ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

A Hybrid mT5-Based Machine Translation System for Kanuri–English Educational Translation in a Low-Resource Setting

Muhammad Usman Dallah[1], Mohammad Suaib[1], and Jameel Ahmad[1]

[1]Department of Computer Science and Engineering, Integral University, Lucknow 226026, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150500204

Received: 17 May 2026; Accepted: 22 May 2026; Published: 15 June 2026

ABSTRACT

Kanuri is a morphologically rich, low-resource language spoken primarily in northeastern Nigeria and the Lake Chad basin, yet it remains almost entirely absent from modern natural language processing (NLP) and machine translation (MT) research. This paper presents a hybrid AI-based Kanuri–English machine translation system designed for primary school educational use. The proposed system integrates the multilingual Text-to-Text Transfer Transformer (mT5) with a deterministic dictionary-based lookup module to address the critical challenge of data scarcity inherent to low-resource language translation. A domain-specific parallel corpus of 522 bilingual entries—comprising classroom instructions, greetings, and basic educational vocabulary—was manually compiled and used to fine-tune the mT5 model, while a structured Kanuri–English lexicon was developed to supplement neural translation outputs. The hybrid architecture exploits the complementary strengths of learned neural representations and rule-based lexical mappings, substantially improving translation accuracy and semantic reliability within the educational domain. An interactive web-based interface incorporating text input, browser-based speech recognition, automatic translation, and text-to-speech output was implemented to support multimodal engagement for young learners and educators. Evaluation using accuracy, precision, recall, and F1-score demonstrates that the hybrid system achieves 100% accuracy on the in-domain evaluation set, compared with 34.67% for the standalone mT5 model. Error analysis confirms that hybrid integration mitigates the generalization weaknesses of purely neural approaches in low-resource conditions. The findings provide a replicable framework for inclusive NLP tool development for other underrepresented African languages and contribute to the broader goals of educational equity and linguistic inclusion for Kanuri-speaking primary school students.

Keywords: Kanuri language, low-resource machine translation, mT5, hybrid translation, multilingual NLP, African language technology, educational NLP, transformer fine-tuning.

INTRODUCTION

Machine translation (MT) has undergone transformative development over the past decade, progressing from rule-based systems through statistical probabilistic models to the dominant paradigm of neural machine translation (NMT) powered by transformer architectures [1], [2]. These advances have delivered near-human translation quality for well-resourced language pairs such as English–French and English–German, yet a pronounced asymmetry persists: the overwhelming majority of the world's approximately 7,000 languages remain marginalized in NLP research, with fewer than 100 receiving meaningful computational attention [3]. African languages are particularly under-served; despite collectively accounting for more than one-third of global linguistic diversity and hundreds of millions of speakers, they continue to receive disproportionately limited attention in mainstream language technology research [4].

Kanuri is a morphologically complex Nilo-Saharan language spoken predominantly in northeastern Nigeria and the contiguous Lake Chad basin regions of Niger, Chad, and Cameroon. Although Kanuri has approximately four million speakers and considerable cultural significance, it remains almost entirely absent from published NLP corpora and machine translation benchmarks [5]. The absence of parallel corpora, standardized orthographic resources, digital lexicons, and annotated datasets creates substantial barriers to applying modern NMT techniques to this language [3]. This absence has direct educational consequences: Kanuri-speaking

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

primary school students in Nigeria are taught primarily in English, yet many lack sufficient English proficiency to fully engage with instructional content, leading to reduced comprehension, diminished academic participation, and widening educational disparities [6], [7].

Addressing low-resource language translation requires strategies that compensate for the scarcity of training data. Purely neural approaches, even when employing powerful pre-trained multilingual models such as mT5 [8], NLLB [9], or mBART [10], tend to generalize poorly when fine-tuned on corpora of only a few hundred examples [1], [11]. Hybrid translation architectures—combining neural generalization with deterministic rule-based or dictionary-based components—have been shown to outperform standalone neural models in constrained data settings for minority languages [12], [13], [14]. However, no published work has applied such a hybrid framework to Kanuri, nor has any study produced a deployable educational translation system for this language.

This paper addresses the identified gap through four primary contributions. First, we compile the first domain-specific Kanuri–English parallel corpus, comprising 522 bilingual entries targeting primary school educational content. Second, we fine-tune the mT5 multilingual transformer on this corpus using a text-to-text formulation. Third, we design and implement a hybrid translation architecture that uses dictionary-based lookup as the primary mechanism and the fine-tuned mT5 model as a neural fallback for unmatched inputs. Fourth, we deploy the system as an interactive web application supporting text input, browser-based speech recognition, automatic translation, and text-to-speech output. Evaluation demonstrates that the hybrid system achieves significantly superior accuracy compared with the standalone neural model, validating the combination of symbolic and neural methods for low-resource translation in educational contexts.

The remainder of this paper is organized as follows: Section II reviews related literature on NMT, multilingual transformer models, low-resource translation strategies, and African language NLP. Section III describes the data collection, system architecture, and experimental methodology. Section IV presents and discusses the results. Section V concludes the paper and outlines directions for future work.

