ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

KOBONGOLA-Lite: Design and Prospective Evaluation Protocol for a Hybrid Offline French–Lingala Neural Translator on Resource-Constrained Smartphones

Kinkete Mfumabi Hervé[1]; Ngoy Déogracias[2]; Ngoma Magema Arsene[2]; Yamone Muisangie Clairette[2]; Kapinga Nsapo Marisa[2]; Mpoyi Kabulu Exauce[2]; Likango Bokondo Christopher[2]; Kapongo Musangu Israël[2]

[1]Faculty Academic Secretary, Faculty of Computer Science and Artificial Intelligence, William Booth University, Kinshasa, Democratic Republic of the Congo
[2]Students, Faculty of Computer Science and Artificial Intelligence, William Booth University, Kinshasa, Democratic Republic of the Congo

Publication Date: 2026/08/31

Abstract—French and Lingala coexist in institutional, educational, and everyday communication in the Democratic Republic of the Congo, where connectivity is uneven and entry-level Android devices remain widespread. This paper specifies KOBONGOLA-Lite, a proposed bidirectional offline French–Lingala neural translation architecture, and defines a prospective protocol for its implementation and confirmatory evaluation. The contribution is deliberately narrower than claims of first bidirectionality or first compression, because recent French–Lingala systems and AfriNLLB already cover these dimensions. KOBONGOLA-Lite instead coordinates four testable components: a governed and stratified Congolese corpus; explicit treatment of natural French–Lingala code-switching; a context-gated lexical mechanism with neural fallback; and physical-device validation of linguistic quality, memory, latency, energy, thermal behavior, reliability, and network silence. Interface prototypes specify the intended user workflows, while the compact encoder–decoder model, tokenizer, domain lexicon, span policies, hybrid reranker, and self-contained mobile bundle remain objects of prospective implementation and validation. The planned study separates external baselines from controlled internal ablations that isolate sequence-level distillation, lexical support, hybrid reranking, ONNX export, and dynamic INT8 quantization. The primary automatic endpoint is chrF++; complementary evidence comprises SacreBLEU, locally validated AfriCOMET, terminology and entity preservation, omission analysis, blinded human assessment, paired document-clustered bootstrap intervals, effect sizes, and an explicitly defined ten-test Holm family. No implementation or confirmatory performance result is claimed at this stage. The protocol establishes a falsifiable and reproducible basis for assessing offline neural machine translation in a low-resource African and edge-computing context.

Keywords: *Code-Switching; Edge AI; French–Lingala Translation; INT8 Quantization; Knowledge Distillation; Low-Resource Languages; Neural Machine Translation; Offline Translation.*

How to Cite: Kinkete Mfumabi Hervé; Ngoy Déogracias; Ngoma Magema Arsene; Yamone Muisangie Clairette; Kapinga Nsapo Marisa; Mpoyi Kabulu Exauce; Likango Bokondo Christopher; Kapongo Musangu Israël (2026) KOBONGOLA-Lite: Design and Prospective Evaluation Protocol for a Hybrid Offline French–Lingala Neural Translator on Resource-Constrained Smartphones. *International Journal of Innovative Science and Research Technology*, 11(8), 2085-2097. https://doi.org/10.38124/ijisrt/26aug847

I. INTRODUCTION

Attention-based architectures and Transformers have substantially improved neural machine translation (NMT) [1], [2]. Their benefits, however, remain unevenly distributed. High-resource languages are supported by large parallel corpora, mature benchmarks, specialized models, and abundant inference infrastructure; low-resource languages continue to face limited data, orthographic variation, uneven dialectal coverage, and a shortage of qualified evaluators. Multilingual pretrained models partly reduce this asymmetry, but their size, computational cost, and direction-dependent quality make local validation indispensable [3].

Lingala is a major Bantu language spoken in the Democratic Republic of the Congo (DRC) and the Republic

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

of the Congo [23]. In Kinshasa and other urban settings, it frequently co-occurs with French within the same interaction, sentence, or syntactic phrase. Such code-switching is not merely noise: it may signal domain, register, social identity, or differential access to technical vocabulary [24]. A system that mechanically normalizes every mixed utterance into a single language may therefore erase part of the communicative intent. A server-dependent solution also remains vulnerable to outages, mobile-data costs, latency, and the transmission of potentially sensitive text.

The scientific problem is consequently not limited to producing French–Lingala translations. It is to determine whether a compact system can preserve useful linguistic quality, handle rare terminology and mixed utterances, operate without network access, and remain within realistic storage, memory, energy, and latency budgets on entry-level smartphones. KOBONGOLA-Lite addresses this problem through a specified architecture, interface prototypes, and a prospective confirmatory protocol; no executable model-integrated APK or validated translation system is claimed in this manuscript.

➢ *Positioning after Recent French–Lingala Work*

Three direct developments constrain any novelty claim. LiSTra documented English-to-Lingala speech translation and a data-construction pipeline [4]. Mandiya et al. studied bidirectional French–Lingala translation in text and speech [5]. AfriNLLB subsequently released efficient French↔Lingala models derived from NLLB-200 through pruning, fine-tuning, quantization, and knowledge distillation, including CTranslate2 variants [6]. KOBONGOLA-Lite therefore does not claim to be the first bidirectional or compressed system for this pair. Its testable contribution lies in the coordinated study of governed Congolese data, natural code-switching, context-sensitive lexical control, explicit neural fallback, and measured on-device behavior.

➢ *Research Question and Contributions*

The main question is: to what extent can a compact architecture that combines sequence-level distillation, contextual lexical bias, span-processing policies, hybrid reranking, and INT8 quantization achieve a verifiable trade-off among French–Lingala quality, robustness to code-switching, and offline inference cost on constrained smartphones?

- A specified Android-oriented architecture in which the model, tokenizer, lexicon, span rules, and decoding logic are intended to be embedded locally.
- A provenance-aware corpus plan with group-wise anti-leakage partitioning before reversal or augmentation.
- A computable hybrid mechanism that applies lexical evidence only above a frozen confidence threshold and otherwise falls back to neural decoding.
- Ablations, non-inferiority margins, mobile acceptance criteria, and statistical decision rules fixed before final-test access.
- A reproducibility package comprising configuration manifests, hashes, a dataset datasheet, a model card, and a deviation log.

II. RELATED WORK

➢ *Low-Resource Multilingual Translation*

Subword tokenization, including byte-pair encoding and SentencePiece, mitigates open-vocabulary problems and is particularly useful for morphologically rich languages [7], [8]. Multilingual pretraining, targeted fine-tuning, back-translation, and distillation can transfer representations from better-resourced languages, but transfer is neither uniform nor automatically beneficial. Adelani et al. showed that a few thousand high-quality human translations can yield substantial improvements for African-language news translation, reinforcing the importance of data provenance and quality rather than unaudited volume [9]. NLLB-200 expanded multilingual coverage and introduced FLORES-200 as a shared benchmark [3]. For the present study, FLORES-200 is an external generalization test, not a substitute for Congolese conversational, domain, and code-switching data.

➢ *Hybrid Lexical Support and Code-Switching*

Code-switched MT cannot be evaluated as ordinary monolingual translation: mixed inputs combine matrix-language structure, embedded spans, borrowings, and discourse-sensitive preservation decisions, and automatic metrics alone may obscure errors visible to bilingual judges [25]. Lexically constrained decoding can guarantee required words or phrases through grid beam search [26] or dynamic beam allocation [27], while terminology-aware training can teach a model to apply constraints without changing the base architecture [28]. These methods establish useful controls but do not by themselves decide whether a candidate term is contextually appropriate in natural French–Lingala switching.

Neural models may produce fluent sentences while omitting an item, substituting a rare term, or altering a number. Conversely, a hard lexical constraint may preserve a term at the cost of grammaticality or contextual accuracy. KOBONGOLA-Lite therefore uses a softer, auditable strategy: a lexicon proposes candidates; a lightweight contextual gate estimates relevance; and the decoder applies a logit bias or reranking feature only above a frozen threshold. Borrowings, proper names, acronyms, and technical expressions are distinguished from genuine language alternation because their output policies differ: translate, preserve, transliterate, normalize, or defer.

