A Comparative Analysis of Baseline English–IsiXhosa Neural Machine Translation Models for Medical domain

Malibongwe Makhonza
University of Cape Town
Cape town, South africa
mkhmal024@myuct.ac.za

ABSTRACT

Neural Machine Translation (NMT) for low-resource languages like isiXhosa face significant challenges due to limited high-quality parallel data. To address these issues, this paper investigates the performance of two NMT models for English (Eng) ↔ isiXhosa translation: a Transformer encoder-decoder model trained from scratch, and a pretrained multilingual MT model finetuned for English-isiXhosa. We compare these models' performance on the FLORES-200 'dev' set and two medical datasets. During development, we tuned our models' settings using both general-domain and medical-domain validation datasets to compare how they performed. Our findings show that the model trained from scratch generally achieved better translation quality on general-domain tasks compared to the fine-tuned pre-trained model. However, when we evaluated the models on the specific medical domain test set (the Blocker dataset), while both models showed lower surface-level accuracy (BLEU scores), the pre-trained model achieved relatively high AfriCOMET scores. This suggests that even if the translations didn't perfectly match the reference wording, the models were still able to produce fluent and semantically correct translations, showing an underlying understanding of the medical context. This work provides insights into how different training approaches impact domain-specific translation quality for low-resource languages in specialized fields like medicine.

KEYWORDS

Neural Machine Translation, isiXhosa, Low-resource languages, Transformers, Domain adaptation

ACM Reference Format:
Malibongwe Makhonza. 2025. A Comparative Analysis of Baseline English–IsiXhosa Neural Machine Translation Models for Medical domain. In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Clear communication is vital in many aspects of life, especially in South Africa, a country rich in linguistic diversity where isiXhosa is the second-most common home language, spoken by over eight million people [17]. Many important services, including healthcare, operate primarily in English. This creates a significant communication gap, as patients who speak isiXhosa may struggle to understand critical medical information, potentially leading to misunderstandings or incorrect treatment [2]. Machine Translation (MT) offers a powerful solution to bridge these language barriers, improving access to information and ensuring better understanding in crucial sectors like healthcare [4].

Despite recent progress in Neural Machine Translation (NMT) with advanced architectures like the Transformer [3], challenges persist, particularly for low-resource languages such as isiXhosa. A major problem is the limited amount of parallel text data (sentences translated both ways) available for training NMT models [5]. While large multilingual models like NLLB-200 include isiXhosa, their performance can be limited in low-resource contexts or when dealing with specific language nuances [7, 8]. Furthermore, traditional overlap-based evaluation metrics like BLEU often do not fully capture the nuances of human translation quality, especially for low-resource languages, which makes assessing true performance challenging beyond simple word match [9, 10].

Our work is fundamentally motivated by the critical need to advance isiXhosa Natural Language Processing (NLP) and, crucially, to establish a robust foundation for future domain-specific adaptation, particularly for vital sectors such as medical texts [2]. By effectively bridging existing communication gaps, we aim to significantly improve access to essential information and empower isiXhosa speakers within various critical contexts. In this paper, we embark on an investigation of two widely recognized yet distinct approaches to machine translation for isiXhosa: training a Transformer model from scratch and fine-tuning a pre-trained multilingual machine translation model. Both methodologies come with inherent trade-offs; while a model trained from scratch potentially offers greater flexibility and tailored performance by learning language patterns specifically from the ground up, it often demands substantial amounts of high-quality parallel data to reach competitive performance. Conversely, fine-tuning a pre-trained multilingual model leverages a vast amount of prior linguistic knowledge acquired from diverse languages, offering a strong starting point, but may struggle to adapt fully to the unique nuances and specific terminology of a new, low-resource domain like isiXhosa medical text. It is currently unclear which of these two strategies will ultimately yield superior results for the challenging task of low-resource medical machine translation. Therefore, this research directly addresses this uncertainty by comparing these two approaches, utilizing the general-domain WMT22 dataset for initial training and fine-tuning, and performing a final, critical evaluation on the specialized medical domain test set, the Blocker dataset, to ascertain which method leads to better domain-specific translation performance for isiXhosa medical texts [1, 2].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

2.1 Neural Machine Translation for isiXhosa

Neural Machine Translation (NMT) has greatly improved how well computers can translate automatically by using deep learning and a technique called "attention" [12]. The Transformer model is currently the most common design because it can handle longer sentences and train faster than older methods [3]. However, NMT models need a lot of parallel data (sentences translated both ways) to work well. If there isn't much data, especially for low-resource languages like isiXhosa, the translation quality can drop significantly unless domain adaptation techniques or other advanced methodologies are employed [**?**]. For widely spoken languages like French or German, this data is usually available. But for languages like isiXhosa, finding large sets of parallel data is a significant challenge [5].

