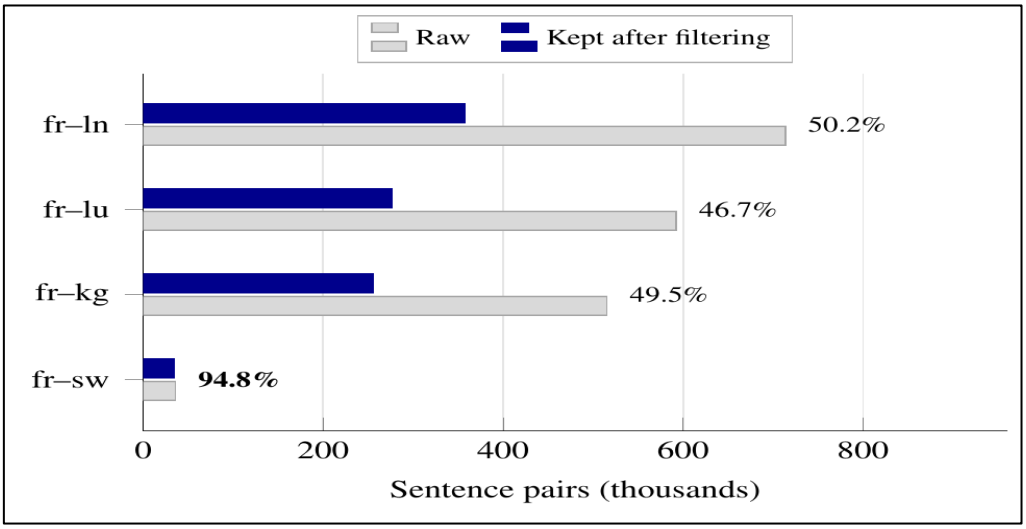


Fig 3 Raw and Retained Volumes Per Language Pair. Automatically Mined Bitext Loses About Half of its Pairs Under Filtering, Whereas the Curated Swahili Sources Keep 94.8%.

Two observations follow. First, cleaning removes about half of the NLLB bitext (retention between 46.7% and 50.2%), mainly through the LASER threshold and the length-ratio filter, whereas the curated Swahili corpus keeps 94.8% of its pairs. This gap of nearly fifty points is a direct measurement, not a conjecture, of the difference in alignment quality between the two source families, in line with audits of web-mined corpora [8]. Second, the volume hierarchy runs against common expectations: after cleaning, Lingala has the largest corpus (357,866 pairs) and Swahili the smallest (33,787), a ratio above ten. Swahili is better resourced than the Congolese languages at web scale, but not in a French-centric parallel corpus that has passed a strict alignment filter.

➢ *Tokenization and Sequence Length*

The shared SentencePiece tokenizer of NLLB-200 [42] (vocabulary of 256k subword units) is retained: a language-specific vocabulary would require re-learning the embedding layer, which is incompatible with a frozen base model. Bantu morphology is agglutinative—a single verb form may encode subject, object, tense, aspect and negation, as in Swahili *hawajatuona* ('they have not yet seen us')—so subword segmentation is essential [43]. More than 98.5% of the sentences of the corpus are shorter than 64 tokens. The maximum length was nevertheless set to $N_{\max} = 128$ tokens on both sides and kept at export, to avoid silent truncation of longer user inputs and any train–inference mismatch. Compared with a typical limit of 512 tokens, this reduces the size of the attention matrices, whose cost is quadratic in sequence length, by a factor $(128/512)^2 = 1/16$.

Language-Specific LoRA Adaptation of NLLB-200

Full fine-tuning of NLLB-200-distilled-600M is avoided because of its cost and of the risk of overfitting small corpora. LoRA [9] keeps each pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ frozen and learns a low-rank update:

$$\mathbf{h} = \mathbf{W}_0\mathbf{x} + \Delta\mathbf{W}\mathbf{x} = \mathbf{W}_0\mathbf{x} + \frac{\alpha}{r}\mathbf{B}\mathbf{A}\,\mathbf{x}, \quad \mathbf{A} \in \mathbb{R}^{r\times k}, \quad \mathbf{B} \in \mathbb{R}^{d\times r}, \tag{7}$$

Where A is initialized from a Gaussian distribution and B with zeros, so that $\Delta W = 0$ at the start of training. Adapters are injected into the four attention projections (q_proj, k_proj, v_proj, out_proj) of every attention block, i.e., encoder self-attention and decoder self- and cross-attention (Fig. 4). The more common restriction to q_proj and v_proj was discarded to widen the adaptation capacity, the target languages being morphosyntactically distant from the languages that dominate pre-training. The rank is $r = 32$ and the scaling factor $\alpha = 64 = 2r$.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

Each adapted 1024×1024 projection adds $r(d + k) = 65{,}536$ parameters. With 144 adapted projections (12 encoder layers × 4 and 12 decoder layers × 8), an adapter contains 9,437,184 trainable parameters, about 1.5% of the ≈615M frozen parameters, i.e., about 38 MB in single precision and about 9 MB after 8-bit quantization.

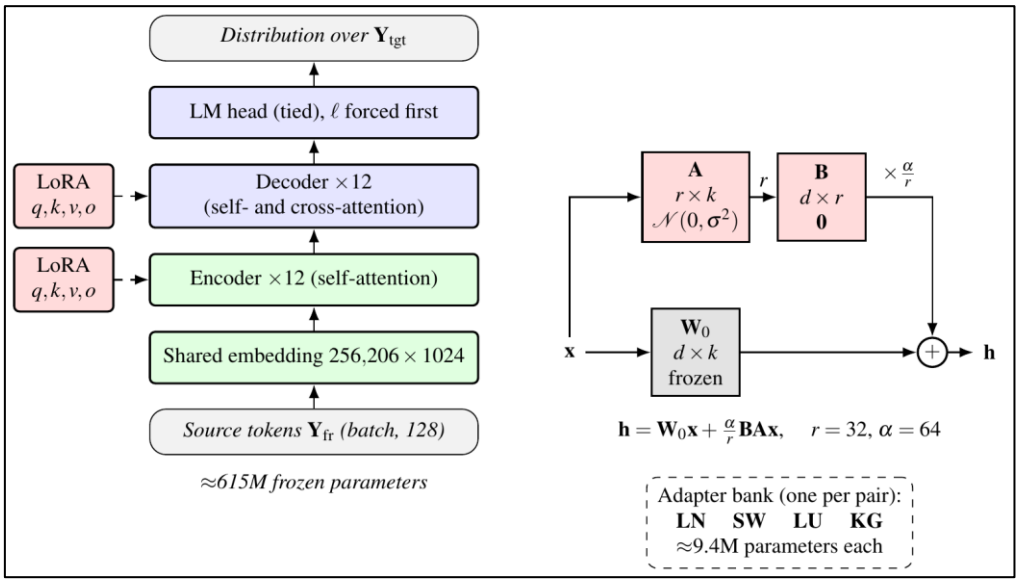


Fig 4 NLLB-200-Distilled-600M with LoRA Adapters Injected into the Four Attention Projections of Every Encoder and Decoder Layer (Left), and the Low-Rank Update Applied to one Frozen Projection (Right).

Four adapters are trained separately, one per language pair, on top of the same frozen base. A temperature-based sampling scheme [23] had been designed for joint multilingual training; it becomes unnecessary with per-pair adapters, since no language can dominate a shared loss. The imbalance does not vanish, however: it turns into an inequality of training budget between adapters (286,292 training pairs for Lingala versus 27,029 for Swahili). Under the assumption that quality grows with parallel data, scores should follow this order; Section IV tests this expectation.

➢ *Training Configuration*

Table 4 lists the training configuration, shared by the four adapters. Training uses AdamW [44], mixed precision (fp16), a linear warm-up followed by the default linear decay of the training library, weight decay and adapter dropout as regularizers, and per-epoch evaluation with selection of the checkpoint that maximizes validation BLEU. Label smoothing and gradient clipping, considered at design time, were not used.