➢ *Compression and African-Language Evaluation*

Knowledge distillation transfers teacher behavior to a smaller student model [10], [11], whereas quantization reduces numerical precision. These operations affect linguistic and hardware performance differently and must be isolated experimentally. MobileNMT illustrates the importance of jointly reporting size, memory, and speed, although its 15 MB and 30 ms results are protocol-specific and cannot be treated as universal thresholds [12]. BLEU

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

remains useful for comparability, but chrF++ is more tolerant of morphological variation [13], [14]. AfriCOMET improves learned evaluation for African languages [15], yet its validity for French–Lingala must be checked locally against human judgments before confirmatory use.

Table 1 Conservative Positioning against the Closest Work

Work	Key prior capability or proposed focus	Role in this study
LiSTra [4]	English→Lingala speech and data pipeline	Lingala resource reference
Mandiya et al. [5]	French↔Lingala text and speech	Direct scope comparator
AfriNLLB [6]	French↔Lingala; pruning, distillation, quantization, CTranslate2	Mandatory baseline and teacher
MobileNMT [12]	On-device NMT; 15 MB / 30 ms in its protocol	Mobile-method reference
KOBONGOLA-Lite	Prospective Congolese corpus, natural switching, hybrid control, ONNX/INT8 device protocol	Hypotheses under test

III. RESEARCH DESIGN AND DECISION RULES

➢ *Study Type and Preregistration*

The study follows a comparative experimental design with controlled ablations. Before final-test access, the author will freeze corpus versions, exclusions, test sets, the primary metric, hypotheses, non-inferiority margins, random seeds, the hyperparameter budget, system configurations, analyses, and stopping rules. This manuscript is a prospective protocol, not an already registered study. The protocol will be time-stamped in a durable repository before confirmatory training or final-test access; the repository identifier and URL will then be added to subsequent versions. Any departure will be recorded and labeled confirmatory or exploratory. The primary comparison unit is the source–reference pair clustered by document or contributor; reversed variants of the same pair remain in the same provenance group. French→Lingala and Lingala→French are always reported separately.

➢ *Confirmatory Hypotheses*

Table 2 Confirmatory Hypotheses and Failure Interpretations

ID	Directional decision rule	Negative interpretation
H1	M2−M0: lower 95% CI bound for ΔchrF++ > 0	Hybrid quality gain unsupported
H2	M1−M0 intersection: lower bound for Δ preservation macro-accuracy > 0 and for Δ human adequacy > −0.15 points on the 1–5 scale	Lexical gate unsupported
H3	M2−M1: upper bound for Δ critical-error rate < 0; critical error is the segment-level event defined in Section V.D	Span-policy/reranking gain unsupported
H4	M3−M2-ORT: lower bound for ΔchrF++ > −1.0 plus every mobile gate	M3 not qualified for mobile deployment
H5	M2−M0: lower bound for Δ blinded human adequacy > 0; gatekeeping p-value includes H1	Human adequacy gain unsupported

Exactly ten directional composite p-values enter one Holm family: H1-FR→LN, H1-LN→FR, ..., H5-FR→LN, and H5-LN→FR. Their construction is frozen in Section VI.E. H2 uses the maximum of its preservation-superiority and adequacy-non-inferiority component p-values. H5 uses the maximum of the corresponding H1 linguistic p-value and its human-superiority component p-value, which implements gatekeeping without a data-dependent change to family size. H4 is assigned p=1.00 whenever any mandatory mobile gate fails. Holm's step-down procedure controls family-wise error at α=0.05 across these ten values; an undefined or untestable comparison is assigned p=1.00.

➢ *A Priori Sample-Size Feasibility*

The local general test initially contains at least 1,200 source–reference pairs per direction. The natural code-switching test contains at least 600 independently identified utterances, and the terminology/entity/number test contains at least 600 enriched pairs. Before training, 10,000 cluster-aware bootstrap simulations will evaluate the initial design under conservative assumptions: a 15-point segment-level chrF++ standard deviation, paired correlation 0.60, an average of five segments per document, and a minimum effect of interest of 1.5 chrF++ points. The design qualifies only if simulated power is at least 90% and the median 95% confidence-interval width for the difference does not exceed 2.0 points. Otherwise, the general test is increased once to 1,800 pairs per direction before training. No resizing is permitted after system outputs on the final test become visible.

Human-endpoint power is evaluated separately by 10,000 simulations of the frozen rating design: 100 document clusters per direction with four sampled segments each, three ratings per output, residual standard deviation 0.90 point, evaluator intraclass correlation 0.10, document intraclass correlation 0.05, and paired system correlation 0.50. H2 is powered for true adequacy difference 0 against the −0.15-point non-inferiority margin; H5 is powered for a +0.20-point mean gain. Each directional test must reach at least 90% simulated power at one-sided α=0.05. If either fails, the human sample is increased once, before output generation, from 400 to 600 segments per direction with the same strata and three ratings; no post-inspection resizing is allowed.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

IV. DATA, ANNOTATION AND GOVERNANCE

➢ *Linguistic Population and Admissible Sources*

The corpus targets contemporary French used in the DRC and documented written Lingala varieties, including standard or literary Lingala, urban Kinshasa usage, and regional varieties when competent speakers can identify them. Dialect labels derive from metadata and contributor self-identification; no variety is treated as intrinsically superior. Admissible data include public-domain or compatibly licensed documents, authorized institutional material, open educational resources, consented community contributions, commissioned human translations, and approved conversational subsets. Private messages, identifiable medical or school records, and content of unknown license are excluded.

Each unit receives a persistent identifier, provenance record, license status, domain, date, creation method, validation state, and dissemination restriction. Raw and normalized versions remain separate. Identifying material is removed before annotation unless explicit consent and a documented justification apply.

Table 3 Planned Corpus Composition

Planned stratum	Target share	Key requirement
Institutional / administrative	15–25%	Explicit reuse rights; public terminology
Education	15–20%	Multiple levels and subjects
Non-clinical public health	10–15%	Terminology review; use warning
Authorized information / media	10–15%	Diversity of styles and dates
Natural conversation	20–30%	Consent, anonymization, authentic switching
Religious / community	≤15%	Cap to limit domain bias
Synthetic / augmented	Separate	Never merged with human-data counts

➢ *Cleaning, Partitioning, and Leakage Prevention*

The transformation chain is retained for every segment. Processing includes language identification, encoding checks, reversible normalization of spacing and punctuation, abnormal length-ratio detection, exact and near-duplicate detection, machine-translation provenance checks, preservation checks for names, numbers, dates and units, and human review of uncertain cases. Partitioning occurs by document, source, author, or speaker before pair reversal, back-translation, paraphrase, or any other augmentation. All derivatives inherit the parent source group. Frozen tests are checked for contamination against FLORES-200 and other public benchmarks; hashes may be released when test text cannot be shared.

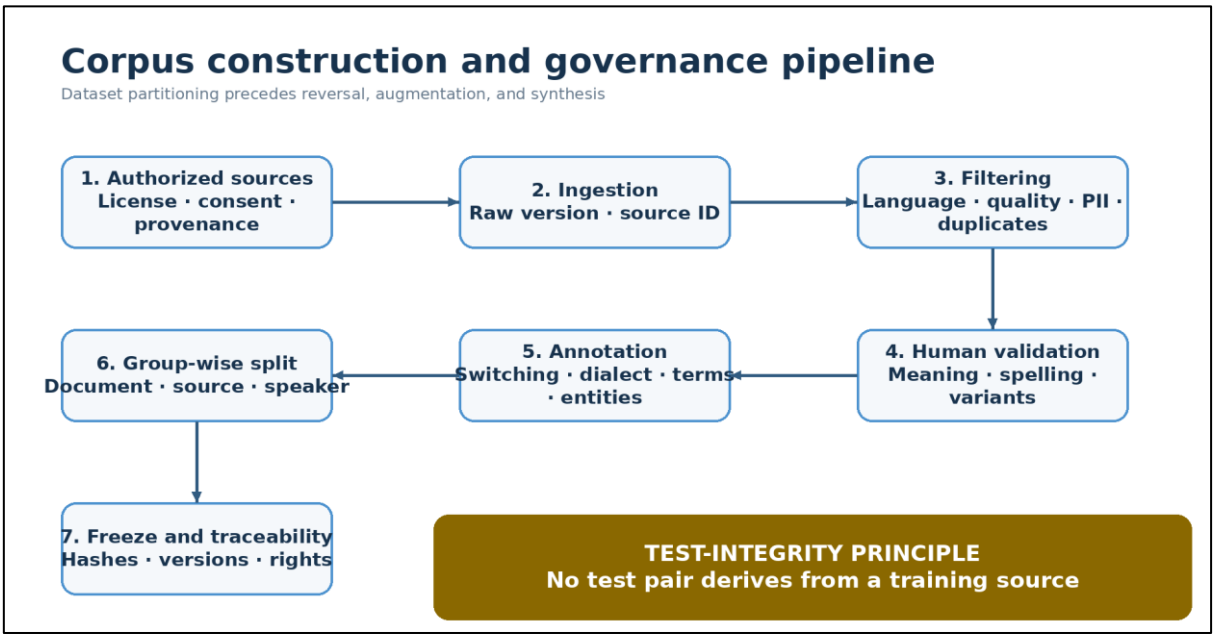

Fig 1 Corpus Construction, Annotation, Partitioning and Freeze Pipeline.