Some progress has been made using multilingual models, such as NLLB-200 [7], which was trained on over 200 languages, including isiXhosa. Multilingual models are particularly beneficial for low-resource languages through cross-lingual transfer, where knowledge learned from data-rich languages is transferred to improve the performance of languages with limited resources. This shared linguistic understanding helps overcome data scarcity by leveraging common patterns and structures across multiple languages. However, even with these models, isiXhosa translation isn't perfect because there's still not much isiXhosa data in their training sets, which limits the extent of direct learning for this specific language [7]. Furthermore, large language models (LLMs) like ChatGPT and LLaMA, which work well for major languages, often struggle with African languages [13]. Studies have shown that these models sometimes leave out or even make up isiXhosa terms [8]. This highlights why it's crucial to build or adjust translation models specifically for isiXhosa, rather than solely relying on generic tools.

Recent projects like the WMT22 competition [1] have helped increase the amount of parallel data available for African languages, including isiXhosa. WMT is an annual event where researchers compare different translation systems using the same public data and evaluation methods. The 2022 competition included English ↔ isiXhosa as one of the supported language pairs [1]. This type of shared benchmark is very helpful for comparing models and creating better starting points for future research.

2.2 Domain Adaptation for Translation

Even when a model is trained on a huge dataset of general language (like WMT22), its performance may decline when translating text from a specialized area like medicine. This is because medical language uses technical words, abbreviations, and phrasings that differ significantly from everyday conversation. Domain adaptation refers to methods that help a translation model adjust to the specific style and vocabulary of a particular subject area.

One common strategy, when sufficient in-domain parallel data is available, is to fine-tune an existing translation model using this specialized data [4]. However, for truly low-resource domains and languages, acquiring adequate in-domain parallel data is often not feasible. In such scenarios, synthetic data generation methods become crucial. Two prominent techniques in this area are back-translation and forward-translation.

Back-translation involves using a machine translation model that translates from the target language to the source language (e.g., isiXhosa to English) to translate a large corpus of monolingual data in the target language (isiXhosa) back into the source language (English). This generates synthetic parallel pairs where the target side is authentic, and the source side is synthetic. These pairs are then added to the training data to enhance the performance of the primary source-to-target model (e.g., English to isiXhosa) [14]. Conversely, *forward-translation* uses an existing source-to-target MT model (e.g., English to isiXhosa) to translate monolingual source language data (e.g., English medical documents from the PriMock57 dataset) into the target language (isiXhosa). The resulting synthetic pairs (original English with computer-generated isiXhosa) are then used to augment the training data, helping the model learn domain-specific vocabulary [14]. Both methods are especially useful when there is not much real human-translated bilingual data available for a specific domain.

Another helpful method is DALI (Domain Adaptation by Lexicon Induction) [18]. DALI utilizes a bilingual dictionary to replace words in a monolingual domain text with translations from the source language, thereby creating new, artificial parallel sentences. This approach is also particularly useful when there is insufficient real medical translation data. A recent study confirmed that DALI performed better than other methods, such as continuous pre-training, for medical translation in languages with very few resources [11]. This demonstrates that simpler, data-focused methods like DALI can still be very effective for domain adaptation, especially in healthcare settings.

Our project builds on these ideas by investigating the application of both back-translation and DALI to adapt general English ↔ isiXhosa models specifically for the medical domain. This helps ensure that the translations are accurate and use the correct medical terminology, which is paramount in healthcare communication.

3 MODELS

This section outlines the different machine translation models investigated in this project to address our research questions regarding isiXhosa medical translation. Our primary goal is to determine the most effective strategy for domain-specific translation in a low-resource setting. To achieve this, we first compare two foundational approaches: training a model from scratch versus fine-tuning a pre-trained multilingual model on general-domain data. Subsequently, these models will be benchmarked against domain-adapted models, developed by our teammates using synthetic data generation and lexicon induction, to comprehensively evaluate which methodology yields the best performance for medical translation tasks.

3.1 Baseline Models

Our first approach involves creating two foundational models for English ↔ isiXhosa translation. The first is a Transformer encoder-decoder model trained entirely from scratch. This model begins with no prior linguistic knowledge and learns translation solely from the provided large general-domain dataset, WMT22 [1]. The motivation for training from scratch is that it allows for greater

control over the model architecture and the data it is exposed to, potentially leading to a more specialized model that is computationally efficient. However, a significant drawback is that achieving optimal performance often requires vast amounts of data, which can be suboptimal for truly low-resource languages.