Table 4 Training Configuration of the LoRA Adapters

Setting	Value
Base model	facebook/nllb-200-distilled-600M (frozen)
Source / target languages	fra_Latn → lin_Latn, swh_Latn, lua_Latn, kon_Latn
Adaptation	LoRA, $r = 32$, $\alpha = 64$, dropout 0.05
Target modules	q_proj, k_proj, v_proj, out_proj
Maximum length (source / target)	128 / 128 tokens
Epochs	12 (fixed budget, see below)
Batch size / gradient accumulation	4 / 8 (effective batch of 32 sentences)
Learning rate / weight decay	1×10^{-4} / 0.01
Warm-up ratio	0.03 (0.06 for fr–sw)
Optimizer / precision	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) / fp16
Model selection	best validation BLEU (eval_bleu)
Decoding	beam search, beam size 4
Random seed	42

- *Training Budget Under an Intermittent Power Supply.*

A full cycle of 30 epochs on the four pairs represents about 36 hours of uninterrupted computation, which is incompatible with the frequent and unplanned power cuts of the Kinshasa grid in the absence of a backup infrastructure. The budget was therefore fixed *a priori* at 12 epochs, a value that fits a reasonably safe power window while letting the loss curves reach a usable regime. All four runs were completed within this budget. This constraint belongs to the very context the study addresses—designing AI systems for the DRC means coping with an intermittent energy infrastructure—and it bounds the scope of the results: the reported scores characterize what the method reaches under this budget, not its ceiling.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

➢ *Model Compression for Edge Deployment*

- *Static Post-Training Quantization*

To fit the cascade into the memory of a phone, both models are quantized to 8-bit integers after training [10], [11]. Static quantization is preferred to dynamic quantization: scales and zero points are computed once, offline, and no calibration is performed on the device. For a weight tensor W, symmetric uniform quantization reads.

$$\begin{aligned} \Delta_W &= \frac{\max|\mathbf{W}|}{2^{b-1}-1}, \qquad b = 8, \\ \mathbf{W}_{\text{int8}} &= \text{clip}\left(\left\lfloor \frac{\mathbf{W}}{\Delta_W} \right\rceil, -128, 127\right), \end{aligned} \tag{8}$$

Where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. Activations depend on the input and require calibration. The calibration set contains 512 sentences drawn from the validation sets, stratified across the four languages. For attention and linear projections, the clipping threshold minimizes the Kullback–Leibler divergence between the FP32 and quantized activation distributions [45],

$$T^* = \arg\min_T D_{\text{KL}}\left(P_{\text{FP32}} \parallel Q_{\text{INT8}}(T)\right), \tag{9}$$

Which is robust to the outliers frequent in attention scores [32]; min–max calibration is kept for the convolutional layers of the Whisper encoder, whose activations are more regular. Layer normalization and the attention softmax remain in FP32 (mixed precision), while linear projections and feed-forward layers, which hold most parameters, are quantized to INT8 [35].

- *ONNX Export and Graph Optimization*

Both models are exported with torch.onnx.export (opset 17), with a dynamic axis on the batch dimension only and $N_{\max} = 128$ for the translation model. The graph is then optimized with the ONNX Runtime transformer optimizer: fusion of multi-head attention into a single kernel, fusion of layer normalization, dead-node elimination and constant folding, and fusion of bias addition with GELU in the feed-forward blocks. These fusions reduce the number of graph nodes, hence the scheduling overhead of the mobile runtime.

Table 5 gives the resulting *static* footprint, computed from parameter counts and bit widths. It is a lower bound of the memory needed at run time, which also includes activation buffers, the decoder key–value cache and the allocation overhead of the runtime; the peak allocation therefore has to be measured on the device (Section V-D).

Table 5 Static Memory Footprint Before and After 8-bit Quantization

Component	Parameters	FP32	INT8
Whisper-small ($\mathcal{M}_{\text{ASR}}$)	≈244M	≈976 MB	≈244 MB
NLLB-200-distilled-600M (frozen base)	≈615M	≈2460 MB	≈615 MB
LoRA adapter ($r = 32$, one language)	≈9.4M	≈38 MB	≈9 MB
Total	**≈868M**	**≈3.5 GB**	**≈0.87 GB**

Analytic estimate from parameter counts, with one target language loaded.

➢ *Mobile Application*

The application is written in Flutter (Dart) and organized in four layers that separate the user interface, the application logic, the inference worker and the native bridge (Fig. 5). Audio capture and log-Mel extraction are implemented in C++ and called through dart:ffi, which avoids the serialization overhead of Flutter platform channels for large binary buffers and reduces L_{audio} in Equation (1); memory allocated on the native side is released explicitly, since it escapes the Dart garbage collector. Inference is confined to a dedicated *isolate*—a Dart execution unit with its own heap that communicates by message passing (SendPort/ReceivePort)—so that the UI thread, responsible for rendering at 60 frames per second, is never blocked.

Only the adapter of the selected language is merged into the frozen base (lazy loading); a change of target language unloads the current adapter, loads the new one from local storage and merges it into the four projections. Given the adapter size, this operation is expected to take tens of milliseconds and occurs only outside the translation loop; it was not instrumented separately. To mitigate thermal throttling, the hop H is temporarily increased whenever the latency of a segment exceeds 1.5 times the mean of the last ten segments, a graceful degradation preferred to an abrupt interruption by the operating system.

The application offers three modes from a single home screen: record-and-translate, typed-text translation, and continuous (live) translation, with a common selector of the target language. It is published on Google Play as *Switch Trans Realtime*.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

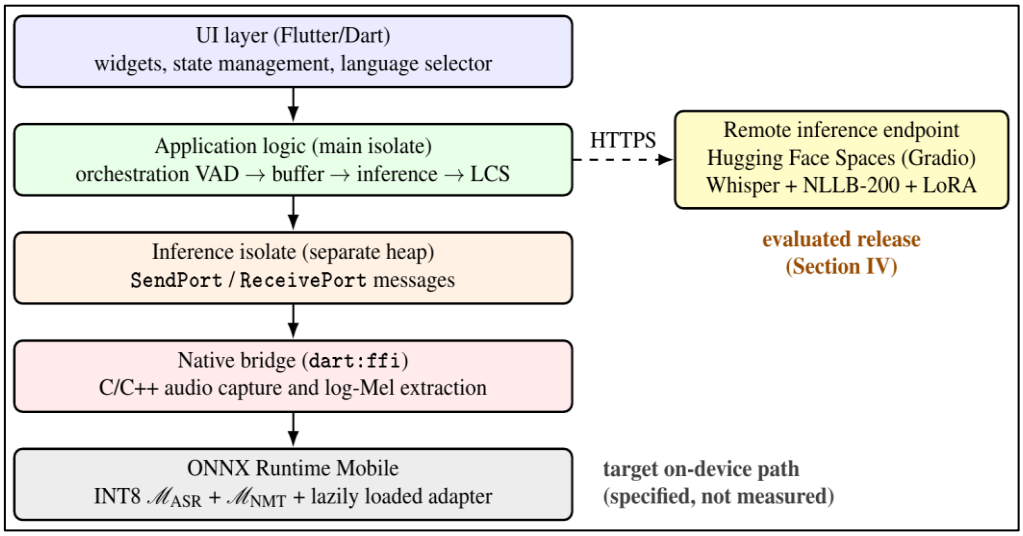


Fig 5 Software architecture of the mobile application. In the evaluated release, the application logic delegates inference to a remote endpoint (dashed arrow); the on-device path through the inference isolate, the native bridge and ONNX Runtime Mobile is the deployment target.

- *Implementation Status*

In the release distributed on Google Play and evaluated in Section IV, the inference of $\mathcal{M}_{ASR}$ and $\mathcal{M}_{NMT}$ is *delegated to a remote service* hosted on Hugging Face Spaces and exposed through a Gradio interface [46]. The application performs capture, windowing, orchestration and display, but does not execute the quantized ONNX graph locally. This choice is consistent with the objective of the study, whose binding constraint is temporal rather than topological: a remote architecture satisfies the real-time requirement as long as the end-to-end latency stays within budget, and it decouples the service from the capability of the handset, an asset in a market dominated by entry-level devices. Offline execution brings other benefits—use in areas without coverage and no data cost—and constitutes a distinct objective. Two consequences follow for the interpretation of the results: the reported latencies include network transport, possible cold starts of the remote container and inference on shared CPUs, so they characterize the deployed system but not the quantized on-device chain; and the components specific to local execution (FFI bridge, inference isolate, lazy adapter loading, adaptive throttling) are implemented but not exercised by the measured path. Fig. 6 summarizes the complete workflow, from raw corpora to deployment.