➢ *Linguistic Annotation*

Annotation covers semantic equivalence, domain, variety, register, orthography, code-switching, entities, numbers, domain terms, and ambiguity. Mixed utterances are annotated at segment and span levels as French-dominant with Lingala insertion, Lingala-dominant with French insertion, balanced alternation, lexicalized borrowing, proper name, acronym, technical term, or indeterminate. Each span receives an intended output policy: translate, preserve, transliterate, normalize, or leave undecided. At least 25% of examples are independently double-annotated. Krippendorff's α is reported for categories and span F1 for boundaries. An α below 0.80 triggers manual revision and recalibration rather than silent removal of disagreements.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

The admissible reference is frozen span by span for the requested target language. A source-language matrix span is translated; an embedded span already in the target language is preserved, with only preregistered orthographic normalization. In balanced alternation, these two rules are applied independently to every annotated span and the reference must preserve the full proposition. A lexicalized borrowing is preserved only when its borrowing status appears in the frozen target-language list; otherwise it is translated. Proper names and acronyms are preserved, except for a prelisted conventional target form. Numbers, dates, and units preserve value and unit, permitting only frozen locale formatting. A technical term takes the validated domain-specific target entry when one exists and otherwise receives an unconstrained human translation. Indeterminate items require two independently accepted reference variants or are excluded from confirmatory code-switch analyses before partition freeze. Thus every test span has one policy, one admissible form set, and an explicit critical/noncritical status before system output is generated.

Table 4 Frozen Data Partitions

Set	Construction	Primary use
Training	Authorized provenance groups after exclusion of all development/test groups	Parameter optimization
General development	Separate groups stratified by direction, domain, length, and variety	Model selection and early stopping
Code-switch development	Groups distinct from the mixed test	Lexical threshold and reranking weights
Local general test	≥1,200 pairs/direction; a priori rule may raise to 1,800	Primary chrF++ and confirmatory comparisons
Natural code-switch test	≥600 natural utterances	H3 and span-policy errors
Terminology/entity/number test	≥600 enriched pairs	H2 and preservation metrics
FLORES-200	fra_Latn and lin_Latn after contamination audit	External generalization

V. PROPOSED KOBONGOLA-LITE ARCHITECTURE

➢ *Functional Overview*

The proposed system accepts French, Lingala, or mixed text and an explicit target direction. Reversible preprocessing is specified to detect language, entities, numbers, and candidate spans. A shared tokenizer will segment the input, and a compact student model will generate neural hypotheses. The lexical module is designed to retrieve candidates compatible with direction, domain, and local context. A confidence gate will determine whether lexical evidence should influence decoding. Hybrid reranking will combine neural likelihood, terminology coverage, span policies, and omission penalties. Detokenization and all supporting operations are intended to execute locally.

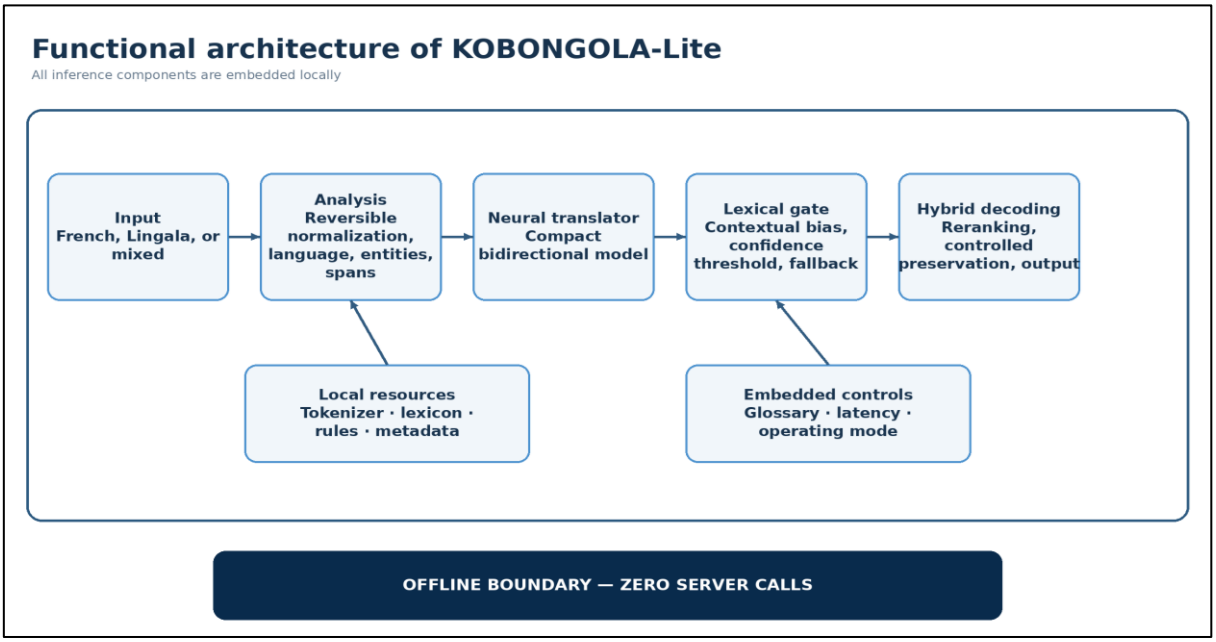

Fig 2 Functional Architecture and Offline Inference Boundary.

➢ *Compact Student and Sequence-Level Distillation*

The confirmatory teacher is the frozen French↔Lingala AfriNLLB Transformers checkpoint used as external baseline B1. Its repository revision and SHA-256 hash are recorded before synthetic data generation. The confirmatory student is a pre-norm bidirectional encoder–decoder Transformer with six encoder layers, four decoder layers, model dimension 384, feed-forward dimension 1,536, six attention heads (64

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

dimensions per head), GELU activation, learned positional embeddings, maximum source and target length 256 subword tokens, residual/embedding dropout 0.20, attention dropout 0.10, and layer-normalization $\varepsilon=10^{-5}$. Encoder input, decoder input, and output projection weights are tied. Explicit target-language tokens are user-defined symbols in a joint 16,000-unit SentencePiece unigram vocabulary trained only on the authorized training partition with NFKC normalization, character coverage 1.0, and byte fallback enabled. Examples exceeding 256 tokens after reversible sentence segmentation are excluded from training and reported; final-test segments are never truncated. The realized parameter count is computed from the frozen graph and recorded in the manifest.

Because teacher and student use different vocabularies, confirmatory distillation is performed at sequence level rather than by direct logit matching. For each authorized training source x_i, the frozen teacher T produces a deterministic synthetic target under fixed decoding configuration γT:

$$\hat{y}KD(i) = Decode(T, x_i; \gamma T). \quad (1)$$

The teacher decoding configuration γT is fixed as follows: source tag fra_Latn or lin_Latn; forced target beginning-of-sentence tag lin_Latn or fra_Latn, respectively; beam size 5; length penalty 1.0; maximum 256 newly generated tokens; deterministic decoding (do_sample=False); early stopping enabled; one returned sequence; and no no-repeat n-gram constraint. The exact teacher checkpoint revision, tokenizer revision, framework version, and SHA-256 hashes are frozen before synthetic generation.