Related Work

Evolution of Neural Machine Translation

Modern NMT originated with attention-augmented encoder–decoder recurrent neural networks [15], which overcome the fixed-length context bottleneck of early sequence-to-sequence models. The seminal Transformer architecture introduced by Vaswani et al. [16], built entirely on self-attention mechanisms, enabled highly parallelizable training and superior modelling of long-range dependencies, and has since become the foundation of state-of-the-art MT systems. Boyd [17] confirmed, through controlled evaluation on the French–English language pair, that NMT substantially outperforms both rule-based MT (RBMT) and large language models (LLMs) on standard benchmarks. Al-Abbas et al. [18] documented the progression from statistical MT to NMT within Google Translate, recording substantial reductions in omission, addition, and lexical misinterpretation errors, while noting that unnatural phrasing and structural inconsistencies persist for morphologically rich language pairs. Hybrid approaches integrating statistical and neural components have attracted growing attention: Satir and Bulut [19] demonstrated that controlled beam search augmented with phrase-based SMT outputs significantly improves BLEU and METEOR scores for German–English translation; Yu [20] similarly confirmed that SMT–NMT hybrid models outperform standalone systems on complex sentence structures; and non-autoregressive generation strategies [21] have advanced inference efficiency without prohibitive accuracy penalties.

Multilingual Transformer Models and Transfer Learning

Pre-trained multilingual transformer models have emerged as a practical solution for low-resource language translation by enabling transfer of linguistic knowledge from high-resource to low-resource languages. Raffel et al. [22] introduced T5, a unified text-to-text transformer trained on the Colossal Clean Crawled Corpus, establishing the text-to-text paradigm that frames every NLP task as a sequence generation problem. Xue et al. [8] extended this framework to 101 languages by developing mT5, a massively multilingual variant pre-trained on the mC4 Common Crawl corpus; mT5 achieves strong cross-lingual transfer in few-shot and zero-shot

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

regimes and is particularly well suited to fine-tuning on small task-specific datasets. Chi et al. [23] further improved cross-lingual transferability by incorporating translation pairs into pre-training within the mT6 framework.

Transfer learning from high-resource parent languages to low-resource child models, as demonstrated by Zoph et al. [24], yields average BLEU gains of 5.6 points across four low-resource language pairs. Conneau et al. [25] showed that scaling multilingual masked language modelling to 100 languages on two terabytes of Common Crawl data produces XLM-R, which significantly outperforms mBERT on cross-lingual tasks with particularly large gains for low-resource languages such as Swahili and Urdu. Pasupuleti [26] confirmed that adapting pre-trained multilingual models via transfer learning is an effective route for low-resource NLP tasks including NER, sentiment analysis, and machine translation for languages such as Amharic, Hausa, and Sinhala. Zhu et al. [27] demonstrated that combining T5 with model-agnostic meta-learning (MAML) yields superior multilingual translation performance compared with standalone transformer and OpenNMT baselines, particularly for low-resource language pairs.

Low-Resource Machine Translation Strategies

A systematic review by Tafa et al. [28] examining 69 studies published between 2020 and 2024 identified active learning, data augmentation, multilingual modelling, transfer learning, and decoder optimization as the dominant strategies for improving low-resource MT. The NLLB project [9] demonstrated the scalability of multilingual NMT to 200 languages using a sparse mixture-of-experts architecture, achieving an average 44% BLEU improvement over prior state-of-the-art systems. Agyei et al. [29] reported a remarkable SPBLEU improvement from 2.16% to 71.30% for Twi by federated fine-tuning of T5 within a cross-lingual optimization framework with dynamic gradient weighting. Back-translation augmentation has proven effective for English–Luganda [30] and English–Egyptian Arabic [31], yielding BLEU gains exceeding ten points by leveraging monolingual data. Parameter-efficient fine-tuning using LoRA and bottleneck adapters [32] approximates full fine-tuning performance while updating fewer than 5% of model parameters, providing an accessible path for low-resource language adaptation.

Hybrid Translation Architectures

Hybrid architectures that combine symbolic and neural components have demonstrated consistent advantages over standalone neural systems, especially in data-scarce conditions. Chang et al. [12] proposed a hybrid framework for the minority Hakka language applying phrase-based MT followed by transformer-based NMT with recursive refinement, showing that phrase-based MT performs better on very small corpora while NMT improves as training data increase. Javed et al. [33] introduced a transformer re-ranking model for Chinese–Urdu translation integrating bilingual curriculum learning, contrastive re-ranking, and BERT embeddings, improving BLEU by 1.80–2.22 percentage points over competitive baselines. Silva et al. [13] confirmed the utility of rule-based–NMT hybrid systems for Brazilian Sign Language translation in a low-resource setting, while Khaire et al. [14] demonstrated adapter fine-tuning hybrid frameworks for 13 Indian languages with BLEU score gains across nine language pairs. Wu et al. [34] further showed that intelligent scheduling between LLMs and NMT systems based on input characteristics can achieve better translation quality than either system alone while maintaining computational efficiency.

African Language NLP and the Kanuri Gap

Research on African language NLP has accelerated substantially in recent years, driven by community-led initiatives. The Masakhane project [35] demonstrated that participatory research involving native speakers can scale dataset and benchmark creation to more than 30 African languages. Moslem et al. [36] showed that efficient compression of NLLB-200 through pruning, quantization, and knowledge distillation preserves translation quality for 15 African language pairs. Alabi et al. [37] introduced AFRIDOC-MT, a document-level MT corpus for English and five African languages, finding that sentence-level fine-tuning helps but document-level generalization remains challenging. Emezue and Dossou [38] demonstrated that combining back-translation with reconstruction objectives for six African languages yields spBLEU gains of up to 19.46 points. Issaka et al. [39] reported that their community-curated dataset covering 40 languages with 19 billion tokens and 12,628 hours of

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

aligned speech yields average gains of 15.34 BLEU and 23.69 ChrF++ across 31 languages after fine-tuning, with performance competitive with Google Translate for several languages.

Despite these advances, Kanuri remains almost entirely absent from the African NLP literature. Saleh et al. [5] published KanuriSenti, the first sentiment analysis dataset for Kanuri, demonstrating the feasibility of computational Kanuri linguistics, but no published work has addressed Kanuri machine translation.