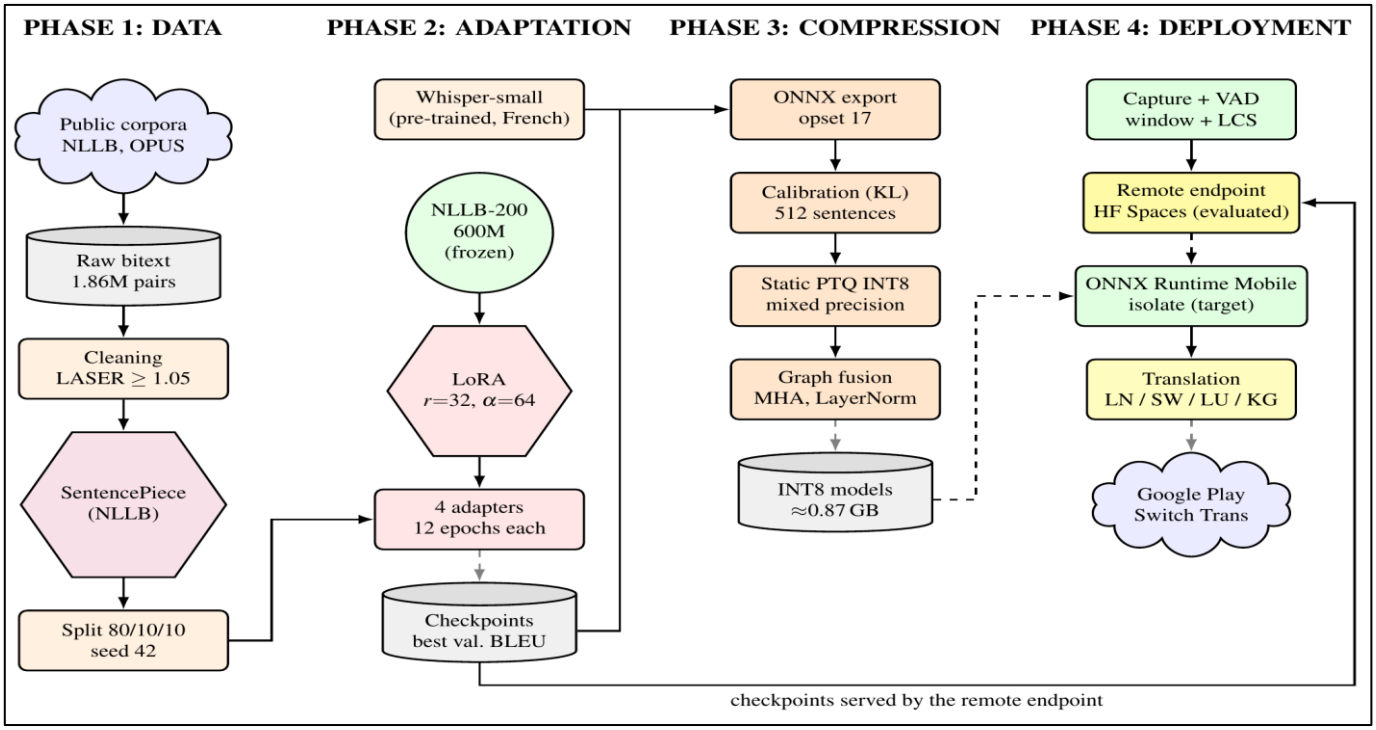


Fig 6 End-to-end workflow, from public corpora to the released application. Shapes encode the nature of each step: cloud, external service; cylinder, stored data or checkpoints; hexagon, tokenization or adaptation; circle, neural network; rectangle, processing step. Dashed arrows lead to the on-device target path.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

➢ *Evaluation Protocol*

- *Translation Quality*

Translation is evaluated on the test partitions of Table 3, which are disjoint from the training and validation sets and of very unequal size: 35,788 sentences for Lingala, 27,659 for Tshiluba, 25,529 for Kikongo and 3380 for Swahili. Each test set inherits the domain of its corpus: the Swahili test set comes from curated collections, the other three from mined bitext. Scores are thus not measured on comparable distributions, and differences between languages cannot be read as pure differences of linguistic difficulty. The metrics are:

✓ BLEU [47] and chrF [48], computed with SacreBLEU [49]; chrF, based on character n-grams, is less penalizing than BLEU for agglutinative languages in which a correctly translated root may carry a wrong inflection, and correlates better with human judgements [50];

✓ Word error rate (WER) of the *translation output* against the reference, computed with jiwer; this is a word-level edit distance on the translation, not the ASR error rate;

✓ Token-level accuracy under teacher forcing on the validation set, reported for completeness only, since it is not comparable with free-running metrics.

- *System Performance*

Latency components of Equation (1) are timestamped by internal instrumentation and displayed on a dedicated *device metrics* screen; L_{post} is not isolated and is merged into $L_{\text{inf}}(\mathcal{M}_{\text{NMT}})$. Energy is measured as the battery drop over a 10 min live-translation session. Memory was to be profiled with the Android Profiler. Measurements were collected on three Android handsets (Table 6): two entry-level TECNO devices sharing the same MediaTek Helio G81 system-on-chip, which isolates the effect of memory and Android version at constant compute, and a high-end Samsung Galaxy S22 Ultra serving as an upper bound.

Table 6 Devices Used for System Measurements

Device	Segment	SoC	RAM	Android
TECNO Spark 50 (KN4)	Entry-level	Helio G81	4 GB	16
TECNO Spark 40 (KM5)	Entry-level	Helio G81	6–8 GB	15
Galaxy S22 Ultra (SM-S908E)	High-end	Snapdragon 8 Gen 1	8–12 GB	16

- *Success Criteria*

The thresholds of H1 ($Q_{\min} = 25$ BLEU) and H2 ($M_{\text{lim}} \approx 1$ GB, $\tau_{\text{rt}} = 2000$ ms, $\epsilon = 0.05$) were set before experimentation by transposing reference points from the MT literature. The 25-BLEU mark was established on high-resource pairs with standard test sets, and nothing guarantees its transferability to French–Bantu pairs evaluated with a single reference; BLEU is sensitive to segmentation, to target-side morphological richness and to the number of references, three factors that all disadvantage the languages studied here. The thresholds are nevertheless kept unchanged. Three complementary indicators are declared in addition, not as replacements: chrF; the gain relative to the unadapted NLLB-200 on the same test set; and the dispersion of scores across languages, whose ranking tests the implicit assumption that quality grows with parallel data. Finally, the latency threshold concerns the *embedded* inference chain; measurements on a build that delegates inference to a remote service include components foreign to the hypothesis and cannot constitute a strict test of it.

➢ *Tools and Technologies*

Training, export and evaluation were implemented in Python with Hugging Face Transformers [51] and PEFT [52], ONNX Runtime tools for export and quantization, SacreBLEU and jiwer for evaluation, and Gradio on Hugging Face Spaces for the inference service. The mobile application uses Flutter (Dart), dart:ffi with C/C++ for audio processing, and ONNX Runtime Mobile for the on-device path.

IV. RESULTS

➢ *Training Convergence*

Fig. 7 shows the training and validation losses over the 12 epochs. The training loss decreases for all pairs, but three distinct dynamics appear. The fr–kg pair has the healthiest profile: both curves decrease together, from 2.39 to 1.85 (training) and from 2.27 to 1.89 (validation), and remain merged over the whole interval, with no sign of overfitting. The fr–lu pair follows a similar trajectory, with a validation loss consistently below the training loss (from 2.35 to 2.05 versus 2.77 to 2.25), the expected behavior when adapter dropout is active during training and disabled during evaluation. In contrast, fr–sw shows a widening gap: the training loss keeps falling to about 1.72 while the validation loss plateaus near 1.84 from the sixth epoch, the onset of overfitting that is consistent with the smallest training set (27,029 pairs, each seen twelve times). The fr–ln validation loss saturates near 2.49 from the seventh epoch.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