Teacher outputs are stored as text, audited for prespecified technical anomalies, and retokenized with the student's SentencePiece model. Development and test sources are never passed through the teacher and reintroduced into training. D0 uses only validated human references. M0 retains the same architecture, tokenizer, partitions, seeds, optimizer, and stopping rules while adding teacher sequences under the fixed objective.

$$LM0(\theta) = 0.5\ Lhum(\theta) + 0.5\ LKD(\theta). \quad (2)$$

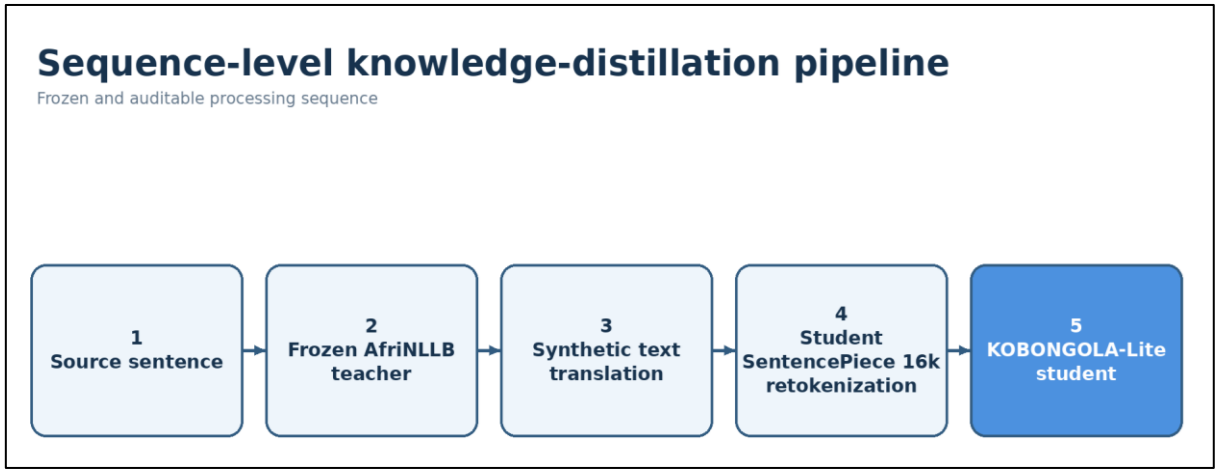

Fig 3 Sequence-Level Knowledge-Distillation Chain.

➢ *Context-Gated Lexical Support*

For a detected source span s and lexical candidate c in context x and domain d, the gate uses exactly five features scaled to [0,1]: (1) freq = $\log(1+n(s,c))/\log(1+\max_{c'} n(s,c'))$, where n is the validated training count; (2) sim = $\cos(v(x),v(g(c)))$, clipped to [0,1], where v is an L2-normalized sparse TF–IDF vector over Unicode character 3–5-grams and g(c) concatenates the candidate gloss with its validated example context; (3) domain = 1 for an exact domain match, 0.5 for a general-domain candidate, and 0 otherwise; (4) policy = 1 when the candidate action equals the frozen span policy and 0 otherwise; and (5) validation = 1 for expert-validated, 0.5 for independently double-validated community entries, and 0 for all other entries. The TF–IDF vocabulary and inverse-document frequencies are learned only from authorized training lexicon contexts, with sublinear term frequency, minimum document frequency 2, a 20,000-feature cap, deterministic frequency-then-lexicographic pruning, and no neural sentence encoder at runtime. With frozen weights w and sigmoid σ:

$$g(c|s,x,d) = \sigma(w^T[freq, sim, domain, policy, validation] + b) \quad (3)$$

Before gate fitting, sim is validated against majority binary relevance judgments on the dedicated tuning panel described in Section VI.D. It is retained only if cross-validated AUROC ≥ 0.75 and expected calibration error ≤ 0.10 using ten equal-frequency bins; otherwise its coefficient is fixed to zero and the gate is refitted. The lexicon influences decoding only if the best candidate score is at least threshold τ. Below τ, or if no candidate meets all direction and policy filters, the system falls back to the unmodified neural path.

➢ *Hybrid Generation and Reranking*

M1 adds only context-gated lexical evidence to M0. M2 adds span-policy features and reranking while retaining the same student checkpoint. For each beam candidate y, the frozen reranking score is:

$$S(y|x) = \lambda N \log P\theta(y|x) + \lambda L\ Clex(y) + \lambda S\ Cspan(y) - \lambda O\ O(y) - \lambda W\ W(y) \quad (4)$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

Let J(x) be eligible lexical items with frozen target-form sets A_j and gate confidences q_j, S(x) the annotated policy spans, and P(x) the protected entities, numbers, dates, units, negations, and critical domain terms. Then $Clex(y)=\Sigma_j q_j I[y \text{ contains a normalized form in } A_j]/\Sigma_j q_j$ (0 if J is empty); $Cspan(y)=|S|^{-1}\Sigma_s I[y \text{ satisfies the frozen action and admissible-form set for } s]$ (0 if S is empty); $O(y)=|P|^{-1}\Sigma_p I[p \text{ is absent or value-corrupted without licensed deletion}]$ (0 if P is empty); and W(y) is the proportion of lexical output tokens outside preserve spans labeled as non-target language by a frozen character 3–5-gram logistic classifier. That classifier is trained only on authorized monolingual training text and must reach macro-F1 ≥0.90 on a separately annotated development sample; otherwise W is removed from reranking and retained only as a blinded human error label. A segment has a critical error when O(y)>0, W(y)>0, or any protected entity/number/date/unit is corrupted. The critical-error rate is $N^{-1}\Sigma_i I[\text{segment } i \text{ has at least one such error}]$; overlapping errors count once per segment.

Weight values and tie-breaking rules are frozen before final-test access. D0→M0 isolates sequence-level distillation; M0→M1 isolates lexical support; M1→M2 isolates span control and reranking; M2→M2-ORT audits export fidelity; and M2-ORT→M3 isolates INT8 quantization.

Reranker weights are selected deterministically on the code-switch development set. Candidate scores are first normalized per source sentence to [0,1]. λN is fixed at 1.0; an exhaustive grid evaluates $\lambda L,\lambda S \in \{0,0.25,0.5,1.0\}$ and $\lambda O,\lambda W \in \{0,0.5,1.0,2.0\}$, for 256 combinations, using five-fold provenance-grouped cross-validation. The selected tuple minimizes the combined critical-error rate subject to mean chrF++ ≥ chrF++(M1)−0.20 and mean human adequacy ≥ adequacy(M1)−0.15 points on the 1–5 scale. Ties are resolved by higher chrF++, then smaller L1 norm $\lambda L+\lambda S+\lambda O+\lambda W$, and finally lexicographic order $(\lambda L,\lambda S,\lambda O,\lambda W)$. Fold assignments, normalized feature values, all grid results, and the selected tuple are frozen before final-test access.

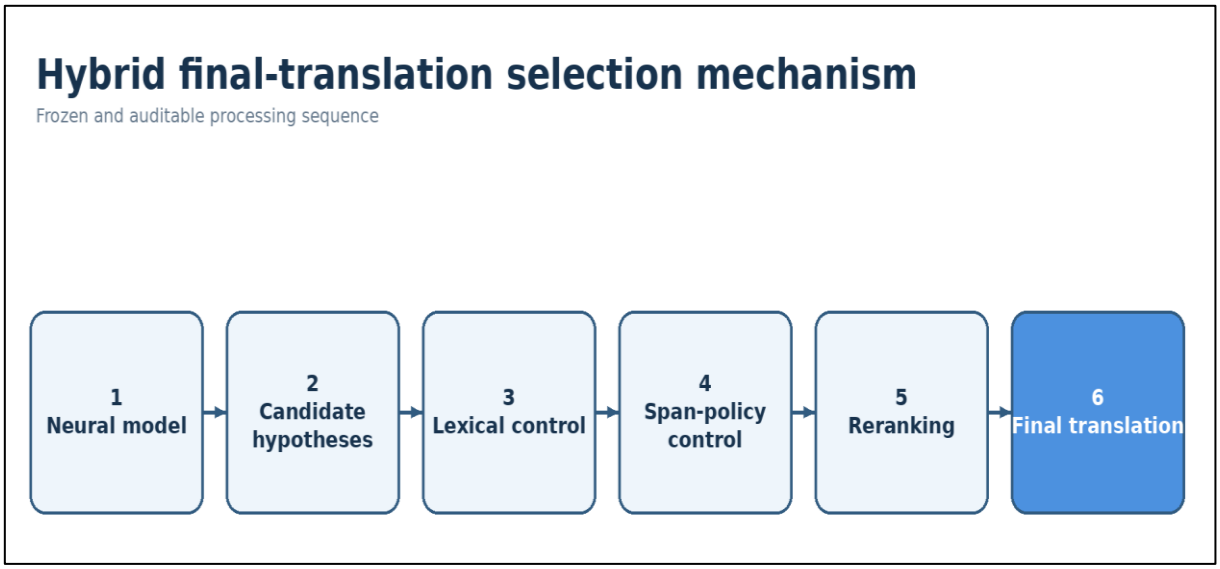

Fig 4 Hybrid Candidate-Generation and Final-Selection Mechanism.

