Improving Neural Machine Translation for isiXhosa in the Medical Domain using Forward- and Back-Translation

Nick Matzopoulos
mtznic006@myuct.ac.za
Department of Computer Science, University of Cape Town
Cape Town, South Africa

Abstract

Machine translation (MT) has advanced significantly by making use of neural architectures and transformer models, yet these improvements remain unevenly distributed across languages. Low-resource languages such as isiXhosa continue to face substantial challenges due to limited parallel corpora. The challenges for isiXhosa are heightened when tasked with domain-specific translation, such as the medical domain, as medical parallel corpora are even more scarce. This paper investigates whether generating synthetic data through forward/back-translation can help overcome the shortage of real parallel medical text, and hence equip models to translate medical text more reliably. To investigate this, this study fine-tuned pre-trained NLLB-200 models using both forward-translation and back-translation on the monolingual English PriMock57 corpus. The results show that synthetic data generation significantly improves translation quality on in-domain medical tests. Forward-translation achieved +2.08 BLEU improvement, while back-translation showed more substantial gains of +7.59 BLEU. To test whether these improvements scale with model capacity, experiments with the larger NLLB-200 1.3B revealed that back-translation remains robust across model sizes, while forward-translation faces challenges with larger models. Notably, the models also demonstrated performance gains on general-domain text, suggesting that in-domain medical fine-tuning enhances overall translation capability rather than narrowing it. When compared to alternate synthetic data approaches, both forward- and back-translation outperformed DALI. While the forward- and back-translation approaches focused on fine-tuning with relatively small amounts of in-domain data, the WMT22-NLLB model, which fine-tuned NLLB-200 on large out-of-domain data, showed catastrophic forgetting, demonstrating that data relevance matters more than volume. Category-specific evaluation of medical terminology revealed that the back-translation model reduced overall Health Term Error Rates from 26.28% to 19.87%. The back-translation models performed particularly well in the Medical Conditions category, reducing error rates from 59.02% to 39.34%, but performed worse than the base model in the Anatomy category. This confirms that forward/back-translation is an effective method for bridging the data gap in low-resource English-isiXhosa medical translation.

1 Introduction

In South Africa's linguistically diverse landscape, a communication gap exists in the medical domain. IsiXhosa is the second-most common home language in the country, spoken as a first language by more than eight million people [25], and by many more as a second language. Despite this widespread use, most medical services operate exclusively in English. When medical staff and patients do not share a language, important medical information can be misunderstood or lost, which increases the risk of misdiagnosis or incorrect treatment [5]. A reliable MT system for English–isiXhosa medical text could help bridge this gap and improve access to quality healthcare for millions of South Africans.

Machine translation systems have evolved through multiple frameworks over the past decades. From early rule-based methods that rely on handcrafted linguistic rules [13], to data-driven statistical approaches [13], and most recently, neural architectures using attention mechanisms [26]. The introduction of attention-based models [2] and the Transformer architecture [26] has led to significant improvements in translation quality for different language pairs. Each framework brought improvements in translation quality but also introduced new challenges, especially for languages with limited parallel data [15].

Large multilingual models like NLLB-200 now support more than 200 languages, including isiXhosa, and demonstrate impressive performance on general-domain text [6]. NLLB's high performance on general-domain text can be attributed to it being trained on a large collection of web-scraped text and their translations, along with Wikipedia articles, and Bible translations [6]. However, these large multilingual models face challenges when translating domain-specific terminology. As a result, domain-specific data is essential to improve translation performance [15].

The main obstacle in building reliable medical translation systems for isiXhosa is the lack of large parallel corpora [18]. The situation becomes even more challenging due to isiXhosa's complex morphological structure. It is an agglutinative language, meaning that words are built by attaching multiple subwords together. This makes accurate translation difficult even with adequate training data [18].

General English–isiXhosa parallel datasets have become more widely available through initiatives like the WMT22 shared task [1], but domain-specific medical corpora remained scarce until recently. Resources like MeMaT [14] and recent work by Blocker et al. [3] have begun to address this gap by creating specialized English–isiXhosa medical datasets. The Blocker et al. dataset, consisting of mock doctor–patient consultations, contains 581 sentence pairs, providing medical domain bilingual data [3]. This size is insufficient for training models, but it is effective as an evaluation dataset to assess medical translation performance.

We address this data scarcity problem by investigating two data augmentation techniques: forward-translation and back-translation [24]. Data augmentation techniques, particularly back-translation, have shown promise in domain adaptation for neural machine translation (NMT) [28]. This study uses these methods to fine-tune the

NLLB-200 model and then measures the performance of these models on medical datasets, MeMaT Medical [14] and Blocker, as well as a general-domain dataset like FLORES-200 [6]. This indicates the model's performance on both medical and general-domain translation tasks. In this paper, unless otherwise stated, "NLLB-200" refers to the distilled 600M-parameter model. We also evaluate a larger NLLB-200 model with 1.3B parameters alongside the distilled 600M version. This investigates whether the gains from forward- and back-translation scale with model capacity, and whether scaling alone can match those gains.

Both forward- and back-translation approaches significantly improved performance over the base NLLB models, with back-translation showing particularly strong gains of +7.59 BLEU on the Blocker medical test set. The effectiveness varies across model sizes. Medical terminology evaluation, using a Health Term Error Rate (HTER) metric, revealed substantial improvements. Notably, the back-translation NLLB-1.3B model was on par with commercial systems overall and surpassed ChatGPT on anatomy terms. These findings confirm that synthetic data generation, through forward- and back-translation, can effectively bridge the gap created by the scarcity of real medical parallel corpora, making domain-specific translation more accessible for low-resource languages like isiXhosa.

2 Related Work

2.1 Neural Machine Translation for isiXhosa

isiXhosa presents unique challenges for NMT. The language has complex morphological structures and limited parallel training data compared to high-resource languages like English. The WMT22 shared task helped establish benchmarks for African languages, including isiXhosa, and highlighted both progress and ongoing challenges [1]. Researchers have responded to these challenges and have come up with effective strategies for improving isiXhosa NMT through multilingual training and transfer learning.

Multilingual training has proven particularly effective [18]. Instead of training models on isiXhosa alone, researchers combine it with related languages to enable cross-lingual transfer, allowing the model to learn shared representations across languages. One of the earliest examples of this is "MMTAfrica" [9], in which a model was trained on six African languages, as well as English and French. This approach worked because related languages often have similar grammar patterns and vocabulary roots, so the model can learn from all of them at once [9]. Nyoni and Bassett [18] showed similar results with their three-language model covering English, isiXhosa, and isiZulu. Their system achieved BLEU score improvements of +9.9 points over single-language baselines [18].