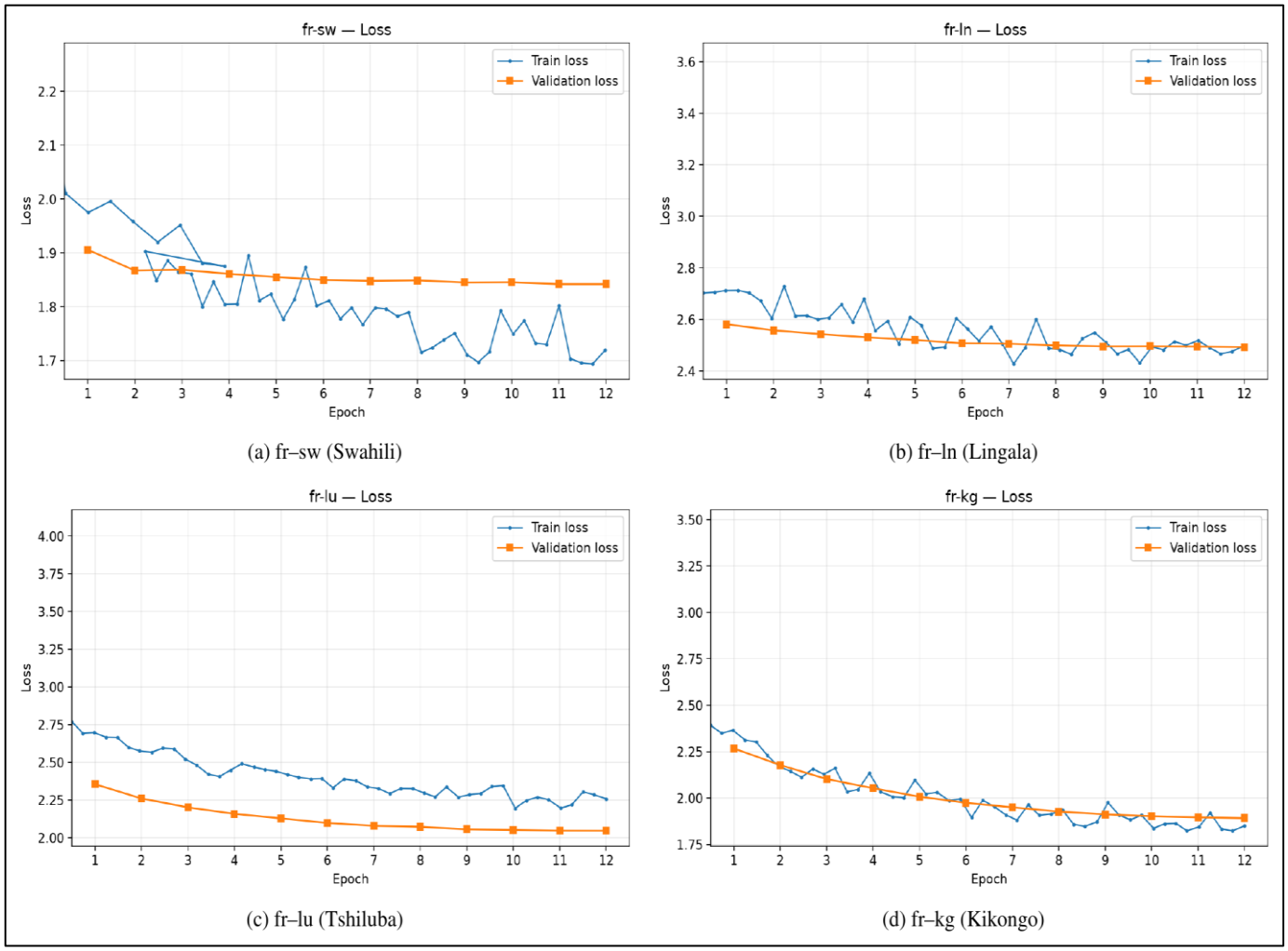


Fig 7 Training and Validation Losses of the Four LoRA Adapters Over the 12-Epoch Budget.

These profiles allow an *a posteriori* assessment of the 12-epoch budget. For Swahili and Lingala it is well calibrated: the former reached the divergence point between training and generalization by the sixth epoch, the latter its plateau by the seventh, and longer training would probably not have helped. For Kikongo and Tshiluba, whose validation losses were still decreasing at the twelfth epoch, a margin of progress remains unexploited, so that their scores are lower bounds of what the configuration permits.

➢ *Translation Quality per Target Language*

Table 7 and Fig. 8 report the scores of the four adapters. Three pairs out of four exceed the threshold $Q_{\min} = 25$: Kikongo (34.08), Lingala (31.50) and Swahili (28.33), with margins of 9.1, 6.5 and 3.3 points. Tshiluba remains below, at 21.49, i.e., 3.5 points short. The mean over the four pairs is 28.85.

Table 7 Translation Quality of NLLB-200-Distilled-600M with Language-Specific LoRA Adapters (French → Target)

Target language	BLEU ↑	chrF ↑	WER (%) ↓	Token acc. (%) ↑
Kikongo (kon_Latn)	**34.08**	50.34	85.21	95.34
Lingala (lin_Latn)	31.50	36.34	95.29	89.95
Swahili (swh_Latn)	28.33	**50.49**	**69.57**	85.39
Tshiluba (lua_Latn)	21.49	38.18	96.94	83.98
Mean	28.85	43.84	86.75	88.67

WER is computed on the translation output against the reference; token accuracy is measured under teacher forcing and is not comparable with free-running metrics.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

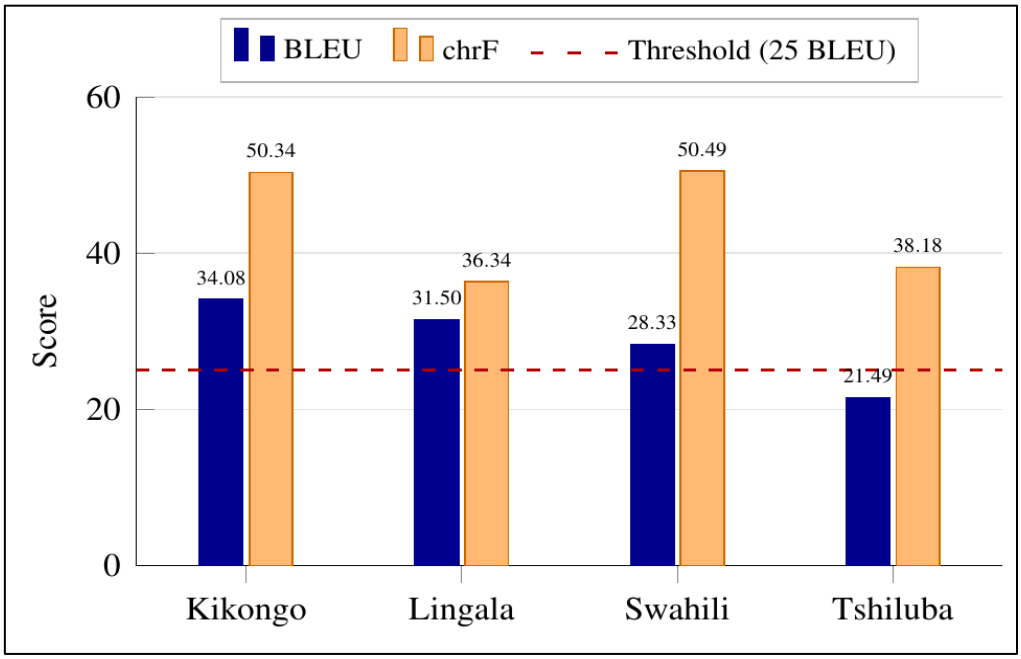


Fig 8 BLEU and chrF per Target Language. The Dashed Line Marks the Threshold $Q_{\min} = 25$ BLEU Fixed Before Experimentation.

- *Metrics Diverge for Lingala*

Lingala obtains the second-best BLEU score (31.50) together with the lowest chrF (36.34) and a WER of 95.29%. Lingala is more analytic than the other three languages—its sentences are on average 7.9% longer than the French sources, against about 10% shorter for Swahili and Tshiluba—and it relies on short, frequent grammatical particles. Matches on such particles raise word n-gram precision, hence BLEU, whereas chrF, computed on character n-grams, rewards the longer content words and morphemes on which errors concentrate. The character-level analysis of Section IV-D is consistent with this reading, which nonetheless calls for human evaluation. Swahili shows the most coherent profile, with the best chrF (50.49) and the lowest WER (69.57%), whereas Tshiluba combines the lowest BLEU and the highest WER (96.94%), which points to a genuine quality deficit rather than to a metric artifact. Had BLEU been the only metric, Lingala would have been counted among the clear successes without revealing its weakness at the sub-word level, which justifies the complementary indicators declared in Section III-J.

- *Token Accuracy Must be Read with Caution*

Values between 83.98% and 95.34% cannot correspond to exact sentence matches, which are rare in machine translation; they measure the proportion of correctly predicted tokens when the decoder receives the reference prefix. This regime differs from free-running generation, in which errors propagate from one token to the next, which explains the apparent contrast with WER. The metric is reported for completeness and is not used to assess the hypotheses.

➢ *Corpus Volume Versus Translation Quality*

Confronted with the training volumes of Table 3, the ranking of the scores shows no relation with the amount of data (Table 8, Fig. 9); over the four pairs, the Spearman rank correlation between training volume and BLEU is exactly zero. The most telling contrast opposes Swahili and Tshiluba: the former, trained on 27,029 pairs, reaches 28.33 BLEU; the latter, trained on 221,263 pairs (8.2 times more), stops at 21.49. No volumetric reading accounts for such a gap.