The second model is built upon the NLLB-200 (No Language Left Behind) model, a powerful pre-trained multilingual model developed by Meta [7]. This NLLB-200 model is then further trained (fine-tuned) using the same WMT22 dataset. The advantage of fine-tuning a pre-trained multilingual model is its ability to leverage a vast amount of multilingual knowledge acquired during its initial training, potentially providing a strong starting point even for low-resource languages. This process essentially specializes the model for isiXhosa while benefiting from cross-lingual transfer. However, a potential issue is that the knowledge from other languages within the multilingual model might interfere with or dilute the specific nuances required for high-quality isiXhosa translation. Both of these models are developed to handle translation in both English → isiXhosa and isiXhosa → English directions, and they serve as crucial benchmarks for our investigation into effective translation strategies.

3.2 Synthetic Data Generation Model

The second approach focuses on creating artificial (synthetic) medical-domain data to improve translation quality. This is achieved using two methods: back-translation and forward-translation. Back-translation involves translating monolingual target language text (isiXhosa) "back" into the source language (English) using a reverse model, generating synthetic parallel pairs [14]. Conversely, forward-translation reverses this process by translating monolingual source language text (English medical documents from the PriMock57 dataset) into synthetic isiXhosa. The NLLB-200 model is utilized for these synthetic data generation processes, and the resulting synthetic data then helps fine-tune the main English ↔ isiXhosa models for improved domain specificity. This model will be developed and evaluated by a teammate as part of the broader project.

3.3 DALI Model

The third approach employs Domain Adaptation by Lexicon Induction (DALI) [18]. This method constructs synthetic parallel sentences by building a bilingual dictionary from various sources, including general and medical terms. This lexicon is subsequently used to translate English medical sentences into isiXhosa through word-to-word mapping. The pseudo-parallel corpus generated by DALI is then used to train or fine-tune an NMT model capable of translating in both English ↔ isiXhosa directions. This method has shown promise in adapting to specific domains, particularly for low-resource languages [11]. This DALI model will also be developed and evaluated by a teammate, and its performance will be critically compared against the models discussed in this paper.

4 EXPERIMENTAL SETUP

4.1 Computational Resources

All our model training and evaluation tasks were carried out on the Centre for High Performance Computing (CHPC) GPU cluster. This cluster provides powerful computing resources essential for deep learning models. The jobs ran on GPUs of type Tesla V100-PCIE-16GB, each equipped with 16 Gigabytes (GB) of dedicated memory (VRAM). These requests included: 8 central processing units (`ncpus=8`), 32 GB of main system memory (`mem=32gb`), and a maximum runtime of 12 hours (`walltime=12:00:00`).

4.2 Datasets

To train and evaluate the models, specific datasets were used and prepared accordingly. The models were primarily trained on the WMT22 English ↔ isiXhosa parallel corpus [1]. This is a large dataset containing about 8 million translated sentence pairs. Before training, this raw data was cleaned and formatted to make it suitable for our machine translation models. The data was filtered to remove all misaligned and sentence pairs with low translation scores.

For validating our model's performance during training, we used the FLORES-200 "dev" set [15]. This widely accepted benchmark dataset, comprising 997 carefully selected sentence pairs, served as a general-domain validation set. In addition to FLORES, we utilized another specialized medical-domain dataset for further validation to assess performance on specialized content during development. Crucially, the Blocker dataset [2] was reserved exclusively for the final evaluation of our models on medical text, providing an unbiased assessment of their ability to translate specialized content.

When preparing data for the fine-tuned NLLB models, we used the autotokenizer tool from the Hugging Face Transformers library. This tool breaks down sentences into smaller pieces (tokens), ensuring they are not longer than 128 tokens, and truncates any parts that exceed this length. After tokenization, a `DataCollatorForSeq2Seq` was used to group these tokenized sentences into batches and add necessary padding, which helps the model process them efficiently.

For the Transformer model trained from scratch, a slightly different data preparation method was employed. Initially, a SentencePiece tokenizer was trained using all the WMT22 data to learn sub-word unit segmentation. This tokenizer was then applied consistently across both the training and validation datasets. Subsequently, these tokenized sentences were converted into the specific binary format required by this model architecture.

4.3 Training Parameters

Both types of models (NLLB and the Transformer model trained from scratch) were trained using specific settings to optimize their learning and ensure efficient use of our computational resources.

