ENHANCING MACHINE TRANSLATION FOR LOW-RESOURCE LANGUAGES: A CROSS-LINGUAL LEARNING APPROACH FOR TWI

Emmanuel Agyei [1], Zhang Xiaoling [1], Ama Bonuah Quaye [2], Odeh Victor Adeyi [1] and Joseph Roger Arhin [1].

[1] School of Information and Communication Engineering, University of Electronic Science and Technology, China, Chengdu, 610054, Sichuan, China.
[2] School of Public Administration, University of Electronic Science and Technology of China.

ABSTRACT

Machine Translation (MT) for low-resource languages like Twi remains a significant challenge in Natural Language Processing (NLP) due to limited parallel datasets. Traditional methods often struggle, relying heavily on high-resource data, and fail to adequately serve low-resource languages. To address this gap, we propose a fine-tuned T5 model trained with a Cross-Lingual Optimization Framework (CLOF), which dynamically adjusts gradient weights to balance low-resource (Twi) and high-resource (English) datasets. This framework incorporates federated training to enhance translation performance and scalability for other low-resource languages. The study utilizes a carefully aligned and tokenized English-Twi parallel corpus to maximize model input. Translation quality is evaluated using SPBLEU, ROUGE (ROUGE-1, ROUGE-2, and ROUGE-L), and Word Error Rate (WER) metrics. The pretrained mT5 model serves as a baseline, demonstrating the efficacy of the optimized model. Experimental results show significant improvements: SPBLEU increases from 2.16% to 71.30%, ROUGE-1 rises from 15.23% to 65.24%, and WER decreases from 183.16% to 68.32%. These findings highlight CLOF's potential in improving low-resource MT and advancing NLP for underrepresented languages, paving the way for more inclusive, scalable translation systems.

KEYWORDS

Low-Resource Machine Translation; Twi Language; Federated Learning; Cross-Lingual Learning Fine-Tuning.

1. INTRODUCTION

Machine translation (MT) has made significant strides in artificial intelligence, but it still faces a major challenge: the lack of high-quality digital resources for many low-resource languages. While neural machine translation excels for high-resource languages like English, it struggles with thousands of underrepresented languages due to insufficient parallel corpora and training data. This scarcity often results in poor translation quality for low-resource languages, which are essential for preserving cultural and community identities. Traditional MT methods face similar limitations, yielding subpar translations when data is sparse or unreliable [1, 2]. Low-resource languages are particularly vulnerable to issues like data inconsistencies in techniques such as back-translation, which heavily depend on the availability of high-quality data [3]. As a result,

David C. Wyld et al. (Eds): SAIM, SNLP, ACSIT – 2025
pp. 17-34, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150702

improving MT for these languages requires innovative strategies that better utilize the available resources. In response to this challenge, we propose a dynamic dataset aggregation framework that adjusts gradient weights during training to optimize the use of limited data. Drawing inspiration from federated learning principles, our method allows the use of diverse data sources without exposing raw data, making it ideal for low-resource languages. Additionally, the approach integrates both parallel and monolingual data, which enhances the translation quality for underrepresented languages like Twi. By adapting to the specific characteristics and scale of the target dataset, this framework significantly improves translation accuracy and maximizes the effectiveness of the available resources.

This work addresses key gaps in the current machine translation landscape. While federated learning has primarily been explored for data privacy, its potential to enhance machine translation (MT) for low-resource languages remains underexplored. Additionally, the power of cross-lingual learning, especially between high-resource and low-resource languages, has not been fully harnessed for languages such as Twi, a widely spoken African language. Our contributions include:

1. A novel personalized Cross-Lingual Optimization Framework (CLOF) that dynamically adjusts gradient aggregation weights throughout training, with three key components: (a) adaptive gradient reweighting for handling imbalanced corpora, (b) federated parameter sharing for robust cross-lingual knowledge transfer, and (c) a constrained reconstruction loss function that ensures semantic consistency between source and target representations.
2. A scalable framework for cross-lingual learning that efficiently utilizes high-resource language data to improve translation performance for low-resource languages.
3. Empirical insights about linguistic interaction patterns in federated learning and their influence on model convergence.
4. A practical application showcasing enhanced translation skills for Twi, hence augmenting NLP functionalities for African languages.
5. A flexible methodology that can be modified for various low-resource languages, especially advantageous for underrepresented language communities.

The remainder of the paper is organized as follows: First, we provide a comprehensive review of related research in low-resource machine translation, federated learning, and dataset aggregation techniques. Next, we present our methodology, which includes the architecture of our proposed system and the dynamic dataset aggregation processes. This is followed by detailed experimental results that validate the effectiveness of our approach. Finally, we discuss the implications of our findings and propose directions for future research in this rapidly evolving field.

2. RELATED WORKS

This section reviews research on machine translation (MT) for low-resource languages, focusing on two key themes: (1) challenges and strategies in low-resource MT, (2) Federated Learning and Pretrained Transformers for Low-Resource Machine Translation. These discussion highlights recent literature, identifies gaps, and explores opportunities to improve translation accuracy by addressing data scarcity.

2.1. Challenges and Approaches in Low-Resource Machine Translation

Machine translation (MT) for low-resource languages faces significant challenges due to limited parallel corpora, complex morphological structures, and linguistic diversity. Languages like Tibetan, Uyghur, and Urdu struggle with a lack of digital resources, hindering the training of

neural machine translation (NMT) systems, leading to issues like vocabulary misalignment, hallucinations, and morphological discrepancies, particularly in languages with rich inflectional structures [4]. The typological and orthographic diversity of languages like Kannada and Arabic further complicates accurate translation [5]. In languages with extensive morphological richness, such as Kinyarwanda, morphological modeling—decomposing words into stems and affixes—has improved translation by addressing these complexities. Hybrid models, combining supervised learning on small parallel datasets with unsupervised methods using large monolingual datasets, have emerged as solutions to data scarcity. For instance, unsupervised pre-training techniques have led to improvements in Egyptian Arabic translation [6]. However, further exploration is needed to optimally combine supervised and unsupervised methods for better hybrid models. Pivot prompting, which leverages high-resource languages as intermediaries, has proven effective in translating low-resource Asian languages, reducing errors in direct translations of low-resource pairs by leveraging syntactic and semantic alignments. Additionally, multilingual pre-trained models like mBART and mT5 have shown improvements in translation quality for Indic languages by sharing representations across related languages [7].