Transfer learning is another solution. Instead of starting from scratch, researchers take large pre-trained models like NLLB-200 and fine-tune them on isiXhosa data. This works because the pre-trained models already understand basic translation patterns, which means they can adapt to isiXhosa's specific features [6]. The approach works best when the original model includes related African languages from the same family, since they share similar word structures and grammar rules [6]. These two strategies have become standard methods for building better isiXhosa translation systems when parallel data is limited.

2.2 Domain Adaptation in NMT

Domain adaptation refers to the process of specializing a general-purpose translation model for specific domains [28]. While modern NMT systems achieve strong results on general-domain text, their performance worsens when faced with domain-specific terminology [28]. Now the main approaches to domain adaptation for isiXhosa NMT will be reviewed: fine-tuning on in-domain data, and synthetic data generation via back-translation and forward-translation.

2.2.1 Fine-tuning on In-Domain Data. Fine-tuning represents the standard approach when specialized parallel data is available [28]. The process involves taking a pre-trained model and continuing training on a smaller, domain-specific dataset [18]. This allows the model to adapt its parameters to handle the specific vocabulary, syntax, and semantic patterns of the target domain [28].

For isiXhosa, fine-tuning has proven particularly effective when starting from multilingual models [18]. Research shows that fine-tuning large pre-trained models on isiXhosa data consistently outperforms training models from scratch [18]. In medical domains, the typical process is to pre-train on general-domain data and then fine-tune on whatever small in-domain isiXhosa corpus is available. This two-step process allows the model to leverage both general-language knowledge and domain-specific knowledge [18].

2.2.2 Synthetic Data Generation. When domain-specific parallel data is scarce or unavailable, synthetic data generation becomes essential for domain adaptation [24]. Two approaches for creating artificial parallel corpora, which leverage existing models and monolingual data, allowing fine-tuning even when real bilingual corpora are limited, are discussed below.

Back-translation is one of the most effective techniques for generating synthetic training data in low-resource scenarios [8]. The process starts with monolingual target-language text and translates it back into the source language using an existing model, creating synthetic training pairs where the target side contains real, natural sentences while the source side is synthetic. The key advantage is that models learn to produce accurate target-language output, since they are trained on real examples of how the target language should look.

Forward-translation takes monolingual source-language sentences and translates them directly into the target language using an existing model [29]. The process begins with authentic source-language text, which is translated into synthetic target-language versions to create parallel training pairs. This approach is useful when high-quality, domain-specific source-language data is available. However, the main limitation is that synthetic target-language quality depends entirely on the initial model's translation ability [8]. If the base model produces poor target-language translations, these errors will be reinforced during training.

Research suggests that back-translation typically produces better results than forward-translation for low-resource languages [4]. This is because back-translation trains models to produce real, correct target-language sentences, while forward-translation relies on potentially flawed synthetic target text. However, both approaches offer valuable ways to synthetically generate parallel data for domain adaptation.

Dataset	Data Type	Size
PriMock57	Monolingual English	57 consultations
FLORES-200 Dev Set	Bilingual pairs	997
Blocker Test Set	Bilingual pairs	581
MeMaT-Medical	Bilingual pairs	181

Table 1: Dataset composition showing data type and size for each source.

3 Experimental Setup

3.1 Datasets

We used several datasets containing both monolingual English medical text and English-isiXhosa text pairs across general and medical domains. The data comes from four main sources: PriMock57 [16], the Blocker dataset, MeMaT, and FLORES-200. The dataset sizes are detailed in Table1.

3.1.1 Training Data. The PriMock57 dataset contains 57 mock doctor-patient consultations originally collected in English [16]. These consultations simulate real medical interactions between healthcare professionals and patients in controlled settings. For this study, 47 consultations (5,598 monolingual sentences) were used to generate synthetic training data through forward- and back-translation processes. The remaining 10 consultations (581 sentence pairs) form the Blocker Dataset described below [3]. These consultations serve as evaluation data and are strictly reserved for testing purposes only, to ensure unbiased evaluation of model performance.

3.1.2 Validation Data. FLORES-200 serves as a standard benchmark for multilingual MT evaluation [11]. This general-domain dataset contains high-quality translations across 200 languages, including English-isiXhosa pairs. The dataset covers diverse text types, including news articles, instruction manuals, and website content [6]. For this evaluation, the development set containing 997 sentence pairs was used as our first validation set. This allowed measurement of how medical-domain adaptation affected performance on general-domain translation tasks.

The MeMaT project produced several parallel corpora focused on different domains, including a specialized medical dataset with English-isiXhosa text pairs [14]. The isiXhosa translations were produced by native isiXhosa speakers, ensuring accurate and human-like target-language text. The MeMaT medical dataset contains only 181 English sentences paired with 181 isiXhosa translations. While this is a relatively small dataset, because it is a medical-domain dataset, this dual-validation approach allowed us to investigate the trade-off between optimizing for domain relevance versus validation set size in model selection.

3.1.3 Test Data. The Blocker dataset was the main resource used when evaluating medical translation quality [3]. This dataset was adapted from PriMock57 and contains 581 English medical sentences paired with 581 corresponding isiXhosa translations. The isiXhosa translations were created by professional human translators. This dataset was used exclusively as the test set to evaluate the final model performance for the medical domain.

3.2 Models

In implementing the forward- and back-translation processes, the NLLB-200 models served as the foundation of this project, providing three distinct roles. First, the base models were used as baselines for evaluation so that improvements could be measured against an existing model for multilingual translation. Second, they were used to generate synthetic data from monolingual text, which was later organized into forward- and back-translation pairs to account for the limited supply of real parallel medical text. Finally, the base models were fine-tuned on the synthetic data to create domain-specific models specialized for English–isiXhosa medical translation. Each of these roles is described in turn below.

Two NLLB-200 variants were used as baselines: the distilled 600M parameter model and the larger 1.3B version [6]. Both models were downloaded from Hugging Face without any modifications to their original weights or configuration. The reason for choosing two NLLB variants was to study the effect of model size on the forward- and back-translation processes. Both models were evaluated on all validation sets and the Blocker test set to establish benchmarks for comparing fine-tuned model performance.

The NLLB models were also used to create synthetic parallel corpora from monolingual text. As illustrated in Figure 1, which lists all stages of the process (Figure 1a–1f), it begins with monolingual English medical text from the 47 PriMock57 consultations, as shown in Figure 1a. This monolingual text is processed by the base NLLB-200 model (Figure 1b), specifically using the English→isiXhosa direction to produce synthetic isiXhosa translations (Figure 1c). These stages (Figure 1a–1c) are the same for both the forward- and back-translation processes. The distinction lies at the pairing stage (Figure 1d), where the ordering of the pairs is reversed, as discussed below.