Table 8 Training Volume, Retention Under Filtering and BLEU, with the Corresponding Ranks

Language	Training pairs	Rank	Retention	BLEU	Rank
Lingala	286,292	1	50.2%	31.50	2
Tshiluba	221,263	2	46.7%	21.49	4
Kikongo	204,223	3	49.5%	34.08	1
Swahili	27,029	4	**94.8%**	28.33	3

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

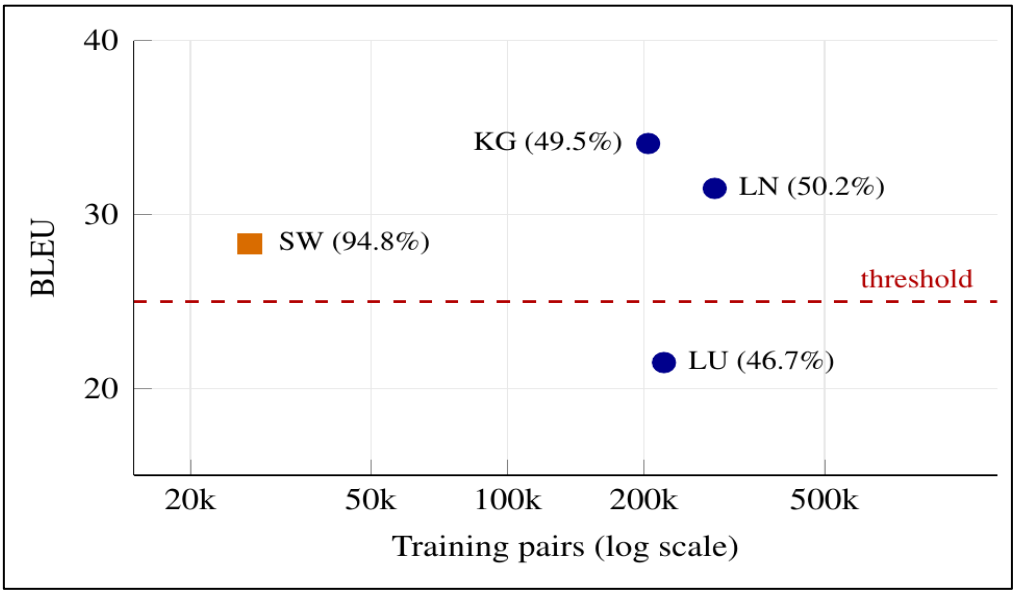


Fig 9 BLEU Versus Training Volume. Circles: Automatically Mined NLLB Bitext; Square: Curated Swahili Sources. RETENTION Under Filtering is Given in Parentheses.

➢ *Character-Level Error Analysis (Lingala)*

Fig. 10 shows the confusion matrix of the fifteen most frequent characters for fr–ln. The diagonal dominates for the vowels *a*, *o* and *i*, but the off-diagonal dispersion is large: the column of unmatched characters alone gathers 279 occurrences for the letter *a*. This dispersion confirms, at the sub-lexical level, the difficulty of restoring Lingala morphology, whose noun-class prefixes and agreement markers carry a substantial share of grammatical information.

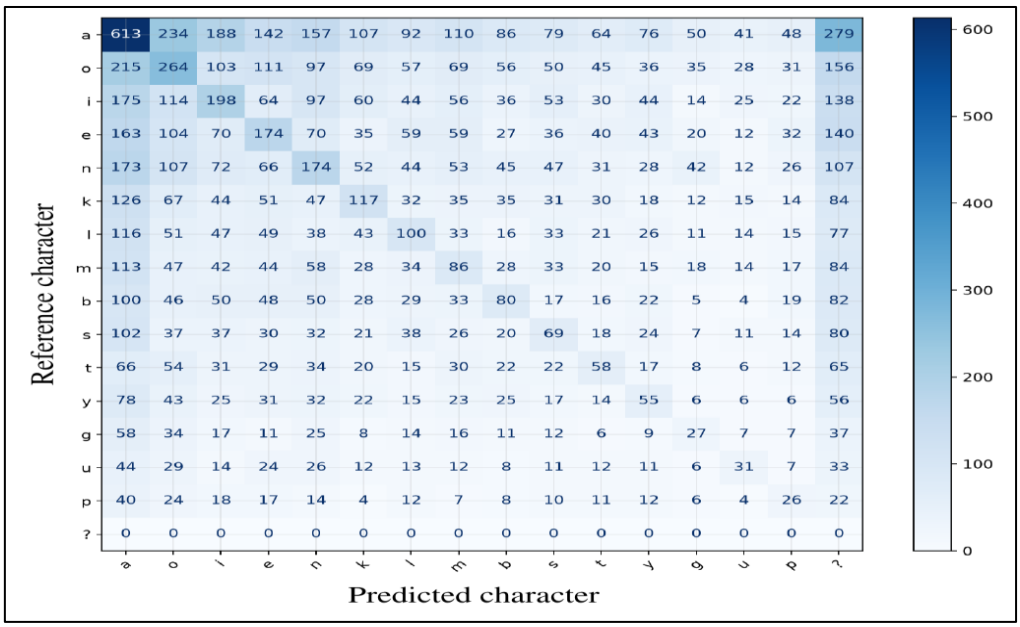


Fig 10 Character Confusion Matrix (15 Most Frequent Characters) between Hypotheses and References for fr–ln.

➢ *Latency on Android Devices*

Measurements come from the internal instrumentation, which recorded 65, 63 and 5 segments on the three devices. The application exposes three input modes with different computational loads: text (translation only), recording (full cascade on a complete utterance) and live (full cascade on the stream, processed in increments of about one second). Table 9 reports the aggregate computed by the application over segments of about 5 s, the only statistic produced automatically.

Table 9 Mean Latency Over Segments of About 5 s, by Device (ms)

Device	L_{audio}	$L_{inf}(\mathcal{M}_{ASR})$	$L_{inf}(\mathcal{M}_{NMT})$	L_{total}	n
TECNO Spark 40 (KM5)	13	1475	9359	10,847	3
Galaxy S22 Ultra	40	851	3235	4126	1

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

The translation stage dominates the decomposition, with 86% of L_{total} on the TECNO Spark 40 and 78% on the Galaxy S22 Ultra. With $n = 3$ and $n = 1$, these values are indicative only and support no statistical inference; moreover, since inference is remote, the difference between the two devices cannot be attributed to local compute.

The live mode is the relevant regime for real-time translation (Table 10, Fig. 11). Three observations stand out. First, the real-time threshold is met *intermittently*: the minimum on the TECNO Spark 40, 1082 ms, lies below $\tau_{\text{rt}} = 2000$ ms, and other segments of the same device come close (1280 ms), but the median remains above the threshold on all three devices. Second, variability is considerable: on the TECNO Spark 50, segments range from 3298 ms to more than 14,000 ms, the latter caused by a spike of the translation stage (13,391 ms) while recognition stayed at 583 ms; together with the fact that the high-end device is not faster than the entry-level ones in this mode, this points to an exogenous cause—load of the remote service or quality of the network link—rather than to a limit of local compute. Third, a preliminary campaign had recorded latencies of about 12,000 ms in text mode on the TECNO Spark 50; the values reported here, three to four times lower, were obtained on an already warm service, and the difference measures the cost of a cold start of the remote container rather than that of inference.

Table 10 Latency in Live Mode, Increments of About 1 s (ms)

Device	n	Minimum	Median	Mean
TECNO Spark 50 (KN4)	8	3298	4414	6700
TECNO Spark 40 (KM5)	8	**1082**	**2637**	**3202**
Galaxy S22 Ultra	3	3142	3956	3741

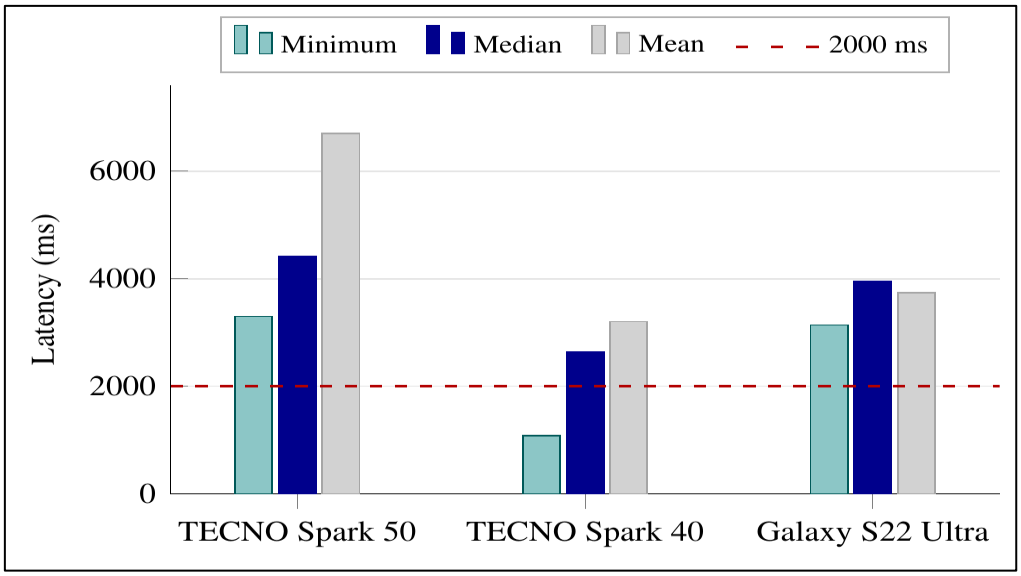


Fig 11 Live-Mode Latency Per Device (Remote-Inference Build). The Dashed Line Marks the Real-Time Budget τ_{rt}.