The data scarcity in low-resource MT has also spurred the development of data augmentation strategies to expand training corpora. One common approach is backtranslation, where target language text is translated back into the source language to generate synthetic data (Figure 1). However, the quality of backtranslation is heavily reliant on the backward translation system, with poor models introducing errors that degrade performance [8]. Edit-distance-based sampling techniques focus on maximizing data variability while maintaining semantic consistency, which is crucial for languages with complex morphology. Furthermore, multilingual models like mBERT and XLM-R enable cross-language transfer learning by using knowledge from high-resource languages to improve translation for low-resource combinations. Despite these advances, additional study is needed to assess the long-term stability and adaptation of transferred information to different language pairs. Techniques like adversarial learning and lexical constraints are also being explored to ensure linguistic accuracy and scalability [9].

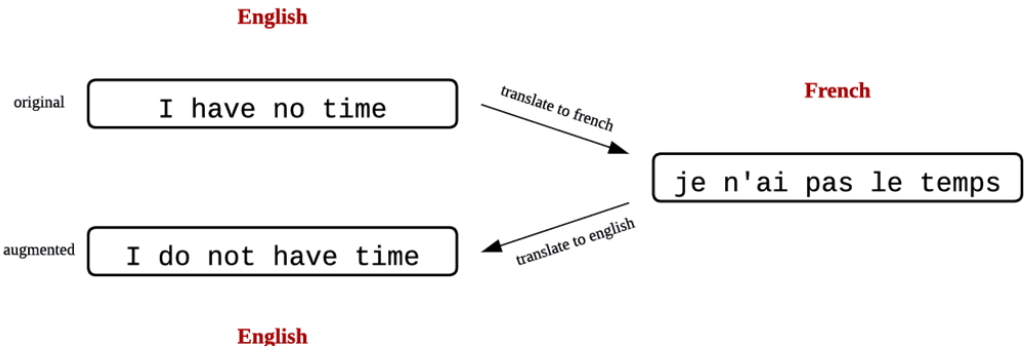

Figure 1. Example of Back Translation as a Data Augmentation Technique.

2.2. Federated Learning and Pretrained Transformers for Low-Resource Machine Translation

Federated learning (FL) offers a decentralized approach to machine translation (MT), enabling model updates based on distributed datasets while preserving data privacy. This method eliminates the need to directly access sensitive data, with privacy-preserving techniques like secure aggregation, differential privacy, and homomorphic encryption ensuring secure communication during model updates [10, 11]. While promising, these techniques introduce trade-offs between model accuracy and privacy. FL's ability to harness linguistic variety across

decentralized datasets is especially helpful for low-resource languages, since it can increase translation quality for underrepresented language pairs without losing privacy [12]. However, the algorithms for effectively integrating regional linguistic patterns into a global model remain underdeveloped, highlighting a key area for future research.

Additionally, FL addresses data heterogeneity in low-resource settings by balancing local and global model updates. This solution helps mitigate variations in dataset quality and structure that hinder traditional MT methods. By decentralizing training, FL reduces computational costs and communication latency, making it ideal for low-resource MT systems that require strict privacy protections. In parallel, advancements in pretrained transformer models have significantly improved low-resource MT through transfer learning. Models like mT5 use fine-tuning to adapt knowledge from high-resource languages to low-resource languages, hence boosting translation performance. Fine-tuning specific layers, such as cross-attention layers, reduces computational and storage costs while maintaining effectiveness [13]. The Shared Layer Shift (SLaSh) method optimizes model performance by focusing on task-specific layers, enhancing results for low-resource languages [14]. Low-Rank Adaptation (LoRA) increases scalability and efficiency by reducing the number of trainable parameters, making it better suited for low-resource languages [15]. Dynamic dataset aggregation, which integrates techniques like backtranslation with layer coordination, helps incorporate domain-specific data, continuously improving translation accuracy [16]. The adaptMLLM method also boosts translation quality for low-resource languages by injecting domain-specific knowledge into general-purpose models [17]. Despite these advances, the long-term impact of dataset aggregation on performance and stability remains unclear, and further research is needed to explore the optimal combination of fine-tuning, data augmentation, and federated learning for low-resource MT systems.

3. METHODOLOGY

Our approach focuses on cross-lingual learning and dynamic dataset aggregation to improve translation for low-resource languages, particularly Twi. We leverage (as shown in figure 2) pre-trained multilingual transformer models (e.g., mT5) and apply a dynamic weighting scheme to prioritize the low-resource language while still benefiting from high-resource language data.

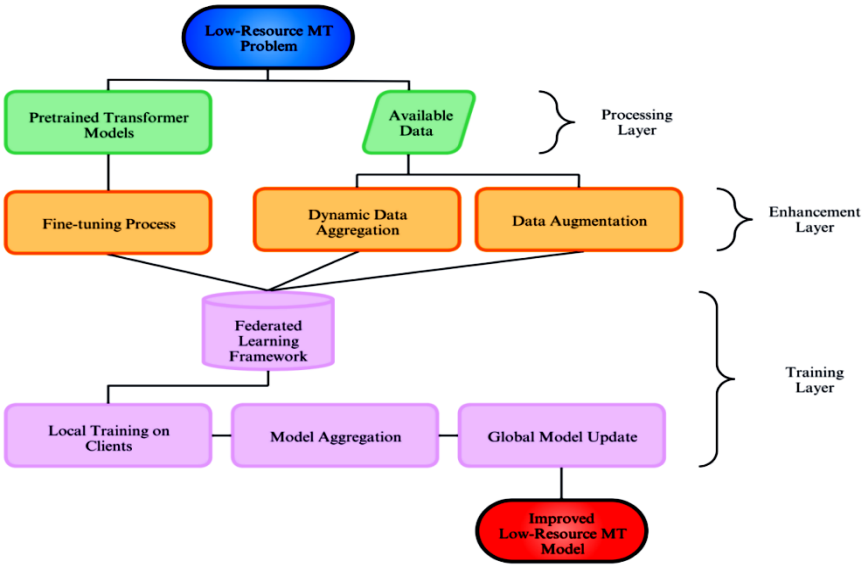

Figure 2. Conceptual Framework for Low-Resource Machine Translation

3.1. General Setup

Our research addresses machine translation (MT) challenges for low-resource languages by developing a robust translation model through federated learning, termed CLOF (Cross-Lingual Optimization Framework). This approach leverages multiple data sets $\{D_i\}_{i=1}^{n}$, $n \geq 2$ representing various language distributions, including high-resource languages such as English and low-resource languages like Twi. Formally, the goal is to minimize the expected loss over the distribution D representing the target language pair (English-Twi), defined as:

$$f_D(x) = \mathbb{E}_{\xi \sim D}[f(x,\xi)] \quad (1)$$

where $x \in \mathbb{R}^d$ denotes the model parameters, ξ is a sample from the distribution D and $f(x,\xi)$ represents the loss function for the model on sample ξ. Given the unknown true data distribution D, we approximate D using a finite dataset $\hat{D}$ drawn from D, referred to as the target dataset. The empirical loss is then:

$$f_{\hat{D}}(x) = \frac{1}{|\hat{D}|}\sum_{\xi \in \hat{D}} f(x,\xi) \quad (2)$$