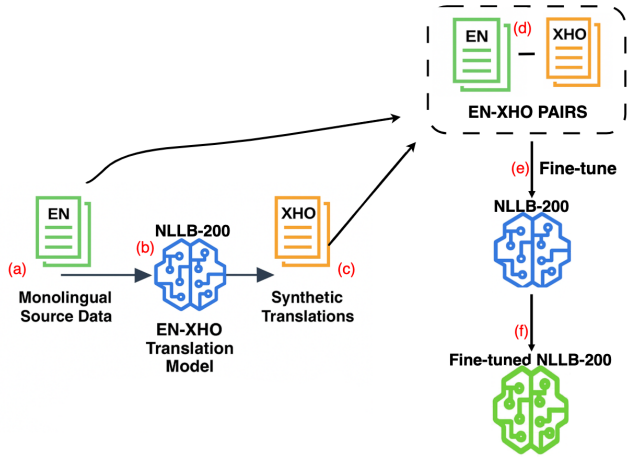


Figure 1: Forward-translation process. Stages (a–f) are annotated in the figure.

Finally, the NLLB models were used as the base models that underwent fine-tuning on the synthetically generated data, with distinct processes for forward- and back-translation. For forward-translation, at Figure 1d the synthetic outputs and original English

Model Base	Set	Model	English → isiXhosa			isiXhosa → English		
			chrF++	BLEU	ACoM	chrF++	BLEU	ACoM
NLLB 600M	FLORES-200	Base-NLLB	46.56	14.19	0.67	55.46	34.01	0.65
		FLORES-Best	47.01 (**+0.45**)	15.85 (**+1.66**)	0.69 (**+0.02**)	56.93 (**+1.47**)	35.72 (**+1.71**)	0.60 (-0.05)
	MeMaT	Base-NLLB	41.54	7.97	0.97	39.29	21.07	0.78
		MeMaT-Best	43.84 (**+2.30**)	8.96 (**+0.99**)	0.84 (-0.13)	41.03 (**+1.74**)	24.58 (**+3.51**)	0.80 (**+0.02**)
NLLB 1.3B	FLORES-200	Base-NLLB	47.31	14.81	0.67	58.49	37.97	0.76
		FLORES-Best	47.73 (**+0.42**)	15.41 (**+0.60**)	0.68 (**+0.01**)	59.44 (**+0.95**)	39.25 (**+1.28**)	0.77 (**+0.01**)
	MeMaT	Base-NLLB	42.99	10.38	0.84	43.36	25.12	0.78
		MeMaT-Best	43.74 (**+0.75**)	9.28 (-1.10)	0.84 (±0.00)	44.08 (**+0.72**)	27.80 (**+2.68**)	0.80 (**+0.02**)

Table 2: Validation sets results: Comparison of base and fine-tuned NLLB-600M and NLLB-1.3B models across FLORES-200 and MeMaT validation sets. Numbers in brackets indicate improvements over the respective base models for each model size.

are paired as (real English, synthetic isiXhosa). These pairs are then used as training data for the English→isiXhosa NLLB-200 model, as shown in Figure 1e. This produces a fine-tuned English→isiXhosa NLLB-200 model (Figure 1f). For back-translation, in Figure 1d the same synthetic outputs are paired as (synthetic isiXhosa, real English) and used as training data for the isiXhosa→English NLLB-200 model, resulting in a fine-tuned isiXhosa→English NLLB-200 model. This ordering is a key advantage of back-translation, since having real English sentences on the target side provides the model with a more reliable training signal [4].

3.3 Hyperparameter Tuning

Hyperparameter tuning becomes especially critical in low-resource settings, where models are more sensitive to training configurations due to limited data availability [30]. Small datasets amplify the impact of hyperparameter choices, making systematic optimization essential for achieving reliable performance [30].

Many different hyperparameter combinations were tested to find the best training settings for the models. Thorough searches for both forward- and back-translation were run to make sure it found the top-performing configurations for each approach.

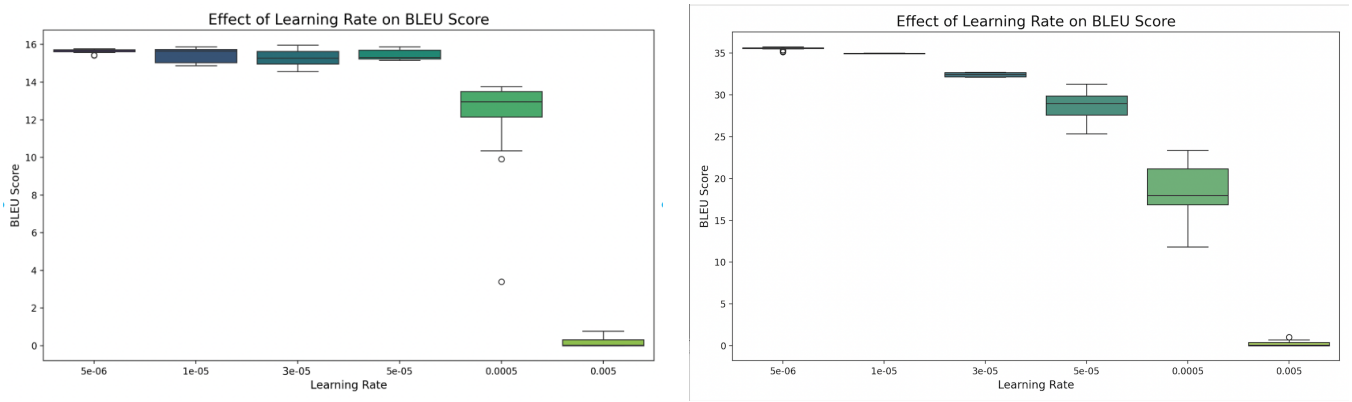
Approximately 160 different configurations for each model were explored, systematically looping through ranges of learning rates, weight decay settings, batch sizes, warmup ratios, label smoothing values[1], and training epochs. Due to limited computational resources, the forward- and back-translation NLLB-200 1.3B models underwent more limited hyperparameter sweeps than the 600M runs. The hyperparameter grids for the 600M- and 1.3B-based model are detailed in Table 3.