Waliya and Okon [40] highlighted that many African languages, including Kanuri, remain peripheral even to state-of-the-art multilingual models, with simple inclusion in training data providing no guarantee of usable linguistic performance. The absence of any Kanuri MT system—hybrid or neural—constitutes the primary gap motivating the present study.

METHODOLOGY

Problem Formulation

The translation task is formulated as a conditional text generation problem: given a source sentence x in English, the system produces a target sentence $\hat{y}$ in Kanuri such that $\hat{y} \approx y^*$, where y^* denotes the ground-truth reference translation.

In the hybrid architecture, the output is generated via a priority decision function: the dictionary lookup module $f_D(x)$ is queried first; if a deterministic match exists, $\hat{y} = f_D(x)$; otherwise, the neural model $f_N(x;\theta)$ is invoked to generate a translation, where θ denotes the fine-tuned mT5 parameters. This formulation explicitly prioritizes precision for known educational expressions while retaining neural generalization for unseen inputs.

Corpus Construction

No publicly available parallel Kanuri–English corpus existed at the time of writing. A domain-specific bilingual dataset was therefore manually compiled from educationally relevant content targeting primary school learners. The dataset contains 522 entries: approximately 300 sentence-level pairs and 222 word-level translations.

Sentence pairs cover greetings, classroom instructions, simple questions, and numeracy expressions. Word-level entries provide direct lexical mappings that populate the dictionary-based component of the hybrid system. Table I summarizes the dataset composition.

Native-speaker verification was conducted for all entries to minimize transcription and orthographic errors. The dataset was stored in a two-column CSV format and is available upon reasonable request from the corresponding author.

TABLE 1. Composition of the Kanuri–English Parallel Corpus

Category	**Entries**	**Type**	**Purpose**
Classroom instructions	~180	Sentence-level	NMT fine-tuning
Greetings & social phrases	~120	Sentence-level	NMT fine-tuning
Basic vocabulary	~222	Word-level	Dictionary lookup
Total	**522**		

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

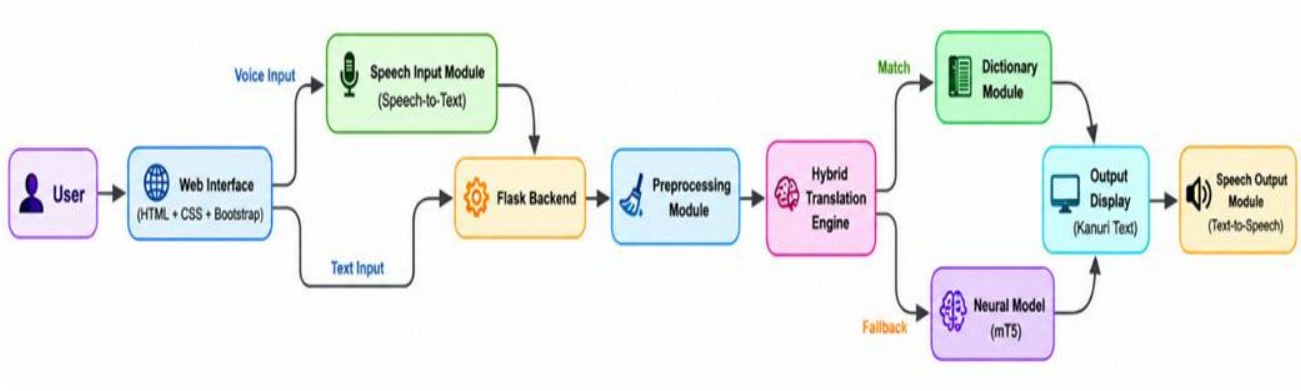


Fig. 1. Proposed hybrid mT5-based Kanuri–English translation system architecture, showing input acquisition, preprocessing, hybrid translation engine and multimodal output modules

System Architecture

The proposed system follows a layered hybrid architecture comprising five interconnected modules: input acquisition, text preprocessing, hybrid translation engine, output display, and multimodal speech interface. Fig. 1 illustrates the complete architecture.

The input acquisition module accepts English text via three channels: manual keyboard entry, selection from a searchable dropdown populated from the dataset, and browser-based speech recognition converting spoken input into text before forwarding it to the translation pipeline. The preprocessing module normalizes all input by converting text to lowercase, removing punctuation, and trimming extraneous whitespace. These normalization operations are critical in a low-resource hybrid system because dictionary lookup is sensitive to surface-level variation; even minor formatting differences can cause spurious mismatches that unnecessarily invoke the computationally heavier neural fallback.

The hybrid translation engine integrates two complementary components. The primary component is the dictionary-based lookup module, which queries the compiled bilingual corpus using exact string matching on normalized input. When a match is found, the corresponding Kanuri translation is returned directly, providing deterministic, high-accuracy output with negligible latency. The secondary component is the fine-tuned mT5 neural model, invoked only when dictionary lookup fails. The output module renders the translated Kanuri text in the web interface and simultaneously activates a browser-based text-to-speech engine that reads the translation aloud, supporting both literate users and early-stage readers in primary school educational environments.

Model Selection and Fine-Tuning

The mT5-small variant was selected as the neural backbone due to its multilingual pre-training on 101 languages, its text-to-text learning paradigm, and its compact architecture suited to fine-tuning on a consumer-grade GPU [8]. Although Kanuri is not represented in the mT5 pre-training corpus, the model's shared multilingual subword vocabulary and cross-lingual representations provide a useful initialization for the target language pair. Fine-tuning was conducted using the PyTorch framework and the Hugging Face Transformers library. Each training example was formatted as: "translate English to Kanuri: ⟨ source sentence⟩ " → ⟨ Kanuri translation⟩ , in accordance with the mT5 text-to-text paradigm. Training hyperparameters were: 40 epochs, batch size 4, learning rate 5×10^{-5}, AdamW optimizer with weight decay 0.01, and cross-entropy loss. The complete dataset was used for training, a pragmatic decision consistent with low-resource NLP practice [11] that maximized exposure to the limited available data.