The NLLB models were fine-tuned with a learning rate of 3×10^{-5}, running for 1 full epoch. This limited number of epochs was necessary due to constraints on available GPU capacity and the computational intensity of training large models. However, it is generally understood that fine-tuning pre-trained models requires significantly less data and fewer training epochs compared to training from scratch, as they leverage extensive prior knowledge. A mixed-precision training (`fp16=True`) was used, which helps speed up training and reduces the amount of memory needed on the GPU. Training progress was checked and logged every 50 steps, and evaluations on the FLORES "dev" set were performed every 2500 steps. The last 5 checkpoints were saved, and the system was set up to

automatically identify and load the model that achieved the best BLEU score during training.

The Transformer model trained from scratch, implemented using the Fairseq toolkit[1], was optimized with the `adam` optimizer. It used a learning rate of 5×10^{-4}, starting with a "warmup" phase of 4000 updates. To prevent the model from simply memorizing the training data, a dropout rate of 0.3 and a `weight_decay` of 0.0001 was applied. The training focused on minimizing the label smoothing loss function, with a label-smoothing value of 0.1. Each training step processed a maximum of 4096 tokens, and the training could run for up to 30 epochs. However, it would stop earlier if the model's performance did not improve for 5 consecutive evaluations (`patience=5`). The last 5 epochs were saved along with a log of the training progress every 100 steps.

Because our Tesla V100 GPUs (16GB VRAM) have limited memory, we used a small `per_device_train_batch_size` of 4 sentences. To simulate the effect of a much larger batch, we combined this with a high `gradient_accumulation_steps` of 64. This gave us an effective batch size of $4 \times 64 = 256$, which greatly improved training speed while preventing memory errors. This balanced approach was vital for completing our training tasks within the allocated 12-hour walltime.

4.4 Evaluation Protocol

To assess the quality of machine translations, we employed a suite of widely recognized metrics: BLEU (Bilingual Evaluation Understudy), *chrF*, *chrF*$^{++}$, and AfriCOMET.

BLEU (Bilingual Evaluation Understudy) is a precision-focused metric that measures the overlap of n-grams between a machine's translation and one or more human-generated reference translations [9]. A higher BLEU score indicates greater similarity to the reference.

chrF (Character n-gram F-score) addresses some limitations of word-based metrics, particularly for morphologically rich or low-resource languages where word forms can vary significantly. It focuses on character n-grams, which can be more robust to word choice variations and useful for languages like isiXhosa with complex word structures [10].

chrF^{++} extends the *chrF* metric by incorporating both character and word n-grams, providing a more comprehensive evaluation of translation quality. This hybrid approach aims for a more robust assessment by balancing the strengths of both character-level and word-level comparisons [19].

Finally, AfriCOMET is a metric specifically designed to evaluate machine translation for under-resourced African languages. It enhances the COMET framework by leveraging multilingual models and techniques tailored to the unique linguistic characteristics and data scarcity challenges of these languages, providing a more nuanced assessment of translation quality, including fluency and semantic adequacy, beyond surface-level lexical overlap [20].

5 RESULTS

This section presents the experimental results obtained from evaluating our machine translation models.

[1]Fairseq is an open-source sequence modeling toolkit developed by Facebook AI Research.

Table 1: BLEU, chrF, and chrF++ scores for English → isiXhosa translation on the FLORES dataset.

Model	BLEU	chrF	chrF++
NLLB - best	**15.2318**	**59.9594**	**53.427**
Trained from Scratch	15.0384	51.5014	45.5447

Table 2: BLEU, chrF, and chrF++ scores for isiXhosa → English translation on the FLORES dataset.

Model	BLEU	chrF	chrF++
NLLB - best	21.1680	**56.7993**	**54.2700**
Trained from Scratch	**26.9002**	51.2570	49.7355

Table 3: BLEU, chrF, and chrF++ scores for English → isiXhosa translation on the MeMaT dataset.

Model	BLEU	chrF	chrF++
NLLB - best	**13.6426**	**51.6638**	**44.1731**
Trained from Scratch	9.2151	43.4204	38.3508

Table 4: BLEU, chrF, and chrF++ scores for isiXhosa → English translation on the MeMaT dataset.

Model	BLEU	chrF	chrF++
NLLB - best	**39.8468**	**64.6963**	**61.7682**
Trained from Scratch	20.5205	43.7708	42.9585

5.1 Comparing training from scratch and finetuning NLLB

The main investigative question focused on comparing the performance of a Transformer model trained from scratch against a fine-tuned NLLB-200 model. On the general-domain FLORES dataset, the NLLB - best model demonstrated slightly superior performance for English → isiXhosa translation across all metrics (BLEU, chrF, and chrF++), as shown in Table 1. However, for isiXhosa → English translation on FLORES, the Trained from Scratch model achieved a notably higher BLEU score, while the NLLB - best model led on chrF and chrF++ (Table 2). The discrepancies in performance between the two models on the FLORES dataset, particularly for English → isiXhosa, were relatively small, indicating a largely similar aptitude for general-domain translation in this direction.