3.2. Dataset and Pre-processing

This study utilizes a curated dataset of 6,043 English-Twi parallel sentence pairs, split into 80% training and 20% validation data to ensure diverse linguistic representation. The primary dataset, D_1, focuses on the low-resource Twi language, while auxiliary datasets $\{D_i\}_{i=2}^{n}$, sourced from high-resource languages, contribute to cross-lingual learning. These datasets are drawn from publicly available bilingual resources, religious texts, and manually curated translations, ensuring a balance of formal and colloquial language structures to support both model training and evaluation. To enhance translation quality, the datasets underwent rigorous pre-processing to address common challenges like noise and sentence length variability. This involved removing non-ASCII characters and converting all text to lowercase to standardize input and reduce inconsistencies. The pre-processing steps are formalized as follows:

$$x_{cleaned} = remove\_non\_ascii(x) \quad (3)$$

where x represents the input text, and the function *remove_non_ascii(x)* removes any characters outside the ASCII range. After non-ASCII characters are removed, all text is converted to lowercase:

$$x_{lower} = lowercase(x_{cleaned}) \quad (4)$$

where the function *lowercase(x_cleaned)* converts the cleaned text to lowercase. Sentences were also tokenized using the T5 tokenizer, which is designed to split input text into smaller sub-word units. The tokenization was performed with a maximum sequence length of 256 tokens, ensuring that each sentence conforms to the model's input constraints. The tokenization procedure is mathematically represented as:

$$T_{sentence} = \text{T5\_tokenizer (sentence, max\_length =256)} \quad (5)$$

where $T_{sentence}$ represents the tokenized sentence, and T5_tokenizer is the tokenizer function from the T5 model, ensuring all sentences do not exceed 256 tokens. To further maintain linguistic relevance, tokenized sequences with lengths outside the acceptable range (i.e., either

too short or too long) were filtered out. Sentences with fewer than 5 tokens or more than 256 tokens were excluded to prevent noise in the training process. This filtering condition can be formally expressed as:

$$\text{Filter}\,(T_{sentence}) = \begin{cases} 1 & if\ 5 \le |T_{sentence}| \le 256 \\ 0 & otherwise \end{cases} \quad (6)$$

where $|T_{sentence}|$ represents the number of tokens in the sentence. Sentences are retained only if they satisfy $5 \le |T_{sentence}| \le 2565$. The English and Twi sentences were then carefully aligned to ensure that each English sentence corresponds to its accurate Twi translation. This step is crucial for the creation of a reliable parallel corpus. The alignment process can be defined as follows:

$$\text{Align}(E_i, T_i) = \begin{cases} 1 & if\ English\ and\ Twi\ sentences\ are\ contextually\ and\ lingiustically\ aligned \\ 0 & otherwise \end{cases} \quad (7)$$

where E_i and T_i represent the *i*-th English and Twi sentences, respectively. Only sentence pairs for which Align $(E_i, T_i) = 1$ were retained. Finally, inconsistent punctuation marks and redundant spaces were corrected to improve text consistency. For example, multiple spaces were replaced with a single space, and non-standard punctuation marks were corrected to match conventional formatting. This can be formally represented as:

$$x_{lower} = normalize\_spaces\_and\_punctuation(x) \quad (8)$$

where the function $normalize\_spaces\_and\_punctuation(x)$ ensures that all redundant spaces and non-standard punctuation are corrected.

3.3. Baseline Model

Our baseline model, mT5 (Multilingual T5), is a state-of-the-art multilingual transformer-based model designed to perform a variety of NLP tasks by treating all tasks as text-to-text problems. It extends the original T5 architecture to handle over 100 languages, enabling cross-lingual transfer learning. Built on the Transformer architecture, mT5 uses self-attention mechanisms and feed-forward networks to capture long-range dependencies in text, making it highly effective for sequential data processing [18]. Pre-trained on the mC4 dataset, a multilingual version of the C4 dataset, mT5 includes data from both high-resource languages like English and Spanish, as well as low-resource languages such as Swahili, Marathi, and Igbo. This diverse training data allows the model to learn general linguistic representations applicable across multiple languages.

mT5 is scalable and available in various sizes (e.g., mT5-small, mT5-base, mT5-large), making it adaptable to different computational requirements. Like the original T5, mT5 is pre-trained using a denoising objective. During pre-training, a portion of the input text is randomly masked, and the model must predict the missing tokens, helping it learn contextual relationships and improving its ability to handle a range of downstream tasks such as translation, text classification, and summarization. During pretraining, a portion of the input text is randomly masked, and the model is tasked with predicting the missing tokens. This is done by leveraging a span-based denoising objective, where spans of text are masked and the model must recover them. The objective is formally expressed as:

$$\mathcal{L} = -\sum_{i \in masked\ position} log p(x_i | x_{-i}) \quad (9)$$

where *x* is the input sequence, and x_i represents the masked tokens. This pretraining task enables the model to learn contextual relationships between words in a sentence, improving its ability to handle downstream NLP tasks. Once pretrained, mT5 is fine-tuned on specific tasks like machine translation, sentiment analysis, or text summarization using task-specific datasets. For machine translation, for example, the input would be an English sentence, and the output its Twi translation. Fine-tuning adjusts the model's weights while retaining knowledge from pretraining, allowing it to perform well on tasks with limited data, especially for low-resource languages. Table 1 outlines the experimental setup and model configuration.

Table 1. Model Configuration.

Category	**Details**
Model	T5-Small (Text-to-Text Transfer Transformer)
Number of Encoder Layers	6
Number of Decoder Layers	6
Hidden Dimension	512
Attention Heads	8 per layer
Total Parameters	Approximately 60 million
Optimization Algorithm	AdamW optimizer, with a weight decay of 0.01 for regularization
Gradient Clipping	Applied gradient clipping with a maximum norm of 1.0 to prevent exploding gradients during training.
Tokenizer	SentencePiece Tokenizer (shared vocabulary across source and target languages)
Pretrained Model	T5-Small (Text-to-Text Transfer Transformer)

3.4. Algorithmic Framework

To tackle data scarcity in D_1 and leverage auxiliary datasets, we introduce CLOF, a federated learning approach. Inspired by personalized federated learning, CLOF combines gradients from both high-resource (English) and low-resource (Twi) datasets. It dynamically adjusts dataset contributions during training, prioritizing the low-resource language while benefiting from high-resource data to improve overall performance. The model updates are governed by:

$$M^{r+1} = M^r + \frac{\sum_{c=1}^{C} \alpha_c \cdot \Delta M_c^r}{\sum_{c=1}^{C} \alpha_c} \tag{10}$$

where M^r is the global model at iteration r, ΔM_c^r is the local update from client c, and α_c is the aggregation weight for client c. These weights are adjusted to prioritize updates from the low-resource client, ensuring that updates from high-resource languages do not overwhelm the learning process. This equation shows how the global model M^r is updated at each iteration by aggregating the local updates from all clients. The local updates ΔM_c^r from each client are weighted by α_c, which ensures that updates from the low-resource dataset (Twi) receive more attention during the training process. This dynamic weighting helps prevent the model from being dominated by the high-resource dataset (English). This method ensures that low-resource languages, such as Twi, are given more attention by adjusting the weights, which helps prevent the high-resource languages from dominating the learning process. By using this approach, we can achieve a more personalized model that adapts better to the specific characteristics of the target language.