➢ *ONNX Export and INT8 Quantization*

M2 is exported to ONNX as M2-ORT and evaluated on 1,000 frozen general-development sentences per direction. Export fidelity receives PASS only if all 2,000 sentences execute successfully, at least 99.5% of detokenized outputs are exactly identical to M2, the absolute corpus-level chrF++ difference is ≤0.10 point in each direction, and no protected entity, number, date, or unit differs. Failure of any condition is FAIL and blocks confirmatory quantization. Only after PASS is dynamic post-training INT8 quantization applied to produce M3. The manifest records source and target hashes, ONNX Runtime version, opset, quantization API and arguments, data types, included and excluded operators, file sizes, CPU settings, and reproducible commands. If M3 violates linguistic non-inferiority, its hardware gains cannot qualify it as the confirmatory mobile model. A later static-quantized or quantization-aware variant receives a new identifier and remains exploratory or requires a new preregistration.

➢ *Interface Prototypes*

The interface prototypes specify the intended principal translation flow and the explanatory views for inspecting mixed spans and lexical decisions. They are design artifacts only: they do not establish that a model-integrated APK currently executes offline translation, nor do they provide evidence that any linguistic or mobile qualification threshold has been met.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

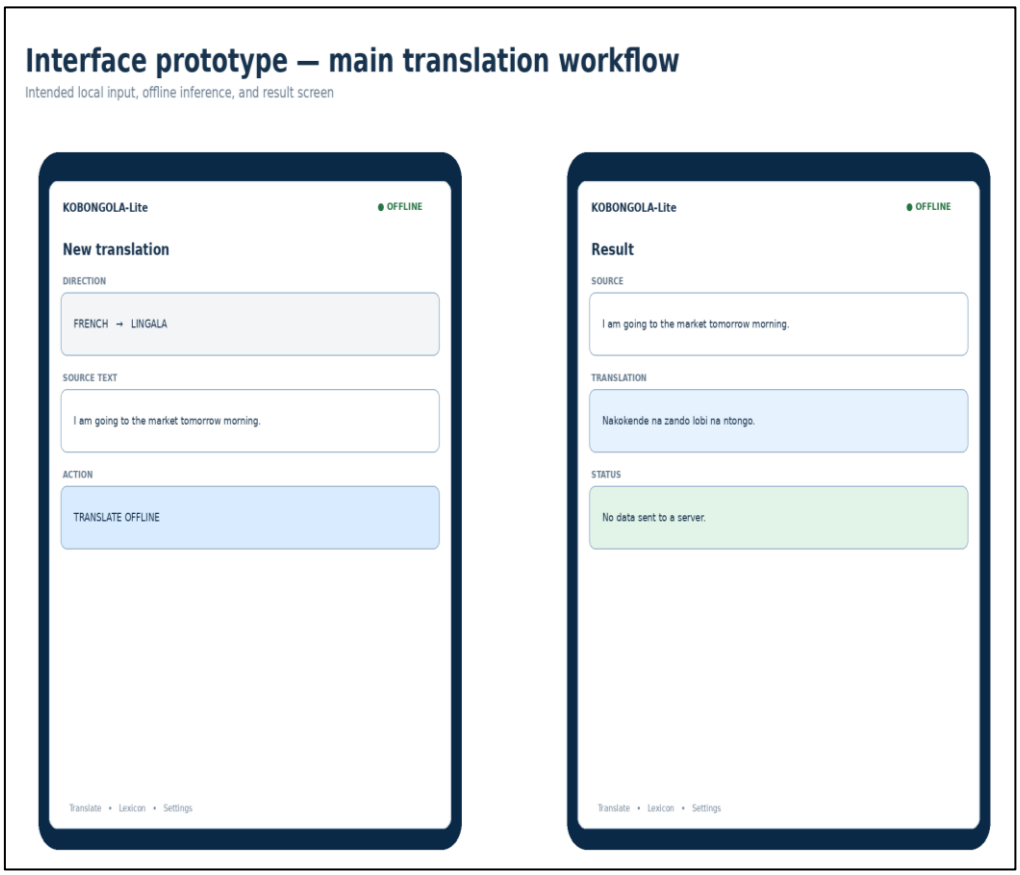

Fig 5 Interface Prototype for the Intended Offline Translation Workflow.

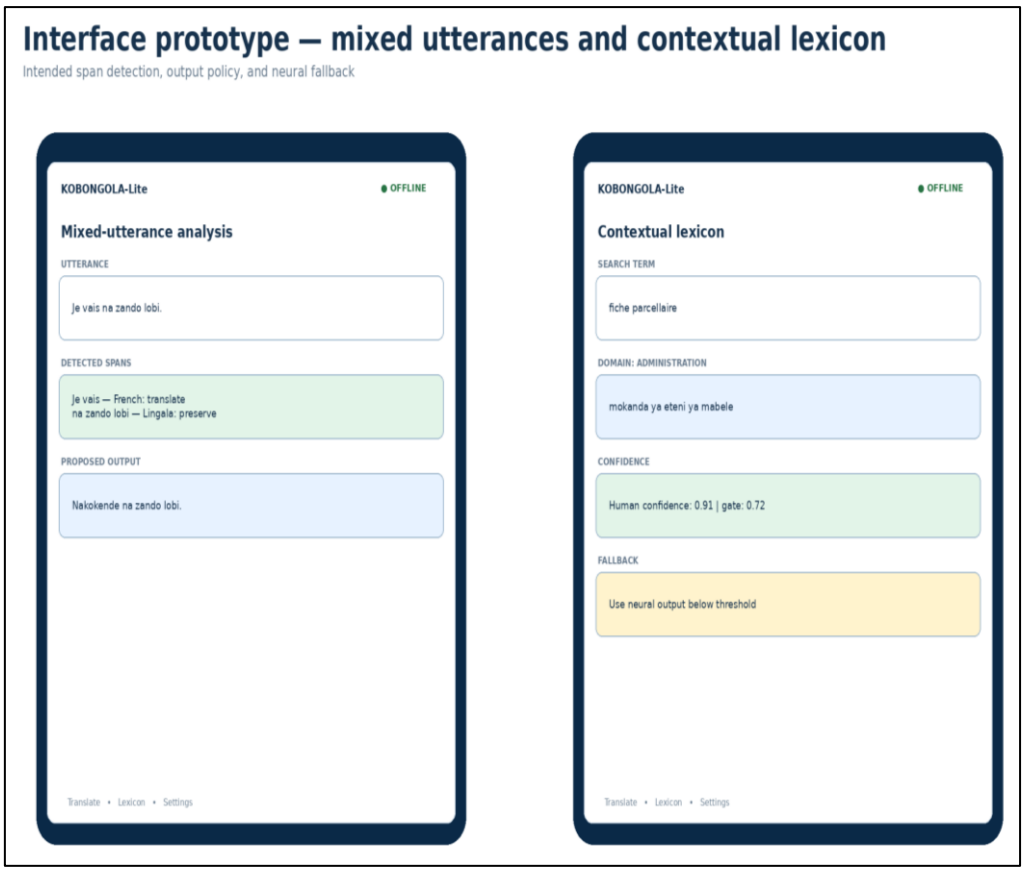

Fig 6 Interface Prototype for Mixed-Utterance Analysis and Contextual Lexical Support.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