The hyperparameter sweeps revealed distinct learning rate sensitivity patterns between translation directions, as shown in Figure 2. Small learning rates from 5×10^{-6} to 1×10^{-5} produce narrow, high-performing boxes in both plots, indicating stable training across different hyperparameter combinations. Back-translation showed greater sensitivity to changes in learning rate, shown by the downward sloping trend in BLEU scores. The box plots show that forward-translation maintains stable performance across a broader range, extending up to 5×10^{-5}, while back-translation performance begins to degrade earlier. Decreases in BLEU scores become visible around 3×10^{-5} to 5×10^{-5} for both directions, but back-translation shows this degradation more prominently. Performance degrades sharply at 5×10^{-4}, where the boxes become wider and lower with increased variance, and collapses almost completely at 5×10^{-3}. This suggests optimal learning rate ranges of 5×10^{-6} to 1×10^{-5} for back-translation and 5×10^{-6} to 5×10^{-5} for forward-translation, with both directions showing learning rate thresholds, where training becomes unstable, above 5×10^{-4}.

Hyperparameter	NLLB-600M Search Space	NLLB-1.3B Search Space
Forward-Translation (English → isiXhosa)		
Learning rate	{$\underline{5\times10^{-6}}$, **$1\times10^{-5}$**, 3×10^{-5}, 5×10^{-5}, 5×10^{-4}, 5×10^{-3}}	{**$\underline{5\times10^{-6}}$**, 1×10^{-5}}
Effective batch size	{32, **$\underline{64}$**}	{**$\underline{64}$**}
Epochs	{3, **5**, $\underline{7}$}	{**5**, $\underline{7}$}
Label smoothing	{0.0, **$\underline{0.1}$**}	{**0.0**, $\underline{0.1}$}
Warmup ratio	{$\underline{0.0}$, **0.1**}	{**$\underline{0.0}$**, 0.1}
Weight decay	{**$\underline{0.0}$**, 0.01}	{**$\underline{0.0}$**, 0.01}
Back-Translation (isiXhosa → English)		
Learning rate	{**$\underline{5\times10^{-6}}$**, 1×10^{-5}, 3×10^{-5}, 5×10^{-5}, 5×10^{-4}, 5×10^{-3}}	{**$\underline{5\times10^{-6}}$**, 1×10^{-5}}
Effective batch size	{32, **$\underline{64}$**}	{**$\underline{64}$**}
Epochs	{3, **5**, $\underline{7}$}	{**5**, $\underline{7}$}
Label smoothing	{**$\underline{0.1}$**}	{$\underline{0.0}$, **0.1**}
Warmup ratio	{**$\underline{0.0}$**, 0.1}	{0.0, **$\underline{0.1}$**}
Weight decay	{0.0, **$\underline{0.01}$**}	{0.0, **$\underline{0.01}$**}

Table 3: Combined hyperparameter search spaces for fine-tuning NLLB-600M and NLLB-1.3B models. Winning hyperparameters are bolded (FLORES-Best) and underlined (MeMat-Best).

[1]Label smoothing was fixed at 0.1 for the back-translation sweep, as a preliminary smaller sweep indicated it performed best.



(a) Forward-Translation (EN→XH) **(b) Back-Translation (XH→EN)**

Figure 2: Effect of learning rate on BLEU score for forward (left) and back translation (right). Each box shows the distribution of BLEU scores across different hyperparameter settings at a fixed learning rate.

After these sweeps, a dual-validation approach was used to explore the trade-off between general-domain and medical performance. Each of the 160 trained models was tested on two different validation sets: the FLORES-200 dev set (general-domain text) and the MeMaT medical set (smaller, in-domain dataset). This allowed the two best models for each translation direction to be selected: a "FLORES-Best" model (highest BLEU on general text) and a "MeMaT-Best" model (highest BLEU on medical text). This dual-validation approach tested how choosing models based on general accuracy versus medical accuracy would influence final performance on the Blocker medical test set.

The best model configurations are shown in Table 4. The results reveal clear patterns across both translation directions. Small learning rates worked best, 5×10^{-6} was optimal for three out of four models, with only FLORES forward-translation preferring the slightly higher 1×10^{-5}.

A batch size of 64 consistently outperformed 32 across all models. Label smoothing at 0.1 beat beat 0.0. Weight decay showed notable direction-specific preferences. Forward-translation models performed best with no weight decay (0.0), while back-translation models benefited from a small amount (0.01). Training duration varied by validation set: FLORES-selected models converged at 5 epochs, while MeMaT-selected models needed 7 epochs for optimal performance.

Model	LR	Epochs	WR	WD	LS
NLLB-200 600M Models					
FLORES-Best (Fwd)	1×10^{-5}	5	0.1	0.0	0.1
MeMaT-Best (Fwd)	5×10^{-6}	7	0.0	0.0	0.1
FLORES-Best (Bck)	5×10^{-6}	5	0.0	0.01	0.1
MeMaT-Best (Bck)	5×10^{-6}	7	0.0	0.01	0.1
NLLB-200 1.3B Models					
FLORES-Best (Fwd)	5×10^{-6}	5	0.0	0.0	0.1
MeMaT-Best (Fwd)	5×10^{-6}	7	0.0	0.01	0.1
FLORES-Best (Bck)	5×10^{-6}	5	0.1	0.0	0.1
MeMaT-Best (Bck)	5×10^{-6}	7	0.1	0.01	0.1

LR = Learning Rate, WR = Warmup Ratio, WD = Weight Decay, LS = Label Smoothing.

Table 4: Best-model hyperparameter configurations for the NLLB-200 based models.

3.4 Baselines

We also compare the forward- and back-translation models against several alternative approaches, two of which were carried out by team member Malibongwe Makhonza. This involved training a Transformer model from scratch using the WMT22 English–isiXhosa parallel dataset, and fine-tuning the pre-trained NLLB-200 model on the same WMT22 data. Both approaches focus on general-domain adaptation rather than medical specialization.

Another team member, Elijah Sherman implemented DALI (Domain Adaptation by Lexicon Induction) [12], an alternative synthetic data generation method to forward- and back-translation. DALI combines dictionary building with domain-specific corpora [12]. First, he extracted a bilingual dictionary from general English–isiXhosa parallel data using fast-align [7] and paired it with PriMock57. Using the dictionary, he mapped key English medical terms from PriMock57 to their isiXhosa equivalents and back-translated the sentences, creating a pseudo-parallel medical corpus [12]. This synthetic corpus was used to fine-tune translation models in both directions (English isiXhosa), allowing the model to handle medical terminology that was absent in the original training data [17].

Blocker Baseline Models: The original study by Blocker et al. established benchmarks using Google Cloud Translate [10] for English-to-isiXhosa translation and ChatGPT [19] with a modified prompt for isiXhosa-to-English translation [3]. The ChatGPT approach modified standard translation prompts by including medical term pairs directly in the prompt context, providing the model with a domain-specific terminology guide for each translation.