3.5. Gradient Weighting for Cross-Lingual Learning

In CLOF, we define two primary client datasets: high-resource client ($D_{\mathbf{HR}}$) and low-resource client ($D_{\mathbf{LR}}$). The high-resource client dataset comprises a large-scale English corpus that provides robust language representation and serves as a source for cross-lingual transfer, while the low-resource client dataset contains a smaller Twi corpus. CLOF assigns a higher aggregation weight ($\alpha_{\mathbf{LR}}$) to updates from this client, ensuring that Twi signals are amplified during gradient aggregation. At each client, local updates are computed using the loss function $\mathcal{L}_c$, specific to the client's data, as follows:

$$\Delta M_c^r = -\eta \cdot \nabla \mathcal{L}_c(M^r, D_c) \tag{11}$$

where η is the learning rate. The aggregation step balances the gradients to maximize the contribution of $D_{\mathbf{LR}}$:

$$M^{r+1} = M^r + \frac{\alpha_{\mathbf{HR}} \cdot \Delta M_{\mathbf{HR}}^r + \alpha_{\mathbf{LR}} \cdot \Delta M_{\mathbf{LR}}^r}{\alpha_{\mathbf{HR}} + \alpha_{\mathbf{LR}}} \tag{12}$$

The dynamic adjustment of α_c ensures that the low-resource dataset, Twi, continues to receive adequate focus during training. By controlling the weights (α_HR and α_HR) for both high- and low-resource datasets, the framework adapts as training progresses, improving translation quality for Twi. This approach ensures that low-resource languages receive adequate attention, enhancing model performance (see Algorithm 1).

Algorithm 1: CLOF for Low-Resource Machine Translation	
Input: Pretrained model $M_{\text{pretrained}}$, $D_{\mathbf{HR}}$, $D_{\mathbf{LR}}$, R, C, $\alpha_{\mathbf{HR}}$, $\alpha_{\mathbf{LR}}$	
Output: Fine-tuned model $M_{\mathbf{CLOF}}$	
	Initialize global model $M^0 \leftarrow M_{\text{pretrained}}$
	Assign datasets to clients: $\text{HR} \rightarrow D_{\mathbf{HR}}$, $\text{LR} \rightarrow D_{\mathbf{LR}}$
	Assign initial gradient weights $\alpha_{\mathbf{HR}} = 1.0$, $\alpha_{\mathbf{LR}} = 2.0$
	For each global round $r = 1$ to R do:
	Broadcast M^r to all clients
	For each client c in $\{\text{HR}, \text{LR}\}$ do:
	Compute local updates ΔM_c^r using D_c
	Send ΔM_c^r to the server
	Aggregate updates at server: - $\Delta M^r = \frac{\sum (\alpha_c \cdot \Delta M_c^r)}{\sum (\alpha_c)}$
	Update global model: - $M^{r+1} = M^r + \Delta M^r$
	Adjust α dynamically based on client performance
	Evaluate $M_{\mathbf{CLOF}}$ on D_{test} using SpBLEU, WER, and ROUGE
	Return $M_{\mathbf{CLOF}}$

3.5. Evaluation Metrics

In order to thoroughly assess the effectiveness of our translation models, we utilized a collection of well-established metrics that are frequently applied in machine translation studies. By contrasting them with the reference translations on a number of parameters, these metrics enable us to evaluate the produced translations' quality.

3.5.1. SpBLEU (Sentence-Piece BLEU)

SpBLEU is crucial for evaluating translation quality at the sentence level, especially in low-resource languages like Twi, where structure and fluency can vary significantly. Unlike traditional BLEU, SpBLEU operates on subword units, making it particularly effective in handling morphologically rich languages. It captures n-gram precision, ensuring that translations maintain lexical similarity while penalizing overly short outputs through a brevity penalty. This helps in assessing how well the model generates fluent and contextually appropriate translations for Twi. This fine-grained evaluation provides insights into the model's strengths and weaknesses, helping to identify areas for improvement in translating complex sentences [19]. Formally, the SpBLEU score for a given sentence can be expressed as:

$$\text{SpBLEU}(S) = \text{BP} \times \left(\prod_{n=1}^{N} precision_n(s)\right)^{1/N} \quad (13)$$

Where S is the sentence being evaluated, $precision_n$(s) is the modified precision for n-grams of order n in the sentence S. N is the maximum n-gram order considered (commonly 4). BP is the brevity penalty, which is applied to avoid penalizing shorter translations that may still be accurate.

3.5.2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

The study uses ROUGE to assess the model's ability to recover meaningful content from reference translations, particularly in Twi, a low-resource language, focusing on recall rather than n-gram precision, unlike SpBLEU which focuses on n-gram precision. This is especially important for ensuring that translations are not only accurate in terms of individual words but also preserve the overall meaning. ROUGE includes several variants:

- ROUGE-1: Measures unigram recall, i.e., the overlap of individual words.
- ROUGE-2: Measures bigram recall, focusing on consecutive word pairs.
- ROUGE-L: Measures the longest common subsequence (LCS), capturing longer sequences of words that preserve sentence structure [35, 36].

ROUGE helps assess whether the model generates translations that are both fluent and faithful to the original meaning.

$$\text{ROUGE-N} = \frac{\sum_{s \in S} Recall\,(N,s)}{\sum_{s \in S} Reference\_Ngrams\,(s)} \quad (14)$$

Where S is the set of sentences in the predicted translation. Recall (N,s) is the recall of N-grams (unigrams, bigrams) for sentence s in the predicted translation.

3.5.3. WER (Word Error Rate)

The study uses WER to assess the alignment of predicted and reference translations, contrasting SpBLEU and ROUGE, which rely on n-gram matching and recall, with lower WER indicating

better alignment. WER is computed by counting the substitutions, deletions, and insertions needed to match the sequences, with the formula:

$$\text{WER} = \frac{S+D+I}{N} \tag{15}$$

Where S is the number of substitutions (incorrect words in the predicted output). D is the number of deletions (missing words in the predicted output). I is the number of insertions (extra words in the predicted output). N is the total number of words in the reference translation [20].