VI. EXPERIMENTAL PLAN

➢ *Systems and Ablations*

Table 5 Experimental Systems and Isolated Contrasts

ID	System	Purpose
B0	facebook/nllb-200-distilled-600M; fra_Latn↔lin_Latn; frozen revision/hash; γT decoding	Exact external multilingual baseline
B1	One exact AfriNLLB Transformers French↔Lingala artifact from [6]; immutable repository ID, revision and SHA-256 in registration	Closest teacher/baseline
B2	Deterministic CTranslate2 conversion of B1, v4.6.0, int8_float16, same tokenizer and γT; conversion manifest/hash frozen	Efficient external comparator
D0	Student trained on human references only	Pre-distillation internal control
M0	D0 + sequence-level teacher targets	Distillation effect
M1	M0 + contextual lexical gate	Lexical-support effect
M2	M1 + span policies and hybrid reranking	Complete linguistic prototype
M2-ORT	Unquantized ONNX export of M2	Export-fidelity control
M3	Dynamic INT8 version of M2-ORT	Quantization and mobile candidate

Registration is blocked until the B0 and B1 repository identifiers, immutable revisions, tokenizer revisions, file-level SHA-256 hashes, licenses, and local cache manifests are populated. B0 and B1 use γT exactly as defined in Section V.B. B2 is not optional: it is generated from the frozen B1 artifact with ctranslate2-converter --quantization int8_float16, beam size 5, length penalty 1.0, maximum output length 256, batch size 1, and CPU execution; the converter version, command, vocabulary mapping, and output hashes are frozen.

➢ *Training Budget and Reproducibility*

The confirmatory architecture and vocabulary are fixed as stated above. Training uses AdamW ($\beta1=0.9$, $\beta2=0.98$, $\varepsilon=10^{-8}$, weight decay 0.01), Xavier-uniform initialization, label smoothing 0.10, and gradient-norm clipping at 1.0. The peak learning rate is 3×10^{-4} under an inverse-square-root schedule with 4,000 linear warm-up updates. The per-device batch is capped at 8,192 non-padding source-plus-target tokens; gradient accumulation produces an effective batch of 32,768 tokens. Each condition is trained for at most 100,000 optimizer updates or 50 epochs, whichever occurs first. General-development chrF++ is evaluated after every epoch. Early stopping uses patience five with a minimum improvement of 0.20 chrF++ point; the checkpoint with the highest mean directional development chrF++ is retained, with the earlier checkpoint winning ties. Mixed-precision arithmetic, hardware type, framework version, and realized accumulation count are recorded.

The lexical-gate vector w is estimated only on the code-switch development set by L2-regularized logistic regression, with provenance group as the cross-validation unit. The regularization constant C is selected from {0.01, 0.1, 1, 10} by five-fold grouped cross-validation using mean log loss. The confidence threshold τ is then selected on out-of-fold predictions to maximize preservation macro-accuracy subject to a mean human-adequacy loss no greater than 0.15 on the 1–5 scale relative to M0; ties select the larger τ. Feature scaling statistics, the selected C, coefficients, intercept, τ, fold assignment, and software version are frozen and hashed before final-test access.

Random seeds are 17, 42, and 73. The three seeds apply to initialization, data-order shuffling, dropout, and any stochastic decoding used during development. Exploratory alternatives may be reported, but they cannot replace the confirmatory configuration after the protocol freeze. All commands, environments, checkpoints, tokenizer artifacts, and random states are versioned.

➢ *Automatic and Specialized Metrics*

The primary automatic endpoint is corpus-level chrF++ on the local general test, reported separately by direction. BLEU is computed and reported with the complete SacreBLEU signature to prevent ambiguous score comparisons [20]. AfriCOMET and COMET are validated separately against the confirmatory human sample: for each direction and metric, Spearman's rank correlation ρ is computed across all source–system cells between the automatic segment score and the cell's mean of three blinded adequacy ratings. A metric is confirmatory only if $\rho\geq0.50$ and the document-clustered 10,000-resample percentile-bootstrap 95% lower bound exceeds 0.30 in both directions; otherwise that metric and all analyses based on it are exploratory [15], [21]. No Pearson, pooled-direction, or post-hoc threshold may substitute for this rule.

Preservation macro-accuracy is defined on the frozen terminology/entity/number test over five classes K={domain term, named entity/acronym, number, date, unit}. For class k, Ak equals the number of eligible annotated source items whose output contains one admissible normalized target form, preserves the underlying value where applicable, and contains no contradictory rendering, divided by the total number Nk of eligible class-k items. An item is eligible only when its frozen policy is preserve, normalize, transliterate, or use a validated domain target; multiple occurrences with one

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

annotation identifier count once. The directional score is MacroAcc=(1/5)Σk Ak; the test construction guarantees Nk>0 for every class. Exact-match and class-specific accuracies are also reported, but only this unweighted five-class macro-average enters H2. Specialized reporting additionally includes span-level omission, unwanted translation, preservation, and wrong-language rates, stratified by domain, documented variety, length, and natural versus synthetic code-switching.

➢ *Human Evaluation*

Tuning judgments are collected on a development-only panel that is disjoint from all confirmatory tests and evaluators. It contains 120 code-switch development segments per direction, provenance-group stratified across the frozen categories. For each eligible span, at most the three highest-frequency lexical candidates are presented with source context and intended target language to three of six qualified bilingual development evaluators, who assign binary contextual relevance; majority vote is the gate label. The same six evaluators rate anonymized M0, M1, and candidate-reranked outputs for adequacy on the anchored 1–5 scale, three ratings per output, under a balanced assignment that prevents an evaluator from seeing competing outputs for one source. Out-of-fold majority labels fit w and τ; out-of-fold mean adequacy supplies the τ constraint and the λ-grid constraint. These judgments are used only for tuning, are released with assignments and exclusions, and never enter H2 or H5.

Human evaluation uses 400 unique source segments per direction (800 total), sampled before system inspection: 200 from the local general test, 100 from the natural code-switching test, and 100 from the terminology/entity/number test in each direction. Outputs from M0, M1, M2, and M3 are evaluated, yielding 3,200 system outputs. Each output receives three independent ratings, for 9,600 rating records. At least 12 qualified French–Lingala bilingual evaluators are recruited; each evaluator completes at most 800 ratings in four sessions of no more than 200, with at least 24 hours between sessions. A balanced incomplete-block assignment ensures that no evaluator sees two system outputs for the same source and that every system, direction, and stratum receives equal coverage.

Evaluators are blinded to system identity, and outputs are independently randomized within session. Meaning adequacy on a 1–5 anchored scale is the primary human endpoint; fluency/naturalness on a separate 1–5 scale is secondary. A structured error taxonomy covers mistranslation, omission, addition, named-entity corruption, number/unit alteration, wrong-language output, inappropriate preservation, morphology, word order, and harmful content. Evaluator variety profile and domain competence are recorded without treating one variety as normative. Krippendorff's α with ordinal distance and intraclass correlation are reported; mixed-effects models include random intercepts for segment and evaluator.

➢ *Statistical Analysis*

Confirmatory automatic contrasts use paired bootstrap resampling clustered by document, with 10,000 resamples, direction-specific one-sided 95% confidence bounds for the directional hypotheses, corresponding two-sided 95% intervals for descriptive reporting, and effect sizes. This follows established paired significance-testing principles for machine translation [22]. Three-seed means and standard deviations are reported for trainable systems. Holm correction is applied exactly to the ten directional composite tests defined in Section III.B. For H4, M3 qualifies only when the lower confidence bound of chrF++(M3)−chrF++(M2-ORT) is strictly above −1.0 in each direction and all mandatory mobile criteria are met. A smaller file or lower latency cannot offset linguistic failure, and linguistic non-inferiority cannot offset hardware or network failure.

For H2 and H5, the prespecified statistic is the direction-specific difference in mean adequacy after averaging the three ratings within each source–system cell. A 10,000-resample paired bootstrap samples document clusters with replacement and retains all sampled segment cells and ratings. H2 adequacy non-inferiority passes only when the one-sided 95% lower bound for mean(M1−M0) is strictly above −0.15 point on the 1–5 scale; its intersection-union decision also requires preservation superiority. H5 passes only when the corresponding lower bound for mean(M2−M0) is strictly above 0. Studentized mixed-effects estimates with random intercepts for document, segment, and evaluator are sensitivity analyses and cannot replace the bootstrap decisions. The power rule is the one specified in Section III.C.