The Blocker study reported results for several systems, including Google Cloud Translate and ChatGPT with a modified prompt [3]. We include these two as commercial baselines. Google Cloud Translate is a production NMT service, while ChatGPT is a large language model (LLM) system. While not strictly comparable, as we do not know their training data or adaptation methods, it helps draw conclusions as to how our models perform compared to the best commercial options.

3.5 Evaluation Protocol

Consistent evaluation across all models is crucial. This requires a standardized approach that ensures that results are reproducible. The same evaluation pipeline was applied to all models in this work for consistency.

The evaluation followed a two-step process for each model on each evaluation set. First, a Python script was used to produce translated text. The script loads a specified model and translates source files line by line. To maximize translation quality and maintain consistency, we generated all translations using beam search with five beams and a maximum generation length of 128 tokens. The generated translations were saved as text files.

The next step calculates the performance metrics using a separate verification script. This takes the model-generated translations and compares them against the validation/test set translations. This leads to the computation of four different evaluation metrics. BLEU[20] and chrF++[21] are computed using SacreBLEU [22]. We also calculate AfriCOMET [27] and a Health Term Error Rate (HTER) from Blocker et al.[3]. These metrics capture different aspects of translation quality and are detailed below.

In selecting the best models, the top models on both the FLORES-200 development set and the MeMaT validation set were considered. For both forward- and back-translation, these were FLORES (FLORES-Fwd-Best and FLORES-Bck-Best) and on MeMaT (MeMaT-Fwd-Best and MeMaT-Bck-Best). For each NLLB system, these four top models, together with the base models, were all evaluated on the Blocker test set. They were evaluated alongside team members' models, larger NLLB models, and two commercial baseline models reported in the Blocker study.

The evaluation metrics were chosen specifically to account for isiXhosa's morphological complexity. While BLEU correlates well with human judgments for many languages, it computes word-level n-gram overlap and therefore struggles on morphologically complex, agglutinative languages like isiXhosa [27]. Character-based metrics are more accurate for such languages, hence chrF++ was employed [21]. Due to its operating at character and sub-word levels, chrF++ is well suited to isiXhosa, where English words may translate to subwords rather than complete word units. Together, BLEU and chrF++ capture lexical accuracy and word-level overlap.

We also used AfriCOMET for evaluation, which is an adaptation of COMET [23]. This uses a pre-trained language model to provide a metric for MT evaluation. COMET was designed to provide consistent and reliable correlations with human judgments for a variety of languages. AfriCOMET extends COMET by making use of multilingual pre-training and fine-tuning on African language data [27]. This makes the metric focused on meaning preservation and fluency.

In addition to the above metrics, we also measured how well the models handled key medical vocabulary. For this, an evaluation method from \citet{blocker2025benchmarking} was used. It includes a list of 68 medical terms grouped into three categories: Anatomy, Condition, and Treatment. A script was used that checks for exact matches of these terms in both the generated translations and the human-translated files. From this, a HTER was calculated, giving a measure of how reliably each model translated medical terms. This terminology-based evaluation was run on each model's Blocker test set translations, and the results are shown in Table 6. HTER uses an exact-match terminology assessment, and so it is useful for assessing translation of medical terms, but it is blind to other aspects of quality, like semantics. However, when used alongside BLEU, chrF++, and AfriCOMET, a well-rounded evaluation is provided.

4 Results and Analysis

This section analyzes the performance of the best models, which were selected using either FLORES-200 or MeMaT validation sets, when evaluated on the Blocker medical test set. The study compares these results against the NLLB base models, team members' models, and two commercial baseline models.

4.1 Does Forward/Back-translation Improve Performance?

4.1.1 Comparison Against Base NLLB-200 600M. Both forward- and back-translation models improved on the base NLLB-200 600M model on the Blocker test set, as shown in Table 5. The biggest improvements were in the isiXhosa→English direction, where the FLORES-Best model achieved +7.59 BLEU and +5.55 chrF++ improvements. AfriCOMET scores showed slight degradation, with the MeMaT-Best model achieving a score of 0.75 compared to the base model's 0.77, which is an indication of a small reduction in meaning preservation and fluency. For the English→isiXhosa direction, the FLORES-Best model achieved more modest improvements of +2.08 BLEU and +1.53 chrF++. For both the forward- and back-translation models, gains in BLEU and chrF++ show improvements at the word and subword level. AfriCOMET scores improved substantially from 0.27 to 0.67 for the FLORES-Best model, indicating better semantic accuracy and fluency.

The fact that back-translation experienced larger performance gains is due to it translating from a low-resource language (isiXhosa) to a high-resource language (English), which is easier than the reverse direction [4]. A likely reason for this is data imbalance: NLLB was pre-trained on far more English than isiXhosa, so it is better at producing English on the target side [6].

Notably, fine-tuning did not lead to overfitting on the medical domain. Performance gains on the general-domain FLORES-200 validation set, as shown in Table 2, indicate that training on medical data improved the model's ability to handle non-medical sentences. For back-translation, the FLORES-Best model improved chrF++ by +1.47 and BLEU by +1.71 over the base model. This suggests the fine-tuning process improved not just medical keyword translation.

4.1.2 Scaling to NLLB-200 1.3B. To test whether forward- and back-translation gains scale with model capacity, we evaluated the NLLB-1.3B-based models, which use the same fine-tuning approach as the 600M models, against the respective base models. The results

		English → isiXhosa			isiXhosa → English		
Model Category	**Model**	**chrF++**	**BLEU**	**ACoM**	**chrF++**	**BLEU**	**ACoM**
NLLB-200 600M	Base-NLLB	42.27	15.78	0.27	42.50	26.87	**0.77**
	FLORES-Best	**43.80 (+1.53)**	**17.86 (+2.08)**	0.67 (+0.40)	**48.05 (+5.55)**	34.46 (+7.59)	0.71 (−0.06)
	MeMaT-Best	42.57 (+0.30)	17.53 (+1.75)	**0.81 (+0.54)**	47.97 (+5.47)	**34.48 (+7.61)**	0.75 (−0.02)
	DALI	42.08 (−0.19)	9.72 (−6.06)	0.35 (+0.08)	39.13 (−3.37)	19.03 (−7.84)	0.56 (−0.21)
	WMT22-NLLB	3.21 (−39.06)	2.08 (−13.70)	0.66 (+0.39)	26.49 (−16.01)	5.83 (−21.04)	0.77 (0.00)
NLLB-200 1.3B	Base-NLLB	**44.86**	**17.70**	**0.87**	46.02	30.07	**0.77**
	FLORES-Best	42.76 (-2.10)	15.61 (-2.09)	0.65 (-0.22)	50.37 (+4.35)	**36.07 (+6.00)**	0.73 (-0.04)
	MeMaT-Best	42.60 (-2.26)	15.60 (-2.10)	0.70 (-0.17)	**50.41 (+4.39)**	35.98 (+5.91)	0.75 (-0.02)
Commercial	Google Cloud Trans	63.79	28.40	–	57.23	28.60	–
	ChatGPT (mod)	52.38	11.40	–	57.59	26.70	–

Table 5: Comparison of NLLB-200 600M and 1.3B models against commercial baselines on the Blocker medical test set. Bold values show the best-performing fine-tuned model within each category. Numbers in brackets indicate improvements/degradation over the respective base model for that category.

in Table 5 shows how model size affects the forward- and back-translation processes.