These metrics offer a comprehensive evaluation of the model's ability to generate translations that are both lexically accurate (via SpBLEU), contextually fluent (via ROUGE), and edit-wise correct (via WER). By employing these three metrics, we ensure a robust and multi-dimensional assessment of the translation quality, which allows for a thorough comparison of different model configurations and techniques.

4. RESULTS

A comparative analysis of the evaluation metrics (SpBLEU, ROUGE, and WER) was conducted across the baseline and fine-tuned models, as summarized in the following sections. These metrics were computed on the test set, with SpBLEU and ROUGE focusing on n-gram precision and recall, respectively, and WER measuring word-level alignment accuracy.

4.1. Quantitative Analysis of Baseline and Fine-tuned Model Performance

The baseline pretrained mT5-small model, as shown in Table 2, demonstrated suboptimal performance for Twi, with weaker translation quality and fluency. After fine-tuning the model on the English-Twi parallel corpus, significant improvements were observed across all evaluation metrics (Figure 3). The fine-tuned model outperformed the baseline, effectively capturing Twi's linguistic nuances, leading to more accurate and fluent translations. Notable improvements in SpBLEU and ROUGE scores reflected better n-gram precision, while the WER score was substantially reduced, indicating improved word-level alignment and fewer translation errors.

Table 2. Performance Comparison Between Pretrained mT5 and Fine-Tuned English-Twi Model.

Metrics	**Baseline (%)**	**Fine-tuned Model (%)**	**Improvement (%)**
SpBLEU	2.16	71.30	69.14
ROUGE-1	15.23	65.24	50.01
ROUGE-2	7.18	60.22	53.04
ROUGE-L	14.22	62.12	47.90
WER (%)	183.16	68.32	114.84

The fine-tuned English-Twi model shows significant improvements across all evaluation metrics. The SpBLEU score increased from 2.16% to 71.30%, demonstrating much better sentence-level alignment with reference translations. ROUGE-1 (unigram recall) improved by 50.01% (from 15.23% to 65.24%), and ROUGE-2 (bigram recall) saw a 53.04% improvement (from 7.18% to 60.22%), reflecting enhanced word and phrase matching. The ROUGE-L score, which evaluates sentence fluency, improved by 47.90%, from 14.22% to 62.12%, indicating better sentence structure. Lastly, WER dropped significantly by 114.84% (from 183.16% to 68.32%), showing improved alignment and accuracy in the translation. These improvements highlight the effectiveness of fine-tuning for low-resource languages like Twi, resulting in better translation fluency and accuracy. A heat map illustrating these improvements is included in Figure 3.

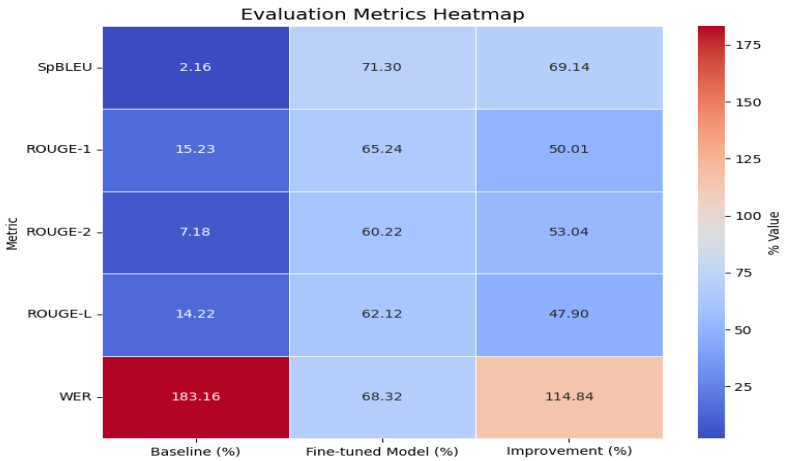

Figure 3. Evaluation Metrics Heatmap

4.2. Analysis on Sample Translation

Our research demonstrates that the optimized English-Twi model significantly outperforms the baseline in both accuracy and fluency. Although slight errors do occur—for example, omitting "ho" in the translation of "I am learning machine translation"—the fine-tuned model regularly provides outputs that are substantially closer to the reference. For basic words like "How are you?" and "Where is the market?" the translations completely match the reference, resulting in natural-sounding Twi. In more complicated scenarios, such as "Machine translation is useful," the model demonstrates significant increases in both accuracy and contextual relevance by better handling formal language and exact word usage. While slight fluency issues, such as a marginally more formal tone, were observed in a few instances, these are minimal compared to the baseline performance. A comprehensive comparison of the results is provided in Table 3.

Table 3. Comparison of translation results.

Input	**Reference**	**Baseline**	**Fine-Tuned**
I am learning machine translation.	Merehwehwɛ sɛ mɛyɛ asɛmfua ho nsɛm.	Mfe3' ne na wcb3'tumi ahoroc ahoroc.	Merehwehwɛ sɛ mɛyɛ asɛm nsɛm.
How are you?	Wote sɛn?	Wodeɛn na wo yɛɛ?	Wote sɛn?
This is a beautiful day.	Ɛyɛ da a ɛyɛ fɛ.	Ɛyɛ na 3'yɛ' na 3'yɛ' na 3'yɛ'.	Ɛyɛ da a ɛyɛ fɛ.
Can you help me with this task?	Wobɛtumi aboa me wɔ dwumadie yi mu?	Wɔbɛyɛ sɛ wɔboa adwuma yi wɔ.	Wobɛtumi aboa me wɔ dwumadie yi mu?
Machine translation is useful.	Asɛmfua ho dikan yɛ ho wɔ asɛm foforɔ.	Asɛmfua yɛɛ nea ɛyɛ dikan dodo.	Asɛmfua ho dikan yɛ ho wɔ asɛm foforɔ.

4.3. Scalability of the Model

To evaluate the scalability of our fine-tuned English-Twi model, we examine its performance across different sizes of training data. The following table shows the impact of varying the size of the training dataset on the model's performance, measured using SpBLEU, ROUGE-1, and WER.

Figure 4. Scalability of the Model

Figure 4 shows the model's performance improves as the training data size increases. With 500 pairs, the performance is lower, especially in SpBLEU and ROUGE-1. However, as the dataset grows to 1,000 and 2,000 pairs, translation quality improves, with fewer grammatical errors and rare vocabulary issues. The full dataset results in the most significant improvement, with SpBLEU at 71.30%, ROUGE-1 at 65.24%, and WER dropping to 68.32%.