Moving to the medical-domain MeMaT validation set, the NLLB - best model consistently outperformed the Trained from Scratch model across all metrics for both English → isiXhosa and isiXhosa → English translation directions (Table 3 and Table 4). This was a surprising result, which could suggests that the pre-trained knowledge embedded in the NLLB model, even after limited fine-tuning, proved beneficial on this particular medical dataset, potentially due to its characteristics or the nature of its medical terminology.

Table 5: BLEU, chrF, and chrF++ scores for English → isiXhosa translation on the Blocker dataset.

Model	BLEU	chrF	chrF++
NLLB - best	2.0849	5.9222	3.211
Trained from Scratch	**12.8898**	**39.8996**	**36.291**

Table 6: BLEU, chrF, and chrF++ scores for isiXhosa → English translation on the Blocker dataset.

Model	BLEU	chrF	chrF++
NLLB - best	5.8351	31.0478	26.489
Trained from Scratch	**14.3724**	**31.4262**	**30.5745**

Some important insights were derived from the evaluation on the Blocker dataset, which served as our dedicated medical-domain test set. Firstly, it is imporatnt to note that both models experienced a significant drop in performance compared to the general FLORES data, highlighting the inherent difficulty of translating specialized medical terminology and the impact of domain shift. However, for both English → isiXhosa and isiXhosa → English translation, the Trained from Scratch model substantially outperformed the NLLB - best model across all metrics (Table 5 and Table 6). For English → isiXhosa, the Trained from Scratch model achieved a BLEU score of 12.8898 compared to NLLB - best's 2.0849. Similarly, for isiXhosa → English, Trained from Scratch scored 14.3724 BLEU against NLLB - best's 5.8351. This stark difference on the Blocker test set is important in answering our primary research question, indicating that for truly domain-specific, low-resource medical text, training a Transformer model from scratch (with its more extensive training epochs) on the WMT22 data proved more effective than fine-tuning the NLLB-200, which might have struggled more with adapting its broad multilingual knowledge to the very specific medical context of the Blocker dataset.

Regarding translation directions, it is important to note that direct quantitative comparison of BLEU scores across different language pairs (e.g., English → isiXhosa versus isiXhosa → English) is not strictly valid due to the differing target languages. However, within each translation direction, model performance was evaluated consistently. We observed that for the isiXhosa → English direction, both models generally achieved higher scores on metrics like BLEU and chrF compared to the English → isiXhosa direction on the FLORES dataset. This could be attributed to the difficulty in generating grammatically correct and fluent isiXhosa, which is a morphologically rich language, when translating from English."'

5.2 Comparing domain-general to domain-specific modelling

This section extends our analysis by comparing the performance of our domain-general models against models that incorporate domain-specific adaptation techniques, utilizing the Blocker medical test dataset. Furthermore, the impact of fine-tuning NLLB models with mixed datasets, combining varying proportions of general-domain WMT22 data with synthetic domain-specific data.

Table 7: BLEU, chrF, and chrF++ scores for all models on the Blocker medical dataset.

Model ID	BLEU	chrF	chrF++
Forward Direction (Eng→isiXhosa)			
WMT22-NLLB	2.0849	5.9222	3.211
Trained from Scratch	12.8898	39.8996	36.291
DALI approach (NLLB-200)	9.72	**47.71**	42.08
Forward/Back-translation approach (NLLB)	**17.86**	43.02	**43.80**
Backward Direction (isiXhosa→English)			
WMT22-NLLB	5.8351	31.0478	26.489
Trained from Scratch	14.3724	31.4262	30.5745
DALI approach (NLLB-200)	19.03	41.01	39.13
Forward/Back-translation approach (NLLB)	**34.48**	**42.76**	**47.97**

5.2.1 Comparison of Domain-General Models with Domain Adaptated models. To assess the effectiveness of domain adaptation strategies, we first compared our domain-general models (Trained from Scratch and WMT22-NLLB) against the DALI approach (NLLB-200) and Forward/Back-translation approach (NLLB) models on the Blocker medical test dataset. The results, summarized in Table 7, show clear evidence of the benefits of domain-specific methodologies.

For the English → isiXhosa translation direction, both domain-adapted models significantly outperformed our WMT22-NLLB model. The Forward/Back-translation approach (NLLB) achieved the highest BLEU score of 17.86, highlighting a strong capacity for producing translations lexically similar to human references. The DALI approach (NLLB-200) secured the top score for chrF (47.71), while the Forward/Back-translation approach (NLLB) also led on chrF++ (43.80), suggesting these models maintained higher character-level and hybrid similarity, respectively, with the reference translations, as seen in Table 7. Our Trained from Scratch model also demonstrated better performance than our WMT22-NLLB model, but it was still surpassed by the best domain-adapted strategies for this direction.