The NLLB-200 1.3B base model performs better than the 600M version across all metrics. For forward-translation, the 1.3B base model achieved 17.70 BLEU compared to 15.78 for the 600M model. For back-translation, the 1.3B base model reached 30.07 BLEU versus 26.87 for the 600M model. This confirms that larger model capacity on its own can improve medical translation quality.

Fine-tuning the larger model produced different results. For the forward-translation process, the 1.3B model degraded performance. BLEU dropped from 17.70 to 15.61, and chrF++ fell from 44.86 to 42.76. Although the 1.3B model generated its own synthetic data, likely better than the 600M's, isiXhosa is morphologically complex and low-resource. And so we believe the synthetic data that it produced still showed recurring linguistic errors, especially in morphology and word formation. The reason for the degradation is inconclusive. Two potential reasons include overfitting to noisy synthetic targets, as well as the limited hyperparameter sweep for the 1.3B models, which may have left better configurations unexplored.

For the back-translation process, the 1.3B model improved results. BLEU rose from 30.07 to 36.07 and chrF++ from 46.02 to 50.37. Here, the target side contains real English sentences. English is a high-resource language and has large amounts of training data, so it provides a clean learning signal. Even if the synthetic isiXhosa source sentences contain noise, the model can still learn reliably from the real English targets. This explains why the larger model shows clear improvements, matching the trends seen in the 600M model. It is also important to note that for both directions the 1.3B fine-tuned models saw similar general-domain gains, with BLEU, chrF++, and AfriCOMET improving over the 1.3B base model, as shown in Table 2.

These results demonstrate that forward- and back-translation effectiveness depends on both model capacity and training data quality. Back-translation remains robust across model sizes because it relies on real target data, whilst forward-translation loses effectiveness at larger scales.

4.1.3 Comparison Against Baselines. To contextualize the results, the forward- and back-translation models are compared against two baselines implemented by team members. One is the WMT22-NLLB model, an alternative fine-tuning approach that fine-tuned the NLLB-200 600M on the WMT22 dataset. The other is DALI, an alternative method for generating synthetic data. The full results are shown in Table 5.

The WMT22-NLLB model shows the cost of out-of-domain fine-tuning. Despite training on a large parallel corpus, the WMT22-NLLB model performed worse on the Blocker test set than the NLLB base model. Its English → isiXhosa chrF++ score collapsed from 42.27 to 3.21, while isiXhosa → English dropped to 26.49 compared to the base model of 42.50. This shows the concept of catastrophic forgetting, where fine-tuning on large out-of-domain data almost completely erased the model's ability to handle medical language. This emphasizes that for domain-specific fine-tuning, the relevance of the fine-tuning data is more important than its size.

For English → isiXhosa, the DALI model achieved 9.72 BLEU compared to our FLORES-Best model's 17.86 BLEU. For isiXhosa → English, the FLORES-Best model outperformed the DALI model, achieving a chrF++ of 48.05 compared to DALI's 39.13. For isiXhosa → English, the AfriCOMET score of 0.56 was considerably worse than the FLORES-Best model's score of 0.71, showing that the back- and forward-translation models not only performed better on a lexical level, shown by BLEU and chrF++ improvements, but also on a semantic level. This weaker performance of DALI, relative to forward- and back-translation, is likely due to the quality of the synthetic data it produces. In contrast to DALI's dictionary-based substitution method, the forward/back-translation process generates synthetic isiXhosa using an existing translation model. As a result, forward/back-translation proves to be a stronger approach to synthetic data generation for this language pair, leading to models that can handle medical translation more effectively.

4.1.4 Comparison Against Commercial LLM and NMT Systems. To provide a broader context, the forward- and back-translation models are also compared against state-of-the-art commercial translation

systems, including Google Cloud Translate and ChatGPT with modified prompts. These systems epresent strong commercial baselines for English-isiXhosa translation. The results are detailed in Table 5.

Google Cloud Translate is the top performer for English → isiXhosa, achieving 63.79 chrF++ compared to our NLLB-200 1.3B-based FLORES-Best model of 42.76. For isiXhosa → English, ChatGPT with modified prompts achieved chrF++ of 57.59 compared to 50.37 for our FLORES-Best model. Again, even though these commercial systems outperform the FLORES-Best models, the results provide clear evidence that forward/back-translation of medical domain sentences significantly improves medical English-isiXhosa translation. This shows that synthetic data generation can effectively compensate for the scarce nature of real parallel medical data.

4.2 Medical Terminology Results

Other than overlap-based evaluation metrics, the study conducted evaluation focusing on medical vocabulary. This terminology-based assessment measures how accurately models translate key medical terms across three categories: Anatomy, Condition, and Treatment. A Health Term Error Rate is calculated using string-match comparisons between generated and reference translations. The results are shown in Table 6.

The fine-tuned NLLB-200 600M models excelled at isiXhosa → English translation, reducing the overall error rate from 26.28% to 19.87%. The models outperformed the base model across every category, with the most noticeable improvement in medical conditions, where the MeMaT-Best model reduced error rates from 59.02% to 39.34%.

For English → isiXhosa translation, while overall improvement was more modest (45.57% to 44.30%), significant gains appeared in the Condition and Treatment categories. Interestingly, anatomy terms proved more challenging for the fine-tuned models, with the base models slightly outperforming them. This is likely because anatomy terms, like arm, leg, and hand, are more common in the general-domain text that NLLB is trained on, unlike condition or treatment terms, which require medical specialization.

4.2.1 Comparison Against Commercial Systems and NLLB-200 1.3B models. Google Cloud Translate is the top English → isiXhosa performer with an overall HTER of 18.99%, while ChatGPT with modified prompts achieved a very low isiXhosa → English HTER of 11.82%. Even though these baseline systems outperform the forward- and back-translation models, the results provide sufficient evidence to conclude that forward/back-translation of medical domain sentences significantly improves medical English-isiXhosa translation.