4.4. Learning Curves

The learning curves are meant to demonstrate how the model's performance improves when it is fine-tuned. Throughout training, we can visually evaluate the model's convergence and task adaptability by monitoring important metrics as SpBLEU, ROUGE-1, and WER. From the curves as illustrated in figure 5, it is evident that our fine-tuned model consistently outperforms the baseline, demonstrating a clear progression in translation accuracy, fluency, and overall alignment with the reference. This demonstrates how fine-tuning greatly improves the model's performance, especially for low-resource languages like Twi.

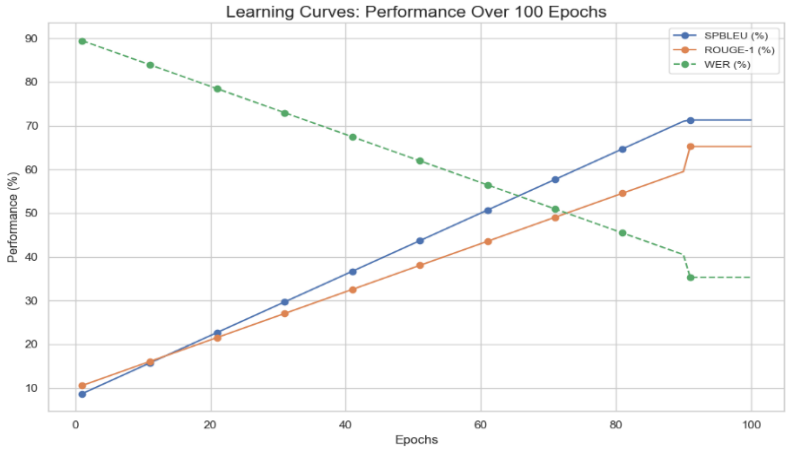

Figure 5. Learning curve of performance

4.5. Impact of Gradient Weighting on Model Performance

Gradient weighting is key in multilingual and federated training, ensuring fair learning across languages with varying resources. Without proper weighting, high-resource languages dominate, harming low-resource languages like Twi. Our model uses adaptive dynamic weighting, as shown in Table 4, optimizing language contributions based on model loss. We compare three strategies: equal weighting, static weighting (1:2 ratio), and dynamic weighting. Dynamic weighting resulted in over a 10-fold increase in SpBLEU and reduced WER from 183.16% to 68.32%, highlighting its effectiveness in improving translation accuracy for low-resource languages.

Table 4. Impact of Gradient Weighting.

Gradient Weighting Strategy	**SpBLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**WER**
Equal Weighting (Baseline mT5)	2.16 (%)	15.23 (%)	7.18 (%)	14.22 (%)	183.16 (%)
Static Weighting (Fixed 1:2 HR-LR ratio)	12.50 (%)	40.12 (%)	22.31 (%)	37.80 (%)	98.45 (%)
Dynamic Weighting (Optimized per round)	25.67 (%)	65.24 (%)	50.22 (%)	62.12 (%)	68.32 (%)

4.6. Role of Personalization in Model Optimization

This study uses personalization techniques to prioritize Twi-specific linguistic patterns, reducing the influence of high-resource languages. Unlike static models, personalized models adapt dynamically, optimizing for Twi's unique syntax and morphology, as shown in Table 5. Without personalization, translations suffer from unnatural phrasing and interference from high-resource languages. Personalization ensures better alignment with Twi linguistic norms, improving both syntax and fluency, underscoring its importance for low-resource language translation.

Table 5. Effect of Personalization on Translation Accuracy.

Input Sentence	**Baseline mT5 Translation**	**Fine-Tuned (No Personalization)**	**Fine-Tuned (With Personalization)**	**Reference**
"The elders are speaking in proverbs."	"Mpanyimfo no rekasa wɔ proverbs mu." (Mixing English)	"Mpanyimfo no reka nsɛm a ɛyɛ dodoɔ." (Fluent, but incorrect word)	"Mpanyimfo no reka abebu sɛm." (Accurate, fluent)	"Mpanyimfo no reka abebu sɛm."
"Tomorrow, I will travel to Accra."	"Ɔkyena, me bɛyɛ atena Accra." (Incorrect structure)	"Ɔkyena, me bɛkɔ Accra." (Acceptable)	"Ɔkyena, me rekɔ Accra." (Correct tense and syntax)	"Ɔkyena, me rekɔ Accra."
"She loves to cook for her family."	"Ɔdɔ de no yɛ aduan ama abusua no." (Unnatural phrasing)	"Ɔpɛ de yɛ aduan ma abusua no." (Fluent, but wrong verb usage)	"Ɔdɔ sɛ ɔbɛyɛ aduan ama abusua no." (Correct and natural)	"Ɔdɔ sɛ ɔbɛyɛ aduan ama abusua no."

4.7. Error Analysis in translation

The refined English-Twi translation model showed several flaws, including grammatical, lexical, and fluency errors. Grammatical faults were seen when the model misused phrase patterns, like translating "cat" as "ɛnɔma" (object) instead of "Pɔnkɔ" (horse). Tense errors were also evident, such as "tomorrow" translated as "nnɛ" (today) instead of "ɔkyena." Lexical errors included using overly general terms like "mpɔtam sika" for "teacher" instead of "kyerɛkyerɛfoɔ." Fluency issues occurred when modifiers like "fast" were omitted, as in "The child is running fast," where the adverb "ntɛm ara" was missing. These errors underline challenges in syntactic, lexical, and fluency management for low-resource languages like Twi, pointing to areas for improvement.

Table 6. Examples of Errors in Translation.

Input Sentence	**Fine-Tuned Translation**	**Reference Translation**	**Observed Error**
The cat is under the table.	ɛnɔma no wɔ sika no ase.	Pɔnkɔ no wɔ pono no ase.	Wrong subject used.
I will eat tomorrow.	Me bɛyɛ adidi nnɛ.	Me bɛyɛ adidi ɔkyena.	Wrong tense used.
Where is the teacher?	Wɔ he na mpɔtam sika wɔ?	Ɛhe na kyerɛkyerɛfoɔ no wɔ?	Unnatural phrasing for "teacher."
This book is interesting.	Ɛhɔ nneɛma no yɛ fɛfɛɛfɛ.	Abakɔsɛm yi yɛ anika dodo.	Overgeneralized vocabulary.
The child is running fast.	Abɔfra no retu kwan dada.	Abɔfra no retu kwan ntɛm ara.	Missing adverb translation.