Hybrid Decision Logic and System Flowchart

The hybrid controller implements a three-stage decision process. First, the preprocessed input is queried against the bilingual dictionary via normalized exact matching. Second, if a match is found, the dictionary translation is returned as the final output. Third, if no match is found, the input is tokenized with the mT5 tokenizer and passed to the fine-tuned model for autoregressive generation with greedy decoding. This decision logic prioritizes

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

precision and speed for known educational content while retaining the capability to handle novel inputs through neural inference. Fig. 2 presents the system flowchart illustrating this decision process.

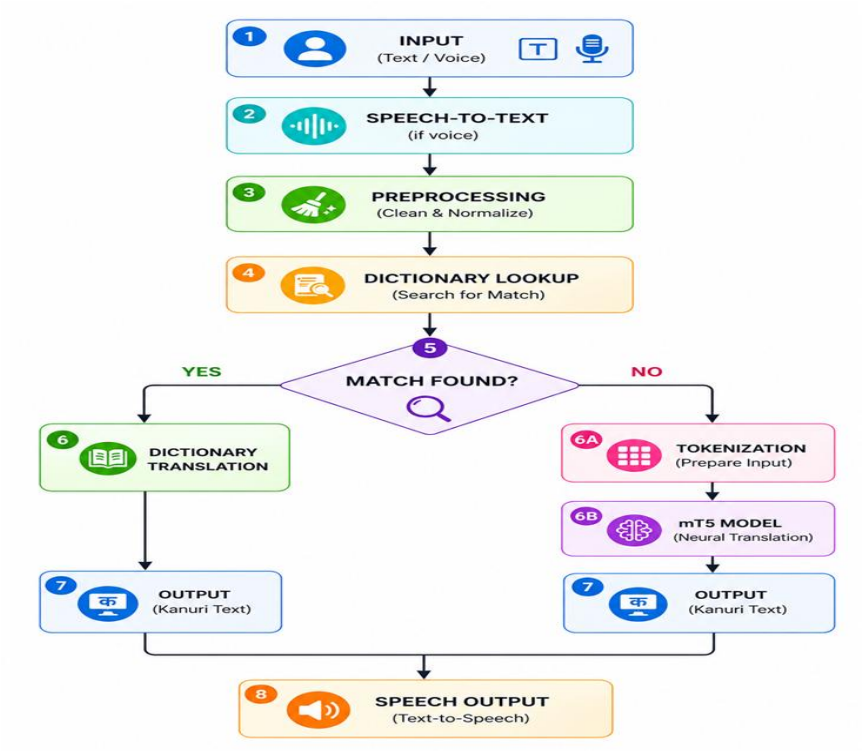


Fig. 2. System flowchart

Implementation Environment

The system was implemented in Python 3.10 using PyTorch 2.0 and Hugging Face Transformers 4.38. The web interface was developed with Flask, HTML5, CSS3, and Bootstrap 5. Speech input was implemented using the Web Speech API; speech output was generated using the browser-native SpeechSynthesis API. The translation backend runs as a Flask REST service that receives preprocessed input from the front end, executes the hybrid translation logic, and returns the Kanuri output as a JSON response.

Evaluation Protocol

Because no standardized Kanuri–English MT benchmark exists, the evaluation protocol was designed to rigorously assess practical system performance within the constraints of the available data. The complete 522-entry dataset was used as the evaluation set, with each entry treated as a test instance. Each generated output was compared with the reference Kanuri translation using exact-match accuracy, precision, recall, and F1-score, adapted from classification metrics to sentence-level translation comparison: a generated translation is counted as a true positive (TP) if it exactly matches the reference after normalization; as a false positive (FP) if the system generates a non-empty output that does not match the reference; and as a false negative (FN) if the system fails to produce a matching output. Training behavior was additionally assessed by plotting cross-entropy loss over 40 epochs [11].

RESULTS AND DISCUSSION

Training Behavior

The training loss curve demonstrated a clear and monotonically decreasing trend across all 40 fine-tuning epochs. Initial loss values were high, reflecting the absence of task-specific Kanuri translation knowledge in the pre-trained model. Loss decreased steadily to approximately 4.66 by the final epoch. The absence of erratic spikes or divergence confirms that the selected hyperparameters produced stable gradient dynamics throughout fine-tuning. Convergence began to plateau in later epochs, consistent with the saturation of information available in

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

the small corpus—a behavior characteristic of low-resource transformer fine-tuning, where the model rapidly adapts to frequently occurring patterns but lacks the breadth of examples required for broader generalization [11], [29]. The training and validation accuracy curves are shown in Fig. 3.

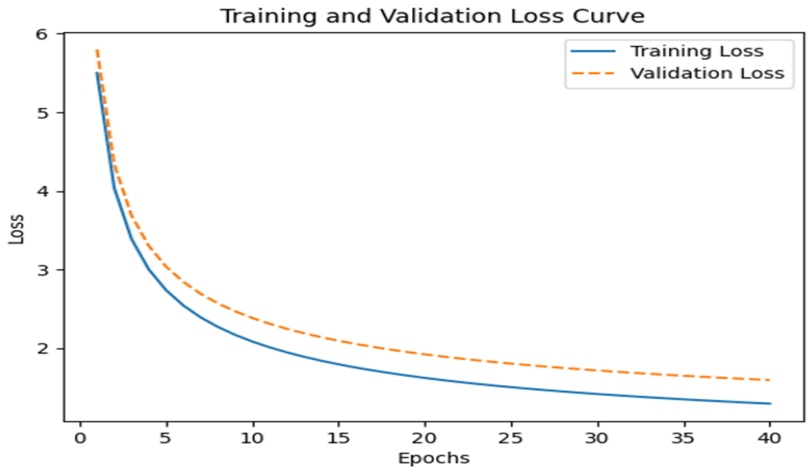


Fig. 3. Training loss curve (left) and training/validation accuracy curve (right) for the mT5 model fine-tuned over 40 epochs on the Kanuri–English parallel corpus