The NLLB-200 1.3B-based fine-tuned models show promising progress toward commercial-level performance, particularly for isiXhosa → English translation. The fine-tuned 1.3B model achieved an overall HTER of 17.95%, approaching the performance of ChatGPT. More notably, the 1.3B fine-tuned model actually outperformed ChatGPT on anatomy terms. This is significant because ChatGPT was identified as the top performer for HTER for isiXhosa → English translation in the original Blocker study. While commercial systems still have higher performance overall, the performance of the 1.3B fine-tuned models show that scaling NLLB capacity while applying forward/back-translation can bring open-source models closer to commercial levels for medical terminology.

4.3 Validation Set Considerations

The dual-validation strategy allows direct comparison of how different model selection approaches affect final performance. It looks at whether optimizing for general-domain accuracy (FLORES-Best) or medical-domain accuracy (MeMaT-Best) leads to better results on the Blocker medical test set, and hence which is more appropriate as a validation set for this setting.

Across three out of the four evaluation metrics, for the 600M-based models, the FLORES-Best models consistently outperformed the MeMaT-Best models on the Blocker test set. In the HTER analysis, the FLORES-Best achieved a 44.30% error rate compared to MeMaT-Best's 45.99%. The pattern held for isiXhosa → English, where FLORES-Best reached 19.87% versus MeMaT-Bck-Best's 20.19%. The trend also applied to the BLEU and chrF++ metrics accross both translation directions. This suggests that FLORES-selected models produce better lexical and term-level accuracy, which is crucial for medical translation where precise terminology matters. These same trends in FLORES-MeMaT model performance applied to the 1.3B-based models.

However, AfriCOMET scores show a different pattern. For English → isiXhosa, the MeMaT-Best model achieved 0.81 compared to FLORES-Best's 0.67. Similarly, for isiXhosa → English, MeMaT-Best reached higher AfriCOMET scores than FLORES-Best. This suggests that MeMaT-selected models may better preserve the meaning of the sentence and read more naturally. AfriCOMET appears to capture semantic improvements that BLEU, chrF++, and HTER do not detect.

BLEU, chrF++, and HTER suggest that FLORES-based model selection works better for optimizing terminology correctness for medical translation. However, AfriCOMET indicates that MeMaT selection may produce better semantic quality. Human evaluation would be needed to decide whether a large, general-domain validation set (FLORES) or a small, domain-specific validation set (MeMaT) better predicts medical translation quality on Blocker.

4.4 Qualitative Analysis

Beyond the quantitative metrics, a detailed examination of actual translation outputs reveals how forward- and back-translation affects medical translation quality (for the NLLB-200 600M-based models). The analysis focuses on specific medical terminology and error patterns, with complete examples provided in Tables 7 and 8.

We observe several examples where the back-translation model generates more accurate translations of medical terms from isiXhosa → English, as detailed in Table 7. Notable examples appear in condition-related terminology: where the base model translated nausea (-caphucaphu) as "angry," the back-translation model correctly identified "nausea."[2] Similarly, the base model confused "cough" (kohlokhohlo) with "rhythm" and mistranslated "high blood pressure" (nehihi ephezulu) as "high fructose corn." We also observe

[2]Note that as a non-isiXhosa speaker, the translations in the tables were provided by Google Translate and verified by human isiXhosa speakers, but may not be 100% accurate as they were not provided by professional translators.

	English → isiXhosa				isiXhosa → English			
Model	**Overall**	**Anatomy**	**Condition**	**Treatment**	**Overall**	**Anatomy**	**Condition**	**Treatment**
NLLB-600M Baseline	45.57%	32.19%	76.56%	44.44%	26.28%	15.25%	59.02%	42.86%
FLORES-Best	44.30%	32.88%	**71.88%**	**40.74%**	19.87%	11.21%	42.62%	**39.29%**
MeMaT-Best	45.99%	34.93%	73.44%	**40.74%**	20.19%	12.56%	39.34%	**39.29%**
NLLB-1.3B Baseline	**43.88%**	**30.82%**	75.00%	**40.74%**	20.19%	12.56%	39.34%	**39.29%**
FLORES-Best	44.73%	33.56%	**71.88%**	**40.74%**	**17.95%**	**10.76%**	34.43%	**39.29%**
MeMaT-Best	44.80%	33.98%	72.34%	**40.74%**	18.42%	11.11%	**33.20%**	**39.29%**
Google Trans	18.99%	14.91%	38.04%	4.55%	14.44%	11.23%	30.71%	0.00%
GPT(mod)	38.14%	32.55%	64.12%	17.39%	11.82%	10.87%	20.71%	0.00%

Table 6: Health Term Error Rate for translations on Blocker dataset.

Medical Term (isiXhosa → English)	**Source (isiXhosa)**	**Reference (English)**	**Baseline Model**	**Fine-Tuned Model**
-caphucaphu (Nausea)	"...uziva unes**icaphucaphu**..."	"...do you feel any **nausea**..."	"...do you feel **angry**..."	"...do you feel any **nausea**..."
-khohlokhohlo (Cough)	"Akukho **kohlokhohlo**..."	"No **cough**..."	"Does your chest have a **rhythm**?"	"No **cough-ing**..."
nehihi ephezulu (High Blood Pressure)	"...kunye **nehihi ephezulu**."	"...and **high blood pressure**."	"...and **high fructose corn**."	"...and **high blood pressure**."
-liso (Eye)	"...isemva kwel**iso**?"	"...is it behind the **eye**?"	[Omitted]	"...is it behind the **eye**?"
-imilenze (Legs)	"...em**milenze**ni..."	"...on your **legs**..."	"...in your **feet**..."	"...in your **legs**..."
unokuxinana kwesayinasi (Sinus Congestion)	"...**unokuxinana kwesayinasi**?"	"...do you have **sinus congestion**?"	"You also say you have a fever."	"So, do you have any **sinus congestion**?"

Table 7: Comparison of Medical Term Translation (isiXhosa → English) for the NLLB-200 600M-based models. Bold text highlights key subwords, their incorrect translations, and their correct translations.

that the back-translation model handled anatomy-related terminology more reliably, correctly translating the terms for "eye" and "legs," which were either incorrectly translated or omitted by the base model. Notably, no examples were found where the base model correctly translated a medical term that the back-translation model missed.

For English → isiXhosa translation, Table 8 shows that the most significant improvements appeared in anatomy-related terminology. We observe cases where the base model made semantic errors, such as translating the term "nose" (impumlo) with the idiomatic term for "gut feeling" or "inner state" (imbilini), while our fine-tuned model correctly identified the term. Similarly, in multiple cases, the base model mistranslated the term "migraine" as "nerve disease" or "neurological disease" (isifo semithambo-luvo). The fine-tuned model did not make this error. In one case, the base model simply omitted the translation of "intloko", meaning "head", whilst the forward-translation model translated this correctly. However, we also observe instances where the model incorrectly translates terms. For example, when translating "clots in the legs," it replaced the correct isiXhosa term for "clots" (amahlwili) with "germs" (iintsholongwane).