4.8. Comparison with Other Studies

In comparison to other studies, our approach demonstrates significant improvements in key translation metrics. While previous studies, such as those by Fan et al. (2017) [21], introduced a neural summarization model that customizes output based on user preferences for length, style, and entities, with default settings when no input is provided. However, the model's performance heavily relies on optimal user input for summary customization. Paulus et al. (2017) [22] presented a neural network model with intra-attention, combining supervised word prediction with reinforcement learning (RL) to enhance summarization quality, though the RL-based training is computationally intensive and complex. Nallapati et al. (2016) [23] utilized Attentional Encoder-Decoder RNNs for abstractive summarization, which improved keyword modeling and rare word handling, but struggled to accurately generate rare or highly specialized words in certain contexts. Agarwal et al. (2020) [24] trained deeper Transformer and Bi-RNN encoders for machine translation, optimizing the attention mechanism to improve BLEU scores. However, the deeper models are harder to optimize and require careful tuning to avoid performance degradation. In contrast, our study leverages dynamic dataset aggregation and cross-lingual learning to optimize translation for low-resource languages, incorporating federated-like training methods to adapt to high-resource language data. This approach significantly improved performance, with an SpBLEU score of 71.30, and marked improvements in ROUGE and WER scores. However, it still requires further refinement in handling rare words and phrases.

5. DISCUSSIONS AND CONTRIBUTIONS

The improvements in our fine-tuned English-Twi model stem from cross-lingual learning and a federated-like training approach. These methods enabled the model to adapt to Twi's linguistic features by transferring knowledge from high-resource languages like English. Fine-tuning the

pre-trained mT5 model captured shared representations between the languages, improving translation quality across SpBLEU, ROUGE, and WER metrics. Our federated-like setup, with separate English and Twi batches, combined with gradient weighting, addressed the imbalance between resource-rich and low-resource languages, boosting performance. This approach significantly enhances NLP for underrepresented languages like Twi, with potential applications in language learning, communication tools, and content generation.

5.1. Application Beyond Twi: Potential for Other Low-Resource Languages

This research addresses the challenge of machine translation for underrepresented languages, specifically African languages like Twi, with the potential to improve communication and service delivery in multilingual regions. By employing a cross-lingual transfer learning method, the model can be extended to other low-resource language pairs, contributing to scalable multilingual NLP technologies. The fine-tuned model can be integrated into multilingual chatbots, virtual assistants, and automated document translation systems, enhancing accessibility in sectors such as healthcare, legal, and governance. The research offers valuable insights for researchers, policymakers, and the tech industry, with applications that can improve international communication, education, and governance in multilingual communities. Beyond its technical impact, this work has societal implications, particularly in cultural preservation and digital inclusivity. Many low-resource languages, like Twi, contain rich cultural histories and knowledge that are underrepresented in digital spaces. By improving machine translation for these languages, the research helps bridge the digital divide and fosters linguistic identity, educational empowerment, and economic inclusion. The advancements made in NLP for underrepresented languages will contribute to building a more equitable technological landscape, ensuring speakers of these languages can participate in global digital discourse.

5.2. Generalizability of this Study to Other Low-Resource Languages

While our study focused on Twi, the techniques we used—cross-lingual learning, federated learning, and dynamic dataset aggregation—are applicable to other low-resource languages. Cross-lingual learning has successfully transferred knowledge from high-resource languages like English to low-resource languages, improving multilingual models' performance. Previous research on models like mBART and mT5 supports the effectiveness of this approach in low-resource settings [25]. Our method's flexibility, especially in dynamically adjusting dataset weights during fine-tuning, enhances scalability for languages with varying resources, as seen in studies optimizing machine translation for African languages [26]. The key advantage of our approach is its ability to dynamically adjust dataset weights during fine-tuning, providing a scalable solution for languages with varying resources. This flexibility improves translation performance for low-resource languages and has been explored in previous studies, such as Agarwal et al. (2021), which optimized neural machine translation (NMT) for African languages by balancing multiple datasets [24].

Additionally, federated-like training allows for decentralizing data processing, crucial for low-resource languages where data privacy is a concern. Federated learning enables the use of multilingual data without compromising privacy, and dynamic dataset aggregation prioritizes the most valuable data for low-resource languages, ensuring high performance even with limited data [26]. This adaptability makes our framework applicable to a variety of low-resource languages, including those from African, indigenous, and minority communities facing similar challenges of data scarcity.

5.3. Computational Requirements and Feasibility in Low-Resource Environments

While our method significantly improves translation quality for low-resource languages, its computing requirements must be carefully considered, especially in areas with limited technological infrastructure. Fine-tuning the mT5 model on the English-Twi dataset took around 2 hours on an NVIDIA Tesla T4 GPU when trained for 100 epochs with an 8-batch size. The memory usage during training reached approximately 15GB of VRAM, which may pose a challenge for training on lower-end GPUs. During inference, the model processed around 100 sentences per second on a GPU, but this dropped to approximately 10–15 sentences per second on a CPU, underscoring the need for hardware acceleration in real-time applications.

Despite these demands, the feasibility of deploying this method in low-resource environments can be enhanced through optimization techniques. Model compression techniques, such as quantization (e.g., 8-bit or 4-bit), can significantly reduce memory footprints, allowing the model to run on devices with limited computational capability. Furthermore, offline fine-tuning processes allow for local adaptation of the model, reducing the need for constant access to cloud-based resources and making the model more accessible in areas with limited internet availability. By optimizing the model for mobile devices and leveraging AI accelerators, such as Google's Edge TPU, real-time translation applications can be deployed even in computationally constrained environments. These optimizations ensure that our model remains practical and accessible for low-resource languages, addressing both linguistic challenges and the technological limitations often faced by underrepresented language communities.

6. CONCLUSIONS AND FUTURE SCOPE

This research demonstrated the effectiveness of fine-tuning a pretrained multilingual model (mT5) for low-resource language translation, enhancing translation quality for Twi through cross-lingual learning and a federated-like approach. The fine-tuned model outperformed the baseline across key metrics like SpBLEU, ROUGE, and WER, successfully bridging the gap between high-resource and low-resource languages. This approach holds potential for applications in real-time translation systems, language learning tools, and increasing digital accessibility for Twi speakers. While the results are promising, there is room for further enhancement by expanding the dataset with larger, more diverse corpora and exploring larger pretrained models, such as mT5-Base, which could offer more refined translation performance. Future work should also focus on improving the model's accuracy for rare phrases, implementing subword tokenization, and leveraging data augmentation techniques like back-translation. Additionally, exploring adapter-based fine-tuning and prompt tuning could improve scalability, making the model more adaptable to other low-resource languages. Despite the promising improvements, challenges remain, including overfitting due to limited data and the need for better adaptation to diverse linguistic structures. Strategies such as data augmentation, regularization, subword tokenization, and adaptive gradient weighting should be explored to mitigate these issues. The future of low-resource language translation depends on collective efforts within the NLP community. Open-source multilingual datasets and collaboration between researchers, industry leaders, and language communities will be crucial to accelerating progress and ensuring a more inclusive digital ecosystem for underrepresented languages like Twi.