Similarly, for the isiXhosa → English translation direction, domain adaptation proved to be quite effective. The Forward/Back-translation approach (NLLB) led yet again across all metrics on the Blocker dataset, achieving a BLEU score of 34.48, chrF of 42.76, and chrF++ of 47.97 (Table 7). The DALI approach (NLLB-200) also registered fairly higher BLEU, chrF, and chrF++ scores (19.03, 41.01, 39.13 respectively) than both our WMT22-NLLB and Trained from Scratch baseline models. These results indicate that for the Blocker medical test set, directly incorporating domain-specific knowledge through techniques like DALI or synthetic data generation consistently leads to better translation quality than relying solely on models trained on general-domain data.

5.2.2 Impact of Mixed Synthetic and General-Domain Data. Following our first comparison, we also looked at what happened when we fine-tuned NLLB models using a mix of datasets. This involved combining general-domain WMT22 data with synthetic domain-specific data from either the DALI or Forward/Back-translation methods.

Table 8: BLEU, chrF, and chrF++ scores for NLLB models fine-tuned with 50% WMT22 and 50% synthetic data on the Blocker dataset.

Model ID	BLEU	chrF	chrF++
Forward Direction (Eng→isiXhosa)			
WMT22/DALI approach synthetic data	4.9324	15.1884	14.0222
WMT22/Forward/Back-translation approach synthetic data	**24.2746**	**54.2830**	**49.0704**
Backward Direction (isiXhosa→English)			
WMT22/DALI approach synthetic data	2.7462	**31.4254**	**29.6935**
WMT22/Forward/Back-translation approach synthetic data	**3.7844**	29.8032	23.3644

Table 9: BLEU, chrF, and chrF++ scores for NLLB models fine-tuned with 30% WMT22 and 70% synthetic data on the Blocker dataset.

Model ID	BLEU	chrF	chrF++
Forward Direction (Eng→isiXhosa)			
WMT22/DALI approach synthetic data	5.6698	15.8235	15.0355
WMT22/Forward/Back-translation approach synthetic data	**24.2881**	**55.2608**	**49.8047**
Backward Direction (isiXhosa→English)			
WMT22/DALI approach synthetic data	28.2591	34.2672	34.1131
WMT22/Forward/Back-translation approach synthetic data	**33.6493**	**36.6500**	**39.0385**

For English → isiXhosa translation, the NLLB model fine-tuned with 50% WMT22 and 50% synthetic data generated by the Forward/Back-translation approach exhibited a huge increase in performance, achieving a BLEU score of 24.2746, chrF of 54.2830, and chrF++ of 49.0704 (Table 8). These scores do not only exceed our domain-general models, but also outperform the standalone domain-adapted models discussed in Subsection 5.2.1, suggesting that this specific combination and approach to synthetic data could be highly effective for generating isiXhosa medical text. The WMT22/DALI approach synthetic data, in this 50-50 mix, yielded lower scores for this direction. When the synthetic data proportion was increased to 70% (30% WMT22), the WMT22/Forward/Back-translation approach synthetic data maintained its leading position, with a BLEU score of 24.2881, chrF of 55.2608, and chrF++ of 49.8047 (Table 9), showing a slight further improvement across metrics compared to the 50-50 mix.

In the isiXhosa → English translation direction, the NLLB models fine-tuned with mixed data also showed great results. For the 50% WMT22 and 50% synthetic data blend (Table 8), the WMT22 mixed with Forward/Back-translation approach synthetic data achieved the highest chrF (31.4254) and chrF++ (29.6935) scores, while the WMT22/DALI approach synthetic data had a slightly higher BLEU of 3.7844. These scores suggest that this mix for backward translation might still face some challenges compared to our domain-general best model on blocker(Table 6). However, improvement was observed with the 30% WMT22 and 70% synthetic data mix (Table 9). Here, the WMT22 mixed with Forward/Back-translation approach synthetic data achieved a BLEU score of 33.6493, along with leading chrF (36.6500) and chrF++ (39.0385) scores. This significant increase shows that a higher proportion of synthetic domain-specific data, particularly from the Forward/Back-translation method, can enhance performance for isiXhosa → English medical translation by a huge amount.

Overall, these mixed data results clearly show that adding synthetic medical-domain data, especially through the Forward/Back-translation approach, could be pivotal for getting high-quality isiXhosa medical machine translation. Such mixed-data fine-tuning strategies can lead to huge performance gains, outperforming models trained on general-domain data or even those with domain-adaptation, by providing the NLLB model with targeted in-domain exposure. This highlights the value of carefully curated mixed datasets in bridging the gap for low-resource, specialized MT tasks.