➢ *Energy Consumption*

Battery drain was recorded over 10 min live-translation sessions (Table 11). Extrapolated to one hour of continuous use, consumption ranges between 3% and 18% of the charge, depending on the device. The Galaxy S22 Ultra is the most frugal, plausibly thanks to a more efficient system-on-chip and radio. The gap between the two TECNO devices, which share the same system-on-chip, rests on a single session for the Spark 40 and may reflect different network conditions, screen brightness or thermal states. The only defensible conclusion at this stage is that energy is not a blocking constraint for conversational use of a few minutes; the number of sessions is too small to establish a reliable energy profile per device.

Table 11 Battery Drain Over a 10 min Live Session

Device	Drain (%)	Per hour (%)	Sessions	Capacity
Galaxy S22 Ultra	0.5	3	2	5000 mAh
TECNO Spark 50 (KN4)	1.0	6	2	5000 mAh
TECNO Spark 40 (KM5)	3.0	18	1	5200 mAh

➢ *Measurements Not Yet Available*

For transparency, we list the measurements foreseen by the protocol that could not be carried out within the study: the peak memory allocation at run time (Android Profiler); the FP32 versus INT8 comparison of WER and BLEU, needed to estimate ϵ; the WER of the speech recognition module on a French test set; the scores of the unadapted NLLB-200 on the same test sets; and the ablation study isolating the respective contributions of LoRA adaptation, INT8 quantization and graph fusion on a single device.

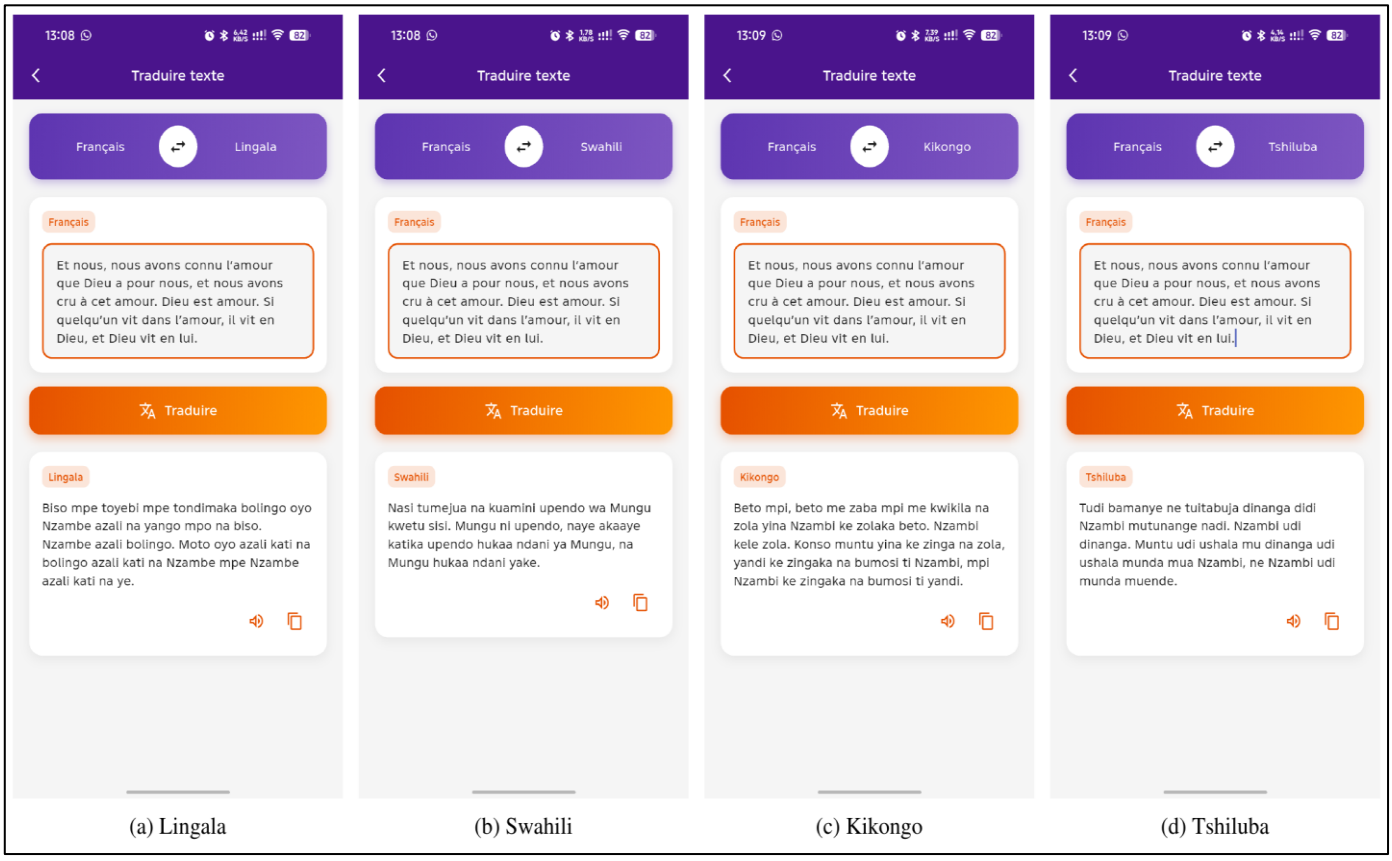
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

V. DISCUSSIONS

➢ *User Interface Demonstrations*

Beyond aggregate scores, the deployed application was inspected qualitatively. The same French source sentence, in a formal register with embedded clauses and taken from religious literature (1 John 4:16), was translated into the four target languages (Fig. 12).

- Transcribed French: “Et nous, nous avons connu l’amour que Dieu a pour nous, et nous avons cru à cet amour. Dieu est amour. Si quelqu’un vit dans l’amour, il vit en Dieu, et Dieu vit en lui.”
- Lingala: “Biso mpe toyebi mpe tondimaka bolingo oyo Nzambe azali na yango mpo na biso. Nzambe azali bolingo. Moto oyo azali kati na bolingo azali kati na Nzambe mpe Nzambe azali kati na ye.”
- Swahili: “Nasi tumejua na kuamini upendo wa Mungu kwetu sisi. Mungu ni upendo, na yeye akaaye katika upendo hukaa ndani ya Mungu, na Mungu hukaa ndani yake.”
- Kikongo: “Beto mpi, beto me zaba mpi me kwikila na zola yina Nzambi ke zolaka beto. Nzambi kele zola. Konso muntu yina ke zinga na zola, yandi ke zingaka na bumosi ti Nzambi, mpi Nzambi ke zingaka na bumosi ti yandi.”
- Tshiluba: “Tudi bamanye ne tuitabuja dinanga didi Nzambi mutunange nadi. Nzambi udi dinanga. Muntu udi ushala mu dinanga udi ushala munda mua Nzambi, ne Nzambi udi munda muende.”



(a) Lingala (b) Swahili (c) Kikongo (d) Tshiluba

Fig 12 The same French Source Sentence Translated into the Four Target Languages in the *Switch Trans* Application.

Three observations emerge. The clause structure of the source is preserved in all four languages: segmentation into propositions, pronominal resumption and the final reciprocal relation are rendered without syntactic collapse. The religious lexicon is transposed with each language’s own term for God—*Nzambe* in Lingala, *Nzambi* in Kikongo and Tshiluba, *Mungu* in Swahili—rather than by transliteration. This sentence, however, belongs to a domain that is heavily represented in the training corpora, and possibly in pre-training, so it shows what the system does in favorable conditions.