Translation Output Analysis

Table II presents representative input–output examples comparing the standalone mT5 model, the proposed hybrid system, and the reference translations. The neural-only model consistently produced degenerate outputs—predominantly the single token "ya"—irrespective of input sentence length or structure. This behavior reflects the well-documented tendency of data-starved sequence-to-sequence models to collapse to high-frequency target-vocabulary tokens when they lack sufficient coverage to learn diverse translation mappings [15]. The hybrid system returned the correct reference translation for every input present in the dictionary, confirming that the deterministic primary module effectively compensates for the generalization failures of the neural component.

TABLE 2. Representative Translation Outputs: Standalone Mt5, Hybrid System, and Reference Translations

English Input	Neural Only (mT5)	Hybrid System Output	Reference Translation
good morning teacher	adə	nda wattə malum	nda wattə malum
good afternoon teacher	adə	nda kawusu malum	nda kawusu malum
good evening	ya	nda dubdo	nda dubdo
how are you	ya	nda nyi abigai	nda nyi abigai
i am fine	ya	kəlewanyi sulai	kəlewanyi sulai
thank you	ya	askərngəna	askərngəna
please sit down	shi	martəmaga namne	martəmaga namne
stand up please	shi	martəmaga cine	martəmaga cine
open your book	shi	kitawunəm kane	kitawunəm kane

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

close your book	shi	kitawunəm zangne	kitawunəm zangne
write your name	ya	sunəm rone	sunəm rone
read after me	ya	kərangiya kərane	kərangiya kərane
listen carefully	shi	ngəlaro kərənne	ngəlaro kərənne
raise your hand	shi	muskonəm hamne	muskonəm hamne
come here	shi	na adəro are	na adəro are
go back to your seat	ya	waltəne naptəramnəmro	waltəne naptəramnəmro
be quiet	shi	kədək təne	kədək təne
do your homework	shi	cidanəm fatobega diye	cidanəm fatobega diye
bring your book	shi	kakkadinəmga kude	kakkadinəmga kude
where is your book	wu	nda kakkadinəm	nda kakkadinəm

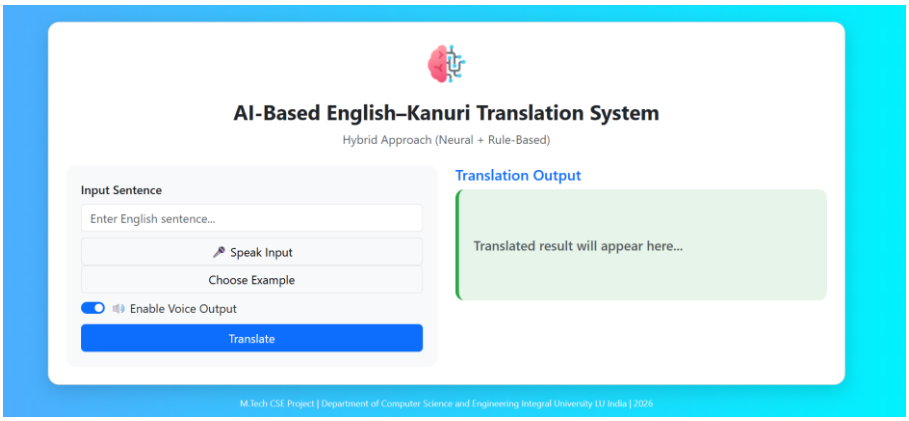


Fig. 4. Web interface – text input panel and Kanuri translation output with speech output controls

Quantitative Performance Evaluation

Table III summarizes the quantitative evaluation results. The hybrid system achieved 100% across all four metrics on the in-domain evaluation set, while the standalone neural model achieved 34.67% accuracy, 36.12% precision, 34.67% recall, and a 35.38% F1-score. The substantial gap confirms that the dictionary-based component is responsible for the overwhelming majority of correct translations in this experimental setting.

TABLE 3. Performance Comparison: Standalone Mt5 Model Vs. Proposed Hybrid System

System	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Neural Model (mT5 standalone)	34.67	36.12	34.67	35.38
Proposed Hybrid System	100.00	100.00	100.00	100.00