Table 10: BLEU and AfriCOMET scores for NLLB - best model on the FLORES dataset.

Direction	BLEU	AfriCOMET
Eng→isiXhosa	15.2318	**0.7719**
isiXhosa→English	21.1680	**0.7966**

Table 11: BLEU and AfriCOMET scores for NLLB - best model on the MeMaT dataset.

Direction	BLEU	AfriCOMET
Eng→isiXhosa	13.6426	**0.7609**
isiXhosa→English	39.8468	**0.6874**

Table 12: BLEU and AfriCOMET scores for NLLB - best model on the Blocker dataset.

Direction	BLEU	AfriCOMET
Eng→isiXhosa	2.0849	**0.6672**
isiXhosa→English	5.8351	**0.7113**

5.3 AfriCOMET Analysis

While BLEU scores, especially on the Blocker medical dataset, were notably low, the corresponding AfriCOMET scores remained relatively high. This huge difference indicates that even when translations did not perfectly match the reference wording at a surface level (low BLEU), the NLLB - best model was often able to produce translations that were fluent and semantically adequate within the medical context, indicating a great understanding of the content.

For instance, on the FLORES dataset, the NLLB - best model achieved a BLEU score of 15.2318 for English → isiXhosa and 21.1680 for isiXhosa → English, with strong AfriCOMET scores of 0.7719 and 0.7966 respectively (Table 10). On the MeMaT medical validation set, we see a similar trend, with a BLEU score of 13.6426 for English → isiXhosa and 39.8468 for isiXhosa → English, alongside AfriCOMET scores of 0.7609 and 0.6874 (Table 11).

Crucially, on the Blocker medical test dataset, the NLLB - best model's BLEU scores dropped significantly to 2.0849 for English → isiXhosa and 5.8351 for isiXhosa → English. However, its AfriCOMET scores remained relatively high at 0.6672 and 0.7113, respectively (Table 12). This outcome highlights the need of advanced metrics like AfriCOMET for better evaluation of MT in low-resource and specialized domains. It is also proof that semantic understanding can persist even when direct lexical overlap with references is low.

5.4 Hyperparameter Tuning for NLLB Fine-tuning

Effective hyperparameter tuning is important when optimizing model performance. For the NLLB models, we conducted a systematic hyperparameter sweep focusing on the English → isiXhosa translation direction on a 200k subset of the WMT22 dataset, evaluated on the FLORES validation set. The goal was to identify the most effective configuration of learning rate, batch size, number of epochs, and other training parameters to achieve the best translation quality. The search space explored for this tuning included varying learning rates, effective batch sizes (controlled by gradient accumulation steps), number of epochs, and warmup ratios, while keeping weight decay and label smoothing fixed at 0.0.

The results of this sweep show that certain hyperparameters had a more pronounced impact on performance. A learning rate of 3×10^{-5} consistently produced favorable BLEU scores, often outperforming higher learning rates of 5×10^{-5} and 1×10^{-4} that sometimes led to more fluctuating or lower word-level accuracy, despite potentially higher character-level scores. Similarly, an effective batch size of 64, achieved through a base batch size of 4 with 16 gradient accumulation steps, proved beneficial for learning on this dataset size. While training for a single epoch yielded good initial results, increasing the number of epochs to 2 generally showed a slight improvement or comparable performance, suggesting that two epochs provided a good balance for model adaptation without overfitting on the validation set. The warmup ratio did not appear to have a dramatic impact within the tested range for this specific dataset.

Based on this sweep, the optimal hyperparameters identified for fine-tuning the NLLB model for English → isiXhosa translation on the FLORES validation set are summarized in Table 13. This configuration which was identified as Experiment 6 during our hyperparameter search, achieved a BLEU score of 19.0784, a chrF score of 60.0976, and a chrF++ score of 53.7882. These optimized parameters, with the exception of the number of epochs due to computational limitations, were then adopted for our NLLB - best model for subsequent evaluations.

Table 13: Optimal Hyperparameters and Performance for NLLB (Eng→isiXhosa on FLORES validation set).

Hyperparameter	Optimal Value
Learning Rate (LR)	3×10^{-5}
Effective Batch Size (EBS)	64
Gradient Accumulation Steps (GA)	16
Number of Epochs	2
Warmup Ratio (WR)	0.0
Weight Decay (WD)	0.0
Label Smoothing (LS)	0.0
Best Performance Metrics	
BLEU Score	19.0784
chrF Score	60.0976
chrF++ Score	53.7882

6 CONCLUSION

This project investigated two key approaches for English ↔ isiXhosa Neural Machine Translation (NMT) in medical and general domains: training Transformer models from scratch and fine-tuning pre-trained multilingual NLLB models. Our goal was to determine which strategy best addressed the challenges of low-resource medical translation.