A counter-example reveals the limits on out-of-domain input (Fig. 13c). The conversational sentence “je vais bien mon très cher ami développeur mobile lineker”, which contains a technical compound and a proper name, becomes “nazali malamu moninga na ngai ya bolingo mobandisi ya mobile lineker”: the greeting and the affective noun phrase are rendered idiomatically and the proper name is preserved, but the technical compound is only partially transposed, the borrowing *mobile* being copied verbatim without marking. This behavior is expected for vocabulary absent from the training corpora and echoes the domain heterogeneity discussed below. These observations are illustrative and are not a substitute for a systematic human evaluation with native speakers.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

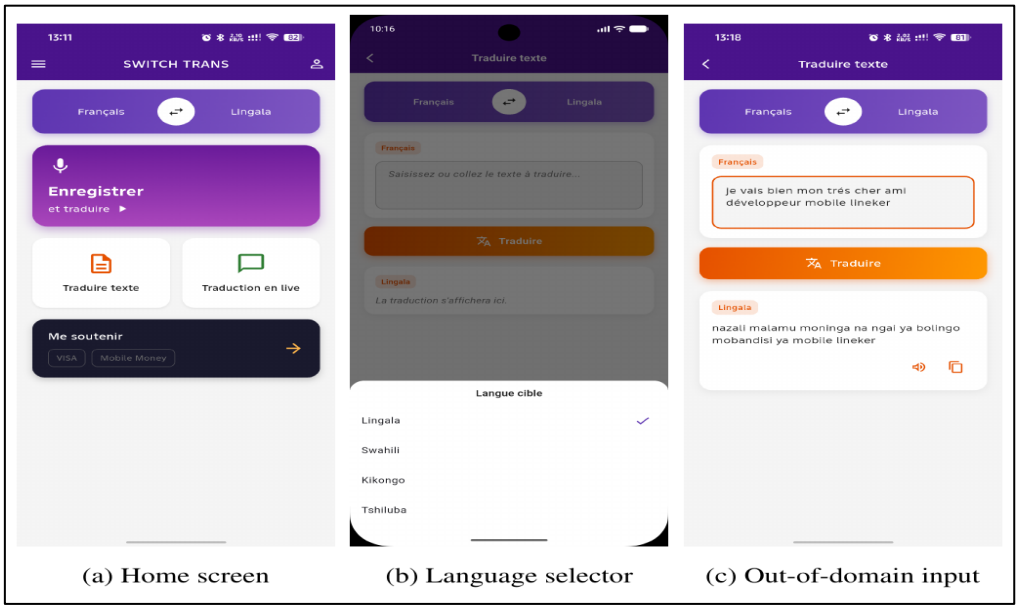


Fig 13 Interface of the *Switch Trans* application: (a) access to the recording, text and live modes; (b) selection of the target language; (c) translation of a conversational, out-of-domain sentence into Lingala.

➢ *Assessment of Hypothesis H1*

H1 posits that LoRA adaptation of a pre-trained multilingual model reaches BLEU ≥ 25 on the DRC national languages. The results support it for three pairs out of four (Kikongo, Lingala, Swahili), and the panel mean (28.85) is also above the threshold; Tshiluba fails at 21.49. Three qualifications delimit the scope of this verdict. First, the training budget was deliberately limited to 12 epochs, and the validation losses of Kikongo and Tshiluba were still decreasing at that point: the only failing pair is precisely one with an unexploited margin, so its 3.5-point shortfall may stem from the budget rather than from the method. Second, without a reference measurement of the unadapted NLLB-200 on the same test sets, the scores cannot be split between the native multilingual capacity of the base model and the contribution of LoRA; H1 concerns the latter, and the data establish a level of performance without attributing its cause. Third, the 80/10/10 split was drawn at the sentence level; for mined bitext that may contain near-duplicates (recurrent religious or administrative formulas), a stricter document-level split such as the book-level split of [2] would rule out any optimistic bias.

➢ *Why Corpus Volume Does Not Predict Quality*

The most substantial finding of this study is not stated by either hypothesis: the ranking of the languages does not follow the ranking of corpus volumes. H1 rested implicitly on the idea that difficulty grows with data scarcity, and this idea is not corroborated. Three factors, which the present data cannot fully separate, account for the observation.

- *Alignment Quality*

The curated Swahili corpus passes the alignment filter at 94.8%, against 46.7–50.2% for the mined bitext. The filter only removes the most obvious failures, so the pairs that survive from an already noisy corpus remain more loosely aligned than those of a curated one. This is consistent with audits showing that mined corpora for low-resource languages contain a large share of defective pairs [8], and with evidence that a few thousand clean, in-domain sentences can matter more than large noisy collections [27].

- *Pre-Training Prior*

Swahili is a medium-resource language within NLLB-200 pre-training, whereas Tshiluba and Kikongo are much less represented; the Swahili adapter starts from a stronger prior.

- *Test-Set Domain*

Each test set inherits the domain of its corpus. The Swahili test set, made of short verified sentences, is intrinsically easier than the Tshiluba test set drawn from comparable documents. Between-language differences are therefore not pure measures of linguistic difficulty.

A reformulated hypothesis—that alignment quality and domain homogeneity, rather than volume, limit adapter quality—is testable by a manual audit of random samples of each corpus, starting with Tshiluba, the only failing pair despite the second largest training set. In the same spirit, the language varieties covered by the data deserve verification: the Swahili sources include Congolese Swahili (swc) while the target code is swh_Latn, and the correspondence between the kon_Latn data and the vehicular Kikongo used as a national language (Kikongo ya Leta) was not examined.

➢ *Assessment of Hypothesis H2*

H2 posits that INT8 quantization keeps the cascade within $M_{\text{lim}} \approx 1\,\text{GB}$ and $\tau_{\text{rt}} = 2000\,\text{ms}$ with a degradation below $\epsilon = 0.05$. It is only partially tested, and the verdict differs by criterion (Table 12). The latency criterion is approached but not guaranteed: in live mode, the minimum falls below the threshold while the medians (2637–4414 ms) do not; the sample sizes are small (three to eight segments per device) and the measured chain delegates inference to a remote service, so it is not the on-device quantized chain that the hypothesis specifies. Energy, not part of the original statement, appears as a non-blocking constraint. The memory criterion

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

remains untested: the static footprint of about 0.87 GB is compatible with M_{lim}, but no run-time peak was recorded, and the footprint concerns a local chain that is not the deployed one. The precision criterion is also untested, the FP32/INT8 comparison not having been carried out. The prevalence of the translation stage in the latency breakdown (78% to 86%) nevertheless designates NMT as the priority target of any further optimization, wherever inference runs.

Table 12 Summary of the Success Criteria Fixed Before Experimentation and of the Corresponding Evidence

Criterion	Threshold	Evidence	Verdict
H1: quality	BLEU ≥ 25	KG 34.08, LN 31.50, SW 28.33, LU 21.49	Met for 3 of 4 pairs
H1: attribution to LoRA	gain over the unadapted model	not measured	Open
H2: latency	≤ 2000 ms	live mode: minimum 1082 ms, medians 2637–4414 ms (remote build)	Met intermittently
H2: memory	≤ 1 GB	static estimate ≈0.87 GB; no run-time peak	Untested
H2: precision	$\epsilon \leq 0.05$	FP32/INT8 comparison not performed	Untested
Complementary: energy	—	0.5–3.0% per 10 min	Non-blocking

➢ *Comparison with Related Systems*

Mandiya et al. [2] report, for French→Lingala, 38.6 BLEU with a BERT-enhanced Transformer on text and 55.4 BLEU for their speech pipeline, on a test set drawn from the Bible with a book-level split. Our 31.50 BLEU for Lingala is measured on a different, mixed-domain test set extracted from mined bitext, with a sentence-level split and a single reference; the two figures are therefore not comparable, and no ranking between the systems can be inferred from them. The two studies are complementary: the former explores bidirectional translation for one language with task-specific models, the present one a single pre-trained model extended to four languages through adapters and packaged for mobile use. A common benchmark—FLORES-200 [26] covers the four target languages—would allow a direct comparison in future work.

VI. CONCLUDING REMARKS AND FUTURE DIRECTIONS

➢ *General Conclusions*

We presented a cascaded system for real-time translation of French speech into Lingala, Swahili, Tshiluba and Kikongo, combining Whisper, NLLB-200-distilled-600M specialized by four language-specific LoRA adapters, a streaming front-end with stable-prefix commitment, an INT8 ONNX compression chain and a Flutter application released on Google Play. Against thresholds fixed before experimentation, the quality target is met for three of the four pairs (Kikongo 34.08, Lingala 31.50, Swahili 28.33 BLEU), with Tshiluba at 21.49 under a training budget constrained by an intermittent power supply. The most informative result is that score rankings are unrelated to corpus volume, while the retention rate under alignment filtering separates curated from mined sources by nearly fifty points: for these languages, the quality of parallel data appears to matter more than its quantity. On the system side, the released build, which delegates inference to a remote endpoint, meets the 2 s budget only intermittently in live mode, at a modest energy cost; the memory and precision criteria of the on-device chain remain to be tested.

➢ *Limitations*

- *Training budget.* Twelve epochs leave an unexploited margin for Kikongo and Tshiluba, whose scores are lower bounds.
- *No baseline nor ablation.* The unadapted NLLB-200 was not evaluated on the same test sets, so the gain attributable to LoRA is not quantified.
- *Evaluation design.* Test sets differ in size (3380 to 35,788 sentences) and domain, rely on a single reference, and were obtained by a sentence-level split. No human evaluation was conducted.
- *Speech recognition.* The WER of Whisper on French, and the effect of recognition errors on translation, were not measured.
- *System measurements.* Latency was measured on a remote-inference build, with few segments per device; peak memory and quantization fidelity were not measured; three handsets were tested, none with a dedicated neural processing unit exploited by the measured path.
- *Linguistic coverage.* Regional varieties (e.g., Lingala of Kinshasa versus Kisangani), code-switching with French and the exact varieties behind the Swahili and Kikongo data were not examined.