Raw p-values are constructed from the same 10,000 paired document-clustered resamples. Every component is written as an effect E for which larger is better, with null boundary m: H1, E=chrF++(M2)−chrF++(M0), m=0; H2-preservation, E=MacroAcc(M1)−MacroAcc(M0), m=0; H2-adequacy, E=adequacy(M1)−adequacy(M0), m=−0.15; H3, E=CER(M1)−CER(M2), m=0, where CER is exactly the segment-level critical-error rate in Section V.D; H4, E=chrF++(M3)−chrF++(M2-ORT), m=−1.0; and H5-human, E=adequacy(M2)−adequacy(M0), m=0. For observed $\hat{E}$ and bootstrap replicate E*b, the one-sided centered-bootstrap p-value is p=(1+Σb I[E*b−$\hat{E}$≥$\hat{E}$−m])/(10,001). Thus no raw p-value can be zero.

For each direction, pH1 is its H1 component; pH2=max(pH2-preservation,pH2-adequacy); pH3 is its CER component; pH4 equals its linguistic component only if every mobile gate passes and otherwise equals 1.00; and pH5=max(pH1,pH5-human). These ten fixed composite values, two per hypothesis, are ordered from smallest to largest and compared sequentially with α/(10−j+1), j=1,...,10; testing stops at the first non-rejection and all remaining hypotheses are not rejected. Holm-adjusted p-values are reported as max up to rank j of (10−i+1)p(i), capped at 1.00, then mapped back to hypothesis and direction. Passing every corresponding unadjusted component confidence bound is

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

necessary but not sufficient; only the Holm-adjusted result determines confirmatory rejection.

VII. PHYSICAL-DEVICE VALIDATION

➢ *Target Profiles and Controlled Workloads*

The confirmatory runtime is ONNX Runtime Mobile on CPU. The primary device is a Samsung Galaxy A03 Core with 2 GB RAM, Unisoc SC9863A system-on-chip, 64-bit ARM CPU, and Android 11 Go; the secondary device is a Samsung Galaxy A14 with 4 GB RAM, Exynos 850 system-on-chip, 64-bit ARM CPU, and Android 13. The exact retail model codes, OS build numbers, security patches, battery-health estimates, storage occupancy, ONNX Runtime version, application hash, thread count (four), CPU execution provider, and power mode are recorded in the registration; substitution is prohibited in confirmatory analysis and creates an exploratory device stratum. Workloads are frozen by direction, sentence length, code-switch status, terminology rarity, domain, and documented variety.

Each frozen workload stratum is executed in 100 timed warm-run translations per direction on each device to estimate p50 and p95 latency. Ten cold starts per device measure initial loading. The 100 warm runs follow a frozen randomized order and are preceded by five untimed warm-up translations that are excluded from analysis. A minimum 15-minute continuous endurance test per direction assesses thermal drift. Battery state, brightness, background processes, starting temperature, and charging status are controlled. Runs affected by thermal throttling are flagged rather than silently discarded.

Energy is measured with a Monsoon High Voltage Power Monitor at 5.0 V and 5 kHz after battery-path calibration. Airplane mode is enabled, brightness is fixed at 200 cd/m², adaptive brightness and battery saver are disabled, the battery state is held at 50–80%, and device temperature must be 22–28 °C before a run. A five-minute idle trace under the same state estimates Pidle. Net energy is $E=\int \max(P(t)-P_{idle},0)dt$ over each batch; joules per sentence and per 100 generated tokens are reported with a 1,000-resample batch-bootstrap 95% interval. Three independent 100-sentence batches per direction and device are required; failed traces are retained and rerun once under the preregistered rule.

Table 6 Mandatory Mobile Measurements and Qualification Criteria

Dimension	Measure	Confirmatory criterion
Storage	Quantized ONNX model; tokenizer; lexicon; rules; APK	M3 neural model ≤120 MB; other assets reported separately
Memory	PSS/RSS before load, after load, and peak	Peak ≤800 MB on the ≤2 GB target
Latency	Cold start, p50, p95, ms/token, tokens/s	20-token standard: p50 ≤1.5 s; p95 ≤3.0 s
Energy	J or mWh per sentence and per 100 tokens	Mandatory method and uncertainty; no threshold
Thermal	Initial/peak temperature and throttling	No interruption or failure in endurance run
Reliability	Crash, timeout, empty output, decode failure	100% completion of the main benchmark
Network	Permissions, packets, DNS, connections	Zero outbound application packet

➢ *Proof of Offline Operation*

The evaluated application contains the quantized model, tokenizer, lexicon, span policies, and all inference dependencies. After installation, no language resource or model parameter may be downloaded. Wi-Fi, mobile data, Bluetooth, and VPN are disabled; the application is closed and relaunched; both translation directions are exercised; the device is rebooted; and the scenario is repeated. The confirmatory Android manifest contains no INTERNET permission. Packet capture or a local firewall independently verifies network silence.

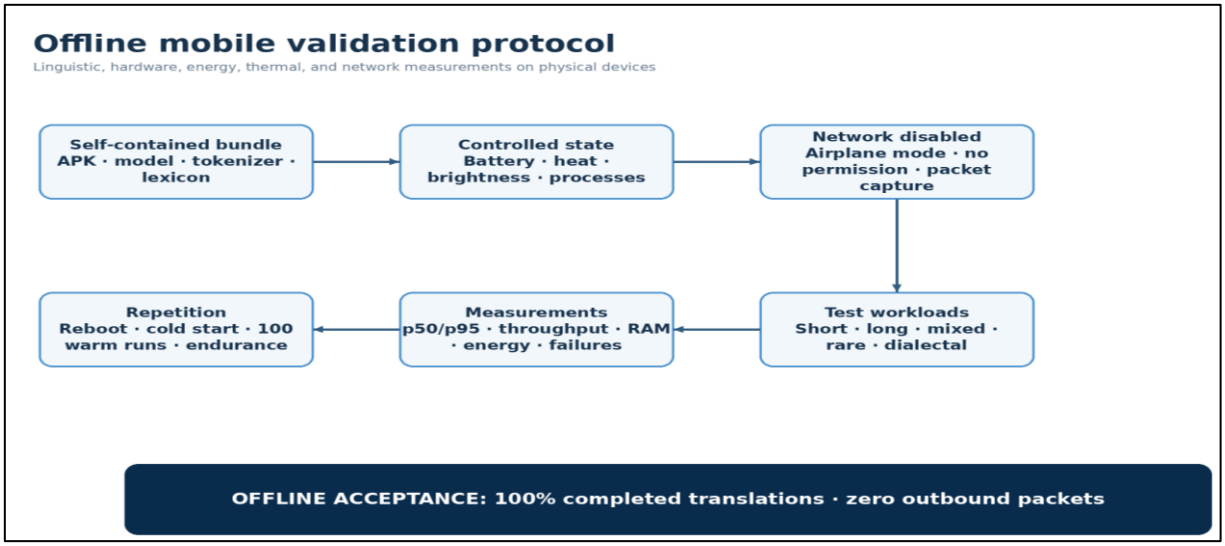

Fig 7 Physical-Device Benchmark and Proof of Offline Operation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

VIII. REPRODUCIBILITY AND QUALITY ASSURANCE

The release package will contain a corpus datasheet and data statement describing motivation, composition, population, collection, annotation, preprocessing, recommended uses, restrictions, maintenance, and risks [16], [17]. Each model receives a model card documenting architecture, data, metrics, limitations, energy use, covered varieties, Android permissions, and warnings [18]. When text cannot be released, metadata, statistics, transformation scripts, and cryptographic hashes are published to the extent permitted by licenses and consent.

Before the final test is opened, the project verifies licenses and withdrawals; audits exact, near, reversed, and derivative duplicates across partitions; freezes manuals, agreement thresholds, metrics, margins, seeds, scripts, and configurations; conducts dry runs only on synthetic or development data; independently verifies hashes and test inaccessibility during tuning; and audits the Android package, permissions, and traffic. The final paper will report every prospectively specified system, including failures, null effects, confidence intervals spanning zero, incompatible devices, and network anomalies. Unplanned analyses are labeled exploratory.