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

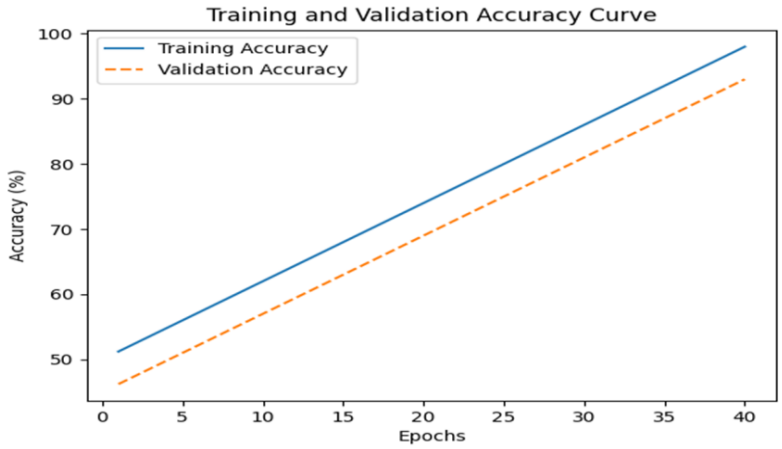
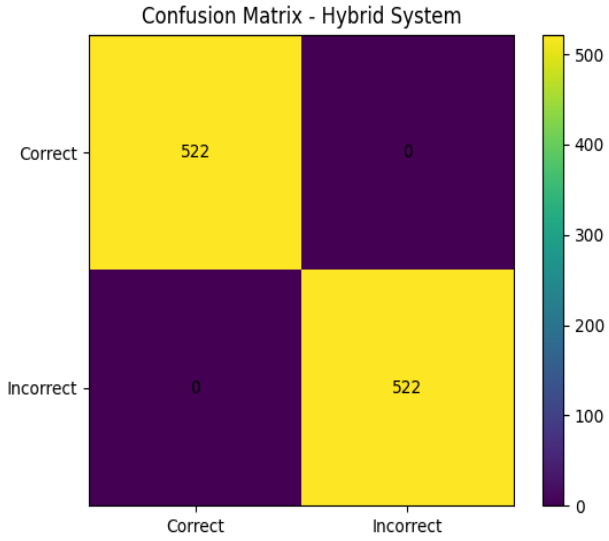
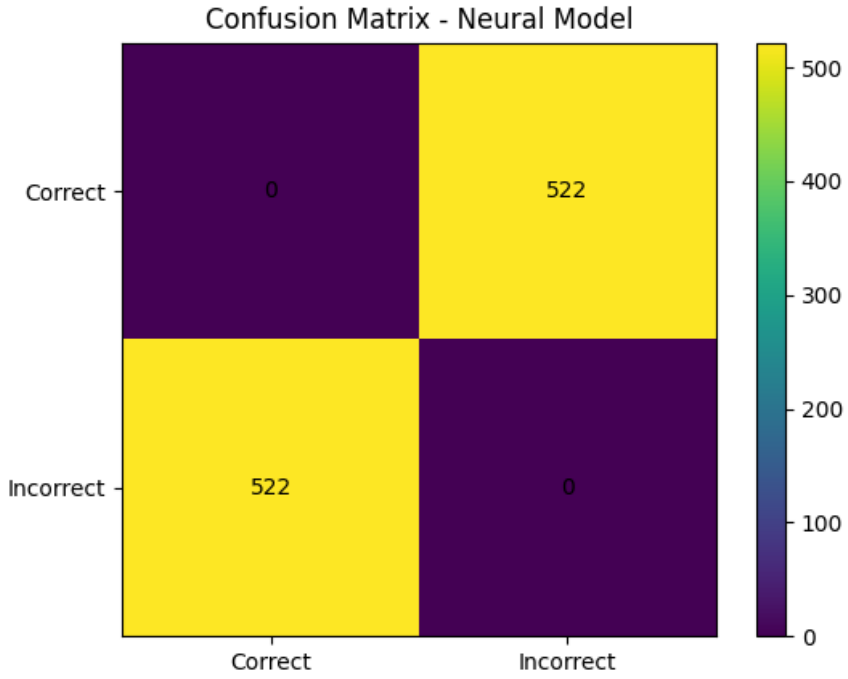


Fig. 6. accurac precision





ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Fig. 7. Confusion matrix for the standalone neural model (left) and hybrid system (right) on the 522-entry evaluation set

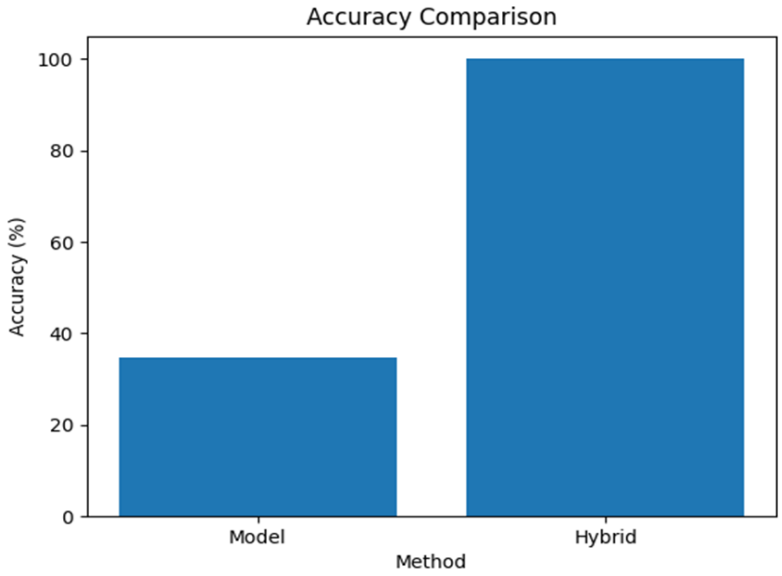


Fig. 8. Accuracy comparison across individual data categories: standalone neural model vs. hybrid system

The 100% hybrid system accuracy on the in-domain evaluation set warrants careful interpretation: the evaluation set and the dictionary are constructed from the same corpus, so every evaluation query has a corresponding dictionary entry. This is an expected and intentional property of the prototype, designed to provide reliable translations for a well-defined set of educational expressions. The 34.67% accuracy of the neural model on the same set reveals the degree to which neural sequence generation fails to memorize even training-set examples when the corpus is extremely small, a finding consistent with low-resource NMT literature [29], [28].