➢ *Perspectives*

- Resume training to convergence for Kikongo and Tshiluba once continuous power or remote computing is available.
- Evaluate the unadapted NLLB-200 on the same test sets, run the planned ablation (LoRA, INT8, graph fusion), and report results on FLORES-200 for comparability.
- Audit the alignment of random samples of each corpus, starting with Tshiluba, and test whether filtering by alignment quality, rather than adding volume, improves the adapters; explore back-translation and morphologically informed augmentation.
- Port the full quantized chain to the device, which would enable offline use in areas without coverage, remove mobile data costs and decouple latency from the variability of the network; measure peak memory and FP32/INT8 fidelity, and consider quantization-aware training if the degradation exceeds ϵ.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

- Conduct a larger evaluation with native speakers, dialectal variants and real acoustic conditions (markets, public transport), and revisit the transferability of the 25-BLEU threshold to Bantu languages evaluated with a single reference.

ACKNOWLEDGMENT

The authors thank the members of the Master's thesis committee of the University of Kinshasa for their comments on an earlier version of this work. This research received no external funding. The authors declare no conflicts of interest.

DATA AVAILABILITY

The training, export and evaluation code is available at https://github.com/Christophe4514/msc-ia-switch-2k26, and the source code of the mobile application at https://github.com/Christophe4514/Switch-trans-realtime. The application is distributed on Google Play under the identifier com.switchtrans.sw_trans_realtime. The parallel data derive from publicly available corpora (NLLB bitext, Tatoeba, GlobalVoices, TED, TICO-19, Wikimedia); the cleaned partitions can be regenerated with the provided script and seed 42.

REFERENCES

[1]. République Démocratique du Congo, *Constitution de la République Démocratique du Congo*. Kinshasa, Democratic Republic of the Congo: Journal Officiel de la RDC, Feb. 18, 2006, art. 1.

[2]. R. E. Mandiya et al., "A transformer-based method for bidirectional French–Lingala machine translation in speech and text," *Appl. Sci.*, vol. 16, no. 7, art. no. 3399, 2026, doi: 10.3390/app16073399.

[3]. W. Nekoto et al., "Participatory research for low-resourced machine translation: A case study in African languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, 2020, pp. 2144–2160.

[4]. A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Machine Learning (ICML)*, Honolulu, HI, USA, 2023, vol. 202, pp. 28492–28518.

[5]. NLLB Team et al., "No language left behind: Scaling human-centered machine translation," arXiv:2207.04672, 2022.

[6]. Ž. Agić and I. Vulić, "JW300: A wide-coverage parallel corpus for low-resource languages," in *Proc. 57th Annu. Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 3204–3210.

[7]. Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[8]. J. Kreutzer et al., "Quality at a glance: An audit of web-crawled multilingual datasets," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 50–72, 2022.

[9]. E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. 10th Int. Conf. Learning Representations (ICLR)*, Virtual, 2022.

[10]. B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 2704–2713.

[11]. A. Gholami et al., "A survey of quantization methods for efficient neural network inference," arXiv:2103.13630, 2021.

[12]. A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *Proc. NIPS Workshop on End-to-End Learning for Speech and Audio Processing*, Barcelona, Spain, 2016.

[13]. R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 2625–2629.

[14]. M. Sperber and M. Paulik, "Speech translation and the end-to-end promise: Taking stock of where we are," in *Proc. 58th Annu. Meeting of the Association for Computational Linguistics (ACL)*, Online, 2020, pp. 7409–7421.

[15]. Seamless Communication et al., "SeamlessM4T: Massively multilingual and multimodal machine translation," arXiv:2308.11596, 2023.

[16]. V. Pratap et al., "Scaling speech technology to 1,000+ languages," arXiv:2305.13516, 2023.

[17]. M. Ma et al., "STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework," in *Proc. 57th Annu. Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 3025–3036.

[18]. D. Macháček, R. Dabre, and O. Bojar, "Turning Whisper into real-time transcription system," in *Proc. 13th Int. Joint Conf. Natural Language Processing and 3rd Conf. Asia-Pacific Chapter of the ACL: System Demonstrations*, Bali, Indonesia, 2023.

[19]. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[20]. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[21]. A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[22]. M. Johnson et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 339–351, 2017.

[23]. N. Arivazhagan et al., "Massively multilingual neural machine translation in the wild: Findings and challenges," arXiv:1907.05019, 2019.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov245

[24]. A. Fan et al., "Beyond English-centric multilingual machine translation," *J. Mach. Learn. Res.*, vol. 22, pp. 1–48, 2021.
[25]. M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 597–610, 2019.
[26]. N. Goyal et al., "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 522–538, 2022.
[27]. D. I. Adelani et al., "A few thousand translations go a long way! Leveraging pre-trained models for African news translation," in *Proc. 2022 Conf. North American Chapter of the ACL: Human Language Technologies (NAACL-HLT)*, Seattle, WA, USA, 2022, pp. 3053–3070.
[28]. N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, Long Beach, CA, USA, 2019, vol. 97, pp. 2790–2799.
[29]. A. Bapna and O. Firat, "Simple, scalable adaptation for neural machine translation," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 1538–1548.
[30]. M. Nagel et al., "A white paper on neural network quantization," arXiv:2106.08295, 2021.
[31]. O. Zafrir, G. Boudoukh, P. Izsak, and M. Wasserblat, "Q8BERT: Quantized 8Bit BERT," in *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2-NeurIPS)*, Vancouver, BC, Canada, 2019, pp. 36–39.
[32]. Y. Bondarenko, M. Nagel, and T. Blankevoort, "Understanding and overcoming the challenges of efficient Transformer quantization," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic, 2021, pp. 7947–7969.
[33]. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.
[34]. S. Gandhi, P. von Platen, and A. M. Rush, "Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling," arXiv:2311.00430, 2023.
[35]. ONNX Runtime Developers, "Quantize ONNX models," ONNX Runtime documentation. [Online]. Available: https://onnxruntime.ai/docs/performance/model-optimizations/quantization.html (accessed Oct. 21, 2026).
[36]. M. T. Mukungu, A. M. Mbayandjambe, T. Tashev, K. Kyamakya, and S. K. Kasereka, "BabaSpeech: A deep learning-based translation of sign language into Lingala text and speech for deaf-mute inclusivity," in *Proc. Int. Joint Conf. Artificial Intelligence*, Singapore: Springer, 2025, pp. 91–105.
[37]. A. M. Mbayandjambe, S. K. Kasereka, K. Kyamakya, and V. T. Ho, "Enhancing printed Lingala script recognition using deep learning techniques," *Procedia Comput. Sci.*, vol. 257, pp. 111–118, 2025.
[38]. Silero Team, "Silero VAD: Pre-trained enterprise-grade voice activity detector," GitHub repository, 2021. [Online]. Available: https://github.com/snakers4/silero-vad
[39]. J. Tiedemann, "The Tatoeba translation challenge – Realistic data sets for low resource and multilingual MT," in *Proc. 5th Conf. Machine Translation (WMT)*, Online, 2020, pp. 1174–1182.
[40]. A. Anastasopoulos et al., "TICO-19: The translation initiative for COvid-19," in *Proc. 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, 2020.
[41]. J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proc. 8th Int. Conf. Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012, pp. 2214–2218.
[42]. T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. 2018 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, 2018, pp. 66–71.
[43]. R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016, pp. 1715–1725.
[44]. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.
[45]. S. Migacz, "8-bit inference with TensorRT," presented at the GPU Technology Conf., San Jose, CA, USA, May 2017.
[46]. A. Abid et al., "Gradio: Hassle-free sharing and testing of ML models in the wild," arXiv:1906.02569, 2019.
[47]. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.
[48]. M. Popović, "chrF: Character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal, 2015, pp. 392–395.
[49]. M. Post, "A call for clarity in reporting BLEU scores," in *Proc. 3rd Conf. Machine Translation: Research Papers (WMT)*, Brussels, Belgium, 2018, pp. 186–191.
[50]. T. Kocmi et al., "To ship or not to ship: An extensive evaluation of automatic metrics for machine translation," in *Proc. 6th Conf. Machine Translation (WMT)*, Online, 2021, pp. 478–494.
[51]. T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020, pp. 38–45.
[52]. S. Mangrulkar et al., "PEFT: State-of-the-art parameter-efficient fine-tuning methods," GitHub repository, 2022. [Online]. Available: https://github.com/huggingface/peft.