The time-stamped registration must precede confirmatory training, final-test decryption, or generation of any final-test system output. It contains the exact B0–B2 and D0–M3 manifests, corpus/test hashes, device identifiers, power protocol, analysis code hash, ten-test Holm family, exclusions, missing-data rules, and abort/retry rules. One amendment window is permitted only before any final-test output exists; each amendment is time-stamped with rationale and preserves the prior version. After final-test access, no confirmatory replacement, margin change, sample resizing, feature redefinition, or device substitution is permitted. Deviations and failed runs remain in the public log and are reported as deviations or exploratory analyses. The repository URL and persistent identifier replace the prospective placeholder before the manuscript is described as preregistered.

IX. ETHICS, SAFETY AND COMMUNITY GOVERNANCE

Human contributors receive understandable information about purpose, retention, dissemination, remuneration, and the right to withdraw before data freeze. Speakers and linguists participate in annotation guidance, quality criteria, error interpretation, and release decisions, consistent with participatory approaches to African-language technology [19]. Results are disaggregated to prevent aggregate scores from hiding degradation in a variety or domain. Differences are described as limitations of the data and model, not as properties of speakers.

KOBONGOLA-Lite is not a medical device, legal tool, or certified interpreter. Translations in health, law, security, or high-stakes administration require review by a qualified person. Offline execution improves control over data flows but does not alone guarantee confidentiality if histories, logs, or backups are stored; such functions are disabled by default and audited. Community benefit includes access to the application, French and Lingala documentation, optional recognition of contributors, an error-reporting mechanism, and a correction pathway.

X. LIMITATIONS

Representative data may remain insufficient across varieties, age groups, domains, and registers. Code-switching boundaries are sociolinguistically contested, and annotation cannot eliminate all ambiguity. Human assessment is costly and sensitive to fatigue, evaluator background, and orthographic preference. Android fragmentation limits generalization beyond tested profiles. AfriCOMET may correlate inadequately with French–Lingala judgments. The multilingual teacher can transmit its biases and errors. Accordingly, the study can establish only the performance of the frozen systems on the predefined populations, domains, test sets, and devices; broader claims require independent replication.

XI. CONCLUSION

KOBONGOLA-Lite is currently a specified architecture accompanied by interface prototypes; this manuscript does not claim an executable model-integrated offline APK or an empirically validated translation system. It defines a falsifiable prospective evaluation that treats AfriNLLB and recent French–Lingala work as direct comparators, isolates distillation, lexical support, span policies, reranking, export, and quantization, and requires simultaneous linguistic and mobile evidence.

If the hypotheses are supported, the study will provide reproducible evidence that locally governed data and explicit hybrid controls can improve the quality–efficiency trade-off for French–Lingala translation on constrained devices. If they are rejected, the same protocol will identify whether the limiting factor is data, lexical intervention, code-switch processing, compression, or hardware execution. In either case, the scientific contribution rests on traceable evidence rather than a technological promise.

DECLARATIONS

Funding—This research received no external funding. Conflicts of interest—The authors declare no conflict of interest. Ethics—This prospective protocol reports no completed human-participant data collection or evaluation. Any future collection, annotation study, or human evaluation will begin only after written approval or documented exemption from the competent ethics body and will follow the consent and governance procedures specified above. Data and code availability—No empirical dataset or trained model accompanies this prospective protocol. Upon execution, code, frozen configurations, analysis scripts, power-simulation code, and nonrestricted metadata will be released; restricted text will be documented and accessed only under

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug847

its license and consent conditions. Author contributions—KINKETE MFUMABI Hervé led conceptualization, methodology, protocol design, architecture specification, visualization, writing—original draft, and project administration. NGOY DÉOGRACIAS, NGOMA MAGEMA ARSENE, YAMONE MUISANGIE CLAIRETTE, KAPINGA NSAPO MARISA, MPOYI KABULU EXAUCE, LIKANGO BOKONDO CHRISTOPHER, and KAPONGO MUSANGU ISRAËL contributed to the conception of KOBONGOLA-Lite and to writing—review and editing. All authors reviewed and approved the submitted manuscript.

REFERENCES

[1]. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.

[2]. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[3]. M. R. Costa-jussà et al., "No Language Left Behind: Scaling human-centered machine translation," arXiv:2207.04672, 2022, doi: 10.48550/arXiv.2207.04672.

[4]. S. Kabongo Kabenamualu, V. Marivate, and H. Kamper, "LiSTra automatic speech translation: English to Lingala case study," in Proc. DCLRL, 2022, pp. 63–67, doi: 10.18653/v1/2022.dclrl-1.8.

[5]. R. E. Mandiya et al., "A Transformer-based method for bidirectional French-Lingala machine translation in speech and text," Applied Sciences, vol. 16, no. 7, art. 3399, 2026, doi: 10.3390/app16073399.

[6]. Y. Moslem, A. K. Wassie, and A. G. Abebe, "AfriNLLB: Efficient translation models for African languages," in Proc. AfricaNLP, 2026, doi: 10.18653/v1/2026.africanlp-main.30.

[7]. R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in Proc. ACL, 2016, pp. 1715–1725, doi: 10.18653/v1/P16-1162.

[8]. T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer," in Proc. EMNLP System Demonstrations, 2018, pp. 66–71, doi: 10.18653/v1/D18-2012.

[9]. D. I. Adelani et al., "A few thousand translations go a long way! Leveraging pre-trained models for African news translation," in Proc. NAACL, 2022, doi: 10.18653/v1/2022.naacl-main.223.

[10]. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[11]. Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in Proc. EMNLP, 2016, pp. 1317–1327, doi: 10.18653/v1/D16-1139.

[12]. Y. Lin, X. Wang, Z. Zhang, M. Wang, T. Xiao, and J. Zhu, "MobileNMT: Enabling translation in 15MB and 30ms," in Proc. ACL Industry Track, 2023, pp. 368–378, doi: 10.18653/v1/2023.acl-industry.36.

[13]. M. Popović, "chrF: Character n-gram F-score for automatic MT evaluation," in Proc. WMT, 2015, pp. 392–395, doi: 10.18653/v1/W15-3049.

[14]. M. Popović, "chrF++: Words helping character n-grams," in Proc. WMT, 2017, pp. 612–618, doi: 10.18653/v1/W17-4770.

[15]. J. Wang et al., "AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages," in Proc. NAACL, 2024, pp. 5997–6023, doi: 10.18653/v1/2024.naacl-long.334.

[16]. E. M. Bender and B. Friedman, "Data statements for natural language processing," Transactions of the ACL, vol. 6, pp. 587–604, 2018, doi: 10.1162/tacl_a_00041.

[17]. T. Gebru et al., "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[18]. M. Mitchell et al., "Model cards for model reporting," in Proc. FAT*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[19]. W. Nekoto et al., "Participatory research for low-resourced machine translation: A case study in African languages," in Findings of EMNLP, 2020, pp. 2144–2160, doi: 10.18653/v1/2020.findings-emnlp.195.

[20]. M. Post, "A call for clarity in reporting BLEU scores," in Proc. WMT, 2018, pp. 186–191, doi: 10.18653/v1/W18-6319.

[21]. R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in Proc. EMNLP, 2020, pp. 2685–2702, doi: 10.18653/v1/2020.emnlp-main.213.

[22]. P. Koehn, "Statistical significance tests for machine translation evaluation," in Proc. EMNLP, 2004, pp. 388–395.

[23]. M. Meeuwis, Lingala. Munich, Germany: LINCOM Europa, 1998.

[24]. C. Myers-Scotton, Social Motivations for Codeswitching: Evidence from Africa. Oxford, U.K.: Oxford University Press, 1993.

[25]. T. Nguyen, L. T. Nguyen, and D. Q. Nguyen, "How effective is machine translation on low-resource code-switching? A case study comparing human and automatic metrics," in Findings of ACL, 2023, pp. 14186–14195, doi: 10.18653/v1/2023.findings-acl.893.

[26]. C. Hokamp and Q. Liu, "Lexically constrained decoding for sequence generation using grid beam search," in Proc. ACL, 2017, pp. 1535–1546, doi: 10.18653/v1/P17-1141.

[27]. M. Post and D. Vilar, "Fast lexically constrained decoding with dynamic beam allocation for neural machine translation," in Proc. NAACL-HLT, 2018, pp. 1314–1324, doi: 10.18653/v1/N18-1119.

[28]. G. Dinu, P. Mathur, M. Federico, and Y. Al-Onaizan, "Training neural machine translation to apply terminology constraints," in Proc. ACL, 2019, pp. 3063–3068, doi: 10.18653/v1/P19-1294.