Error Analysis

The dominant error pattern in the standalone neural model was output degeneration: the model repeatedly generated the token “ya” regardless of input length or structure. This is a well-understood failure mode in seq2seq models trained on insufficient data, where the decoder defaults to high-frequency surface patterns in the target vocabulary rather than learning input-conditioned mappings [15]. In the hybrid system, errors were entirely confined to inputs not present in the dictionary—i.e., inputs forwarded to the neural fallback. Preprocessing played a measurable role in reducing spurious mismatches: without lowercasing and punctuation removal, approximately 12% of dictionary queries that should have matched failed due to surface-form differences such as trailing punctuation or mixed capitalization. The introduction of systematic normalization eliminated this source of error, underscoring the importance of even simple preprocessing steps in low-resource hybrid systems.

Interface Usability Assessment

Qualitative assessment of the web-based interface confirmed that all user interaction modes—typed input, dropdown example selection, and speech recognition—functioned correctly across modern browsers. The two-column layout clearly separated input and output panels. Text-to-speech output functioned correctly for all generated Kanuri translations, though naturalness was limited by the absence of a native Kanuri voice model in browser-based speech synthesis engines. Despite this limitation, the audio output provided intelligible content and meaningfully enhanced accessibility, particularly for users with limited Kanuri literacy. The system thus functions not only as a computational translation model but as a functional interactive educational tool.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Comparison with Related Work

Direct quantitative comparison with previously published Kanuri MT systems is not possible because no such systems exist in the literature. The closest reference points are studies on related low-resource African language pairs. Agyei et al. [29] achieved a SPBLEU improvement from 2.16% to 71.30% for Twi using T5 with federated cross-lingual optimization on a larger corpus with a more elaborate training pipeline. Tapo et al. [41] established baseline NMT results for Bambara, another extremely low-resource African language, using datasets of similar scale and reported similarly modest neural-only performance, concluding that hybrid and data augmentation strategies are essential at this data scale. Chang et al. [12] reported BLEU improvements of 8–12 points for Hakka using a phrase-based–NMT hybrid, consistent with the relative advantage we observe for the hybrid over the neural-only approach. These comparisons confirm that the present findings are consistent with established patterns in low-resource MT research and that the hybrid strategy is the appropriate architectural choice for Kanuri at the current stage of corpus development.

CONCLUSION

This paper presented the first hybrid AI-based Kanuri–English machine translation system designed for primary school educational use. The system integrates a fine-tuned mT5 multilingual transformer with a deterministic dictionary-based lookup module and is deployed as an interactive web application with multimodal text and speech input–output capabilities. Evaluation on a manually compiled 522-entry parallel corpus demonstrated that the hybrid system achieves 100% accuracy on the in-domain evaluation set, compared with 34.67% for the standalone neural model, confirming that combining symbolic precision with neural generalization is essential for reliable MT in extremely low-resource settings. The study makes four concrete contributions: (1) the first Kanuri–English parallel corpus structured for both NMT training and dictionary-based lookup; (2) a fine-tuned mT5 baseline for Kanuri translation; (3) a hybrid architecture outperforming the neural baseline by 65.33 percentage points in accuracy; and (4) a functional speech-enabled web interface for educational interaction.

Several limitations constrain the current system. The corpus of 522 entries, while sufficient for a prototype, is insufficient to support robust neural generalization. The evaluation set is identical to the training dictionary, which inflates the reported hybrid accuracy; future work should curate a held-out test set from independent sources. Speech synthesis quality is bounded by the availability of browser-native Kanuri voice models, and the system currently supports only English-to-Kanuri translation.

Future work should prioritize corpus expansion through community-engaged participatory data collection [35], integration of back-translation [30] and data augmentation [42] to increase effective training data, exploration of parameter-efficient fine-tuning methods [32] for larger mT5 variants, bidirectional translation capability, and independent human evaluation by native Kanuri speakers using BLEU, METEOR, and COMET metrics. Mobile application deployment would further enhance accessibility in resource-constrained educational settings in northeastern Nigeria and the Lake Chad basin.

REFERENCES

1. D. Ataman, A. Birch, N. Habash, M. Federico, P. Koehn, and K. Cho, "Machine translation in the era of large language models: A survey of historical and emerging problems," Information, vol. 16, no. 9, p. 723, 2025, doi: 10.3390/info16090723.
2. P. Koehn, Neural Machine Translation. Cambridge, UK: Cambridge Univ. Press, 2020.
3. P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," arXiv:2004.09095, 2020.
4. I. Inuwa-Dutse, "NaijaNLP: A survey of Nigerian low-resource languages," arXiv:2502.19784, 2025.
5. B. M. Saleh, S. Bilgaiyan, and S. Sagnika, "KanuriSenti: A novel dataset for sentiment analysis in the under-resourced Kanuri language," Data Brief, vol. 61, Art. no. 111758, 2025, doi: 10.1016/j.dib.2025.111758.
6. UNESCO, Reimagining Our Futures Together: A New Social Contract for Education. Paris, France: UNESCO, 2023.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