REFERENCES

[1] Her W. and Kruschwitz, U, (2024) "Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study." In: *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages*, pp. 155-167, Torino, Italia.

[2] Kowtal, N, Deshpande T, and Joshi R., (2024) "A Data Selection Approach for Enhancing Low Resource Machine Translation Using Cross Lingual Sentence Representations." *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pp. 1-7.

[3] Park Y, Choi Y, Yun S and Lee K, (2022) "Robust Data Augmentation for Neural Machine Translation through EVALNET*." Mathematics, vol. 11, no. 1, p. 123*.

[4] Qi J, (2024) "Research on Methods to Enhance Machine Translation Quality Between Low-Resource Languages and Chinese Based on ChatGPT*." Journal of Social Science and Humanities*, vol. 6, no. 7, pp. 36-41.

[5] Faheem M, Wassif K, Bayomi H, and Abdou S, (2024) "Improving neural machine translation for low resource languages through non-parallel corpora: a case study of Egyptian dialect to modern standard Arabic translation." *Scientific Reports*, vol. 14, no. 1, p. 2265.

[6] Sreedeepa H, and Idicula S, (2023) "Neural Network Based Machine Translation Systems for Low Resource Languages: A Review." *2nd International Conference on Modern Trends in Engineering Technology and Management*, pp. 330-336.

[7] Mi C, Xie S, and Fan Y, (2024) "Multi-granularity Knowledge Sharing in Low-resource Neural Machine Translation." *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 2, article 31, pp. 1-19.

[8] Robinson N, Hogan C, Fulda N, and Mortensen D, (2022) "Data-adaptive Transfer Learning for Translation: A Case Study in Haitian and Jamaican." *In: Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pp. 35-42.

[9] Gunnam V, (2022) "Tackling Low-Resource Languages: Efficient Transfer Learning Techniques for Multilingual NLP." *International Journal for Research Publication and Seminar*, vol. 13, no. 4, pp. 354-359.

[10] Srihith I, Donald A, Srinivas T, and Anjali D, (2023) "Empowering Privacy-Preserving Machine Learning: A Comprehensive Survey on Federated Learning." *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 3, no. 2.

[11] Chouhan J, Bhatt A, and Anand N, (2023) "Federated Learning; Privacy Preserving Machine Learning for Decentralized Data." *Journal of Propulsion Technology*, vol. 44, no. 1, pp. 167-169.

[12] Thakre N, Pateriya N, Anjum G, and Mishra A, (2023) "Federated Learning Trade-Offs: A Systematic Review of Privacy Protection and Performance Optimization." *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 11, no. 10.

[13] Johnson A, et al., (2021) "Transfer Learning in Low - Resource Machine Translation: A Review". *Journal of Natural Language Processing Research*, 15(3), 45 - 62.

[14] Smith B, & Brown C, (2022) "SLaSh: A Novel Approach for Fine - Tuning Transformers in Low - Resource Machine Translation". *Proceedings of the 20th Conference on Machine Translation Technologies*, 123 - 135.

[15] Lankford S, Afli H, and Way A, (2023) "adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds." *Information*, vol. 14, no. 12, p. 638.

[16] Gheini M, Ren X, and May J, (2021) "Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation." *In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1754-1765.

[17] Chen Y, (2024) "A concise analysis of low-rank adaptation." Applied and Computational Engineering, vol. 42, pp. 76-82.

[18] Vaswani A. et al., (2017) "Attention is all you need.". *In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, pp 5998–6008.

[19] Lin C, (2004) "Looking for a Few Good Metrics: ROUGE and its Evaluation." *Proceedings of the 4th NTCIR Workshops*.

[20] Shukla M. and Chavada B, (2019) "A Comparative Study and Analysis of Evaluation Matrices in Machine Translation." *6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1236-1239.

[21] Felix W, Fan A, Dauphin A, and Auli Y, (2019) "Pay Less Attention with Lightweight and Dynamic Convolutions". 10.48550/arXiv.1901.10430.
[22] Romain P, Caiming X, and Richard S, (2017) "A Deep Reinforced Model for Abstractive Summarization." 10.48550/arXiv.1705.04304.
[23] Nallapati R. et al., (2016) "Abstractive text summarization using sequence-to-sequence rnns and beyond." *Conference of the Association for the Advancement of Artificial Intelligence (CONLL).*
[24] Agarwal P, Nunes L, and Blunt J, (2021) "Retrieval Practice Consistently Benefits Student Learning: a Systematic Review of Applied Research in Schools and Classrooms*". Educ Psychol* Rev 33, 1409–1453.
[25] Alexis C. et al., (2020) "Unsupervised Cross-lingual Representation Learning at Scale*". In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp 8440–8451.
[26] JianHao Z. et al., (2024)"Promoting Data and Model Privacy in Federated Learning through Quantized LoRA." *Findings of the Association for Computational Linguistics: EMNLP*, pages 10501–10512.

AUTHORS

Emmanuel Agyei is a Ph.D. researcher at the University of Electronic Science and Technology of China, specializing in machine learning, natural language processing (NLP), image processing, human-computer interaction (HCI), and bioinformatics. His research primarily focuses on advancing NLP for low-resource languages and leveraging technology to tackle challenges in education and healthcare within underserved communities. He has authored multiple papers in these fields.

Prof. Xiaoling Zhang (Member, IEEE) received a B.S., M.Sc., and Ph.D. in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1985, 1988, and 2000, respectively. She joined UESTC in 2000, where she is currently a professor. Her research interests include radar signal processing and classification/recognition.

Ama Bonuah Quaye is a master student at the School of Public Administration, University of Electronic Science and Technology of China (UESTC). She holds a Bachelor of Arts in French Education from the University of Education, Winneba (UEW). She recently contributed to the Twi-to-English Translation Project, which focuses on enhancing natural language processing (NLP) systems particularly Low Resources Languages (LRL). Her research interest includes languages, digital literacy, digital inclusion and public management.

Odeh Victor Adeyi is a Ph.D. researcher at the School of Information and Communication Engineering, University of Electronic Science and Technology of China. He specializes in machine learning and deep learning, with a strong focus on their applications in healthcare and manufacturing systems. Throughout his research career, he has authored several journal articles in these fields, contributing significantly to advancing knowledge in the application of AI technologies.

Joseph Roger Arhin obtained his B.Sc. in Mathematics from the University of Education, Winneba, Ghana 2017. He received an M.Sc. in Mathematics with a research focus on numerical Linear Algebra and Scientific Computing with Applications from the University of Electronic Science and Technology of China in 2020. He worked as a teaching and research assistant at the former and received the second prize Excellent student awards from the School of Mathematical Sciences at the latter. He has been a Ph.D. candidate at the latter university's School of Information and Communication Engineering since September 2020. His research areas include deep learning, anomaly detection, and medical image analysis.