Our primary findings demonstrate that for general-domain translation on FLORES, both baseline models performed comparably (Tables 1, 2), with NLLB slightly better Eng→isiXhosa and Trained from Scratch better isiXhosa→Eng BLEU. On the MeMaT medical validation set, the fine-tuned NLLB - best model consistently outperformed the Trained from Scratch model (Tables 3, 4). However, a critical insight from the Blocker medical test dataset (Tables 5, 6) revealed that the Trained from Scratch model significantly outperformed the fine-tuned NLLB - best model in both directions for truly domain-specific text.

Comparing these baselines with domain-specific models, it was clear that adaptation methods greatly enhanced performance. Both the DALI approach and the Forward/Back-translation approach consistently surpassed our domain-general baselines on the Blocker dataset (Table 7). The Forward/Back-translation approach, in particular, achieved the highest BLEU scores for both English → isiXhosa (17.86) and isiXhosa → English (34.48) medical translation.

Furthermore, integrating mixed synthetic and general-domain data largely improved results. Fine-tuning NLLB models with synthetic data, especially from the Forward/Back-translation approach, led to substantial performance gains on the Blocker dataset, achieving BLEU scores of 24.2881 for English → isiXhosa and 33.6493 for isiXhosa → English with a 30-70 mix (Table 9). These results strongly reinforce that strategically integrating synthetic domain-specific data is crucial for achieving high-quality medical machine translation for isiXhosa.

Finally, while BLEU scores on the Blocker dataset were low, the AfriCOMET analysis (Tables 10, 11, 12) showed relatively high scores for the NLLB - best model. This suggests that the model could produce fluent and semantically adequate medical translations, even when exact lexical overlap was minimal.

Based on these findings, future work should focus on generating more high-quality isiXhosa medical data, exploring advanced domain adaptation techniques, and optimizing the ratios of mixed synthetic and general-domain data for fine-tuning. Continued use of nuanced evaluation metrics like AfriCOMET is also vital for comprehensive assessment in low-resource and specialized domains.

REFERENCES

[1] Adelani, D. I., Whitenack, D., Neubig, G., et al. (2022). Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT 2022)*, pp. 773–800.

[2] Blocker, A., Meyer, F., Biyabani, A., Mwangama, J., Datay, M. I., and Malila, B. (2025). Benchmarking isiXhosa Automatic Speech Recognition and Machine Translation for Digital Health Provision. In *Proceedings of the Workshop on Patient-oriented Language Processing*.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008.

[4] Koehn, P., and Knowles, R. (2017). Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39.

[5] Nyoni, E., and Bassett, B. A. (2021). Low-Resource Neural Machine Translation for Southern African Languages. *arXiv preprint arXiv:2104.05579*.

[6] Emezue, C. C., et al. (2021). MMTAfrica: Multilingual Machine Translation for African Languages. In *ACL Workshop on African NLP*.

[7] Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

[8] Robinson, N., Ogayo, P., Mortensen, D. R., and Neubig, G. (2023). ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. *arXiv preprint arXiv:2309.07423*.

[9] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pp. 311–318.

[10] Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395.

[11] Marashian, A., Rice, E., Gessler, L., Palmer, A., and von der Wense, K. (2024). From Priest to Doctor: Domain Adaptation for Low-Resource Neural Machine Translation. *arXiv preprint arXiv:2412.00966*.

[12] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

[13] Ojo, J., and Ogueji, K. (2023). How Good Are Commercial Large Language Models on African Languages? *arXiv preprint arXiv:2305.06530*.

[14] Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 489–500.

[15] Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10, pp. 522–538.

[16] Elmadani, K. N., Meyer, F., and Buys, J. (2022). University of Cape Town's WMT22 System: Multilingual Machine Translation for Southern African Languages. *arXiv preprint arXiv:2210.11757*.

[17] Eberhard, D. M., Simons, G. F., and Fennig, C. D. (Eds.). (2022). *Ethnologue: Languages of Africa and Europe, Twenty-Fifth Edition*. SIL International.

[18] Hu, J., Xia, M., Neubig, G., and Carbonell, J. (2019). Domain Adaptation of Neural Machine Translation by Lexicon Induction.

[19] Popović, M. (2017). chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pp. 612–618.

[20] Wang, J., et al. (2024). AfriMTE and AfriCOMET: Enhancing COMET to Embrace Under-Resourced African Languages. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.