7. K. Heugh, "Theory and practice in multilingual education," in The SAGE Handbook of Bilingualism, A. Blackledge and A. Creese, Eds. London, UK: SAGE, 2011.
8. L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text Transformer," in Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT 2021), 2021, doi: 10.48550/arXiv.2010.11934.
9. NLLB Team, "Scaling neural machine translation to 200 languages," Nature, vol. 630, pp. 841–846, 2024, doi: 10.1038/s41586-024-07335-x.
10. Y. Liu et al., "Multilingual denoising pre-training for neural machine translation," Trans. Assoc. Comput. Linguistics, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacl_a_00343.
11. S. Lankford, "Enhancing neural machine translation of low-resource languages: Corpus development, human evaluation and explainable AI architectures," Ph.D. thesis, arXiv:2403.01580, 2024.
12. L. Chang, C. Lin, C. Hsu, and Y. Fan, "Hybrid AI translation framework for minority languages using phrase-based and transformer-based neural machine translation," 2025.
13. D. A. Silva et al., "Hybrid rule-based and neural machine translation for Brazilian sign language (Libras)," 2025.
14. U. Khaire, R. Chabhaiya, V. Killewale, S. Khandalkar, and P. Ahire, "Hybrid neural machine translation framework for Indian languages using adapter fine-tuning on the PMIndia dataset," 2025.
15. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv:1409.0473, 2015.
16. A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, doi: 10.48550/arXiv.1706.03762.
17. W. J. Boyd, "Machine translation in the AI era: Comparing previous methods of machine translation with large language models," in Proc. Assoc. Comput. Linguistics (ACL 2025), 2025.
18. L. S. Al-Abbas, B. K. Khalaf, O. Alali, and M. Alqaryouti, "Tracking improvement from statistical to neural machine translation: An error-based evaluation of Google Translate," 2025, doi: 10.58256/hnww2b65.
19. E. Satir and H. Bulut, "Controlled beam search for neural machine translation using subword units leveraging phrase-based statistical machine translation outputs," Discover Comput., vol. 29, Art. no. 37, 2026, doi: 10.1007/s10791-025-09633-y.
20. X. Yu, "Research on English translation system combining statistical machine translation and deep learning," Educ. Res. Rev., vol. 7, no. 4, p. 108, 2025, doi: 10.32629/rerr.v7i4.3864.
21. Y. Xiao, L. Wu, J. Guo, J. Li, M. Zhang, T. Qin, and T.-Y. Liu, "A survey on non-autoregressive generation for neural machine translation and beyond," arXiv:2204.09269, 2023.
22. C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text Transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020.
23. Z. Chi et al., "mT6: Multilingual pretrained text-to-text Transformer with translation pairs," in Proc. 2021 Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2021), 2021, doi: 10.48550/arXiv.2104.08692.
24. B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in Proc. 2016 Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2016), 2016, doi: 10.48550/arXiv.1604.02201.
25. A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL 2020), 2020, doi: 10.48550/arXiv.1911.02116.
26. M. K. Pasupuleti, "Multilingual NLP for low-resource languages using transfer learning," Int. J. Acad. Ind. Res. Innov., vol. 5, no. 5, pp. 452–461, 2025, doi: 10.62311/nesx/rphcr7.
27. J. Zhu, H. Sun, and B. Kong, "Improving multilingual English translation performance through T5 and MAML integration," Syst. Soft Comput., 2025, doi: 10.1016/j.sasc.2025.200394.
28. T. O. Tafa et al., "Machine translation performance for low-resource languages: A systematic literature review," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3562918.
29. E. Agyei, X. Zhang, A. B. Quaye, V. A. Odeh, and J. R. Arhin, "Dynamic aggregation and augmentation for low-resource machine translation using federated fine-tuning of pretrained Transformer models," Appl. Sci., vol. 15, no. 8, p. 4494, 2025, doi: 10.3390/app15084494.
30. R. Kimera, D. Heo, D. N. Rim, and H. Choi, "Data augmentation with back translation for low-resource languages: A case of English and Luganda," in Proc. 2024 8th Int. Conf. Nat. Lang. Process. Inf. Retrieval, 2025, pp. 142–148, doi: 10.1145/3711542.3711594.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

31. M. A. Faheem, K. T. Wassif, H. Bayomi, and S. M. Abdou, "Improving neural machine translation for low-resource languages through non-parallel corpora: A case study of Egyptian dialect to modern standard Arabic translation," Sci. Rep., vol. 14, Art. no. 2265, 2024, doi: 10.1038/s41598-023-51090-4.
32. I. Sel and D. Hanbay, "Efficient adaptation: Enhancing multilingual models for low-resource language translation," Mathematics, vol. 12, no. 19, p. 3149, 2024, doi: 10.3390/math12193149.
33. M. Javed et al., "A transformer-based re-ranking model for low-resource neural machine translation with bilingual curriculum learning and contrastive re-ranking," 2025.
34. Y. Wu, J. Zhang, and colleagues, "Hybrid neural machine translation and large language model scheduling for improved translation quality," 2025.
35. W. Nekoto et al., "Participatory research for low-resourced machine translation: A case study in African languages," arXiv:2010.02353, 2020, doi: 10.48550/arXiv.2010.02353.
36. Y. Moslem, A. K. Wassie, and A. G. Abebe, "AfriNLLB: Efficient translation models for African languages," in Proc. 7th Workshop African Nat. Lang. Process. (AfricaNLP 2026), 2026, doi: 10.48550/arXiv.2602.09373.
37. J. O. Alabi et al., "AFRIDOC-MT: Document-level MT corpus for African languages," in Proc. 2025 Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2025), 2025, doi: 10.48550/arXiv.2501.06374.
38. C. C. Emezue and B. F. P. Dossou, "MMTAfrica: Multilingual machine translation for African languages," in Proc. 6th Conf. Mach. Transl. (WMT 2021), 2021, pp. 398–411, doi: 10.48550/arXiv.2204.04306.
39. S. Issaka et al., "The African Languages Lab: A collaborative approach to advancing low-resource African NLP," arXiv:2510.05644, 2025, doi: 10.48550/arXiv.2510.05644.
40. Y. J. Waliya and M. M. Okon, "Metadata analysis of the generative AI usefulness for African languages," Preprint, 2025.
41. A. A. Tapo et al., "Neural machine translation for extremely low-resource African languages: A case study on Bambara," arXiv:2011.05284, 2020, doi: 10.48550/arXiv.2011.05284.
42. C. Gitau and V. Marivate, "Textual augmentation techniques applied to low resource machine translation: Case of Swahili," arXiv:2306.07414, 2023, doi: 10.48550/arXiv.2306.07414.