The qualitative analysis provides insight into how forward- and back-translation affects medical translation quality, particularly for anatomy- and condition-related terms. These examples highlight the specific ways in which forward- and back-translation can improve translation accuracy while also revealing areas where errors still occur.

5 Future Work

Future work should explore training on mixed synthetic and real data, which could provide more robust training signals by combining the volume benefits of synthetic data with the authentic

Medical Term	Source (English)	Reference (isiXhosa)	Baseline Model	Fine-Tuned Model
Migraines (**-migraine**)	"...have any **migraines**..."	"...one-**migraine**..."	"...onesifo **semithambo**-luvo..."	"...onesifo se-**migraine**..."
Head (**intloko**)	"...have my **head**..."	"...**intloko** yam..."	[Omitted]	"...**intloko** yam..."
Nose (**-mpumlo**)	"...my **nose** is ok."	"...**impumlo** zilungile."	"...**imbilini** yam ilungile."	"...**impumlo** yam ilungile."

Table 8: Comparison of Medical Term Translation (English → isiXhosa) for the NLLB-200 600M-based models. Bold text highlights root subwords in isiXhosa terms.

linguistic and semantic patterns of real parallel corpora. Additionally, including general-domain data into the training mix could address the anatomy term performance degradation, since terms like "arm" and "hand" appear frequently in everyday text.

6 Conclusions

This study addressed whether forward- and back-translation of medical domain sentences improves medical English-isiXhosa translation. The results provide a clear answer.

Both forward-translation and back-translation effectively adapted NLLB-200 600M to the medical domain. Back-translation showed particularly strong improvements, increasing chrF++ and BLEU significantly whilst reducing HTER. Forward-translation achieved more modest gains, with chrF++ improvement and error rate reductions. However, the introduction of AfriCOMET metrics revealed important nuances in translation quality. While back-translation showed slight AfriCOMET degradation, forward-translation demonstrated substantial semantic improvements. These results confirm that fine-tuning on synthetically generated medical data via forward- and back-translation are viable methods for creating models that can effectively and correctly handle the translation of medical terminology. This shows that domain-specific improvements are achievable without real parallel corpora.

The scaling experiment with NLLB-200 1.3B revealed that model size affects domain adaptation differently depending on translation direction. Back-translation remained robust across model sizes, with the 1.3B model achieving even stronger improvements. However, forward-translation degraded performance on the larger model, possibly due to overfitting on noisy synthetic targets. This demonstrates that forward- and back-translation effectiveness depends on model capacity.

It is also important to note that domain specialization didn't worsen general translation capability. In fact, our fine-tuned models improved performance not only on medical translation but also on general-domain text, suggesting medical training improved the models' sentence translating abilities in general. However, it is important to note certain trade-offs. While the forward- and back-translation process resulted in improvements on condition and treatment translation, it slightly reduced anatomy term performance, so the model actually performs worse on certain medical-critical terms. This shows that domain adaptation needs careful consideration of which capabilities to prioritize for real-world deployment.

This study showed that the forward- and back-translation processes appear to be more effective in creating a model that can effectively and correctly handle the translation of medical terminology, for this language pair, than the DALI process. We also identified that when facing data scarcity, training data relevance proves far more important than size, as seen by the catastrophic forgetting we observed with the WMT22-NLLB model. Notably, the 1.3B fine-tuned models showed promising progress toward commercial-level performance, particularly for isiXhosa → English translation.

The validation set analysis provided insights into model selection strategies. FLORES-Best models consistently outperformed MeMaT-Best models on lexical metrics, suggesting that general-domain validation better optimizes medical terminology correctness. However, MeMaT-Best models achieved higher AfriCOMET scores, indicating better semantic preservation. This suggests that different validation approaches optimize for different aspects of translation quality.

References

[1] David I. Adelani, D. Whitenack, Graham Neubig, et al. 2022. Findings of the WMT'22 Shared Task on African Language Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT 2022)*. 773–800.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs.CL]

[3] Avin Blocker, Fémi Meyer, Sakhile Biyabani, Joyce Mwangama, Mmasibidi I. Datay, and Bonga Malila. 2025. Benchmarking isiXhosa Automatic Speech Recognition and Machine Translation for Digital Health Provision. In *Proceedings of the Workshop on Patient-oriented Language Processing*.

[4] Nikolay Bogoychev and Kamal Chowdhury. 2019. Domain-specific MT for low-resource languages. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. European Association for Machine Translation, 78–86.

[5] B. Clark, A. Schreuder, and C. Mathews. 2020. Language barriers and their impact on health-care delivery in South Africa. *South African Medical Journal* 110, 8 (2020), 780–784.

[6] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672 [cs.CL]

[7] Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 644–648.

[8] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 489–500.

[9] C. C. Emezue, B. F. P. Dossou, and the Masakhane Team. 2021. MMTAfrica: Multilingual Machine Translation for African Languages. In *Proceedings of the Sixth Conference on Machine Translation (WMT 2021)*. 398–411.

[10] Google Cloud. 2025. Cloud Translation API. https://cloud.google.com/translate/docs

[11] N. Goyal, C. Gao, V. Chaudhary, P. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics* 10 (2022), 522–538.

[12] J. Hu, M. Xia, G. Neubig, and J. Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1363–1373.

[13] J. Huang, J. Li, and H. Xue. 2020. A Comparative Study of RBMT, SMT, and NMT for Low-Resource Domain Adaptation. *Information* 11, 5 (2020), 259.

[14] C. Maria Keet, Zola Mahlaza, Kenneth Heafield, Alexandra Birch, Palash Pal, and Nosithembele Kwebulana. 2018. MeMaT: Medical machine translation corpus (isiXhosa–English). https://github.com/mkeet/MeMaT Corpus developed by the University of Cape Town and the University of Edinburgh. CC-BY license..

[15] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*. 28–39.

[16] A. P. Korfiatis, F. Moramarco, R. Sarac, and A. Savkov. 2022. Primock57: A dataset of primary care mock consultations. arXiv:2204.00333 [cs.CL]

[17] A. Marashian, E. Rice, L. Gessler, A. Palmer, and K. von der Wense. 2021. From Priest to Doctor: Domain Adaptation for Low-Resource Neural Machine Translation. arXiv:2109.08833 [cs.CL] Corrected from incomplete entry..

[18] E. Nyoni and B. A. Bassett. 2021. Low-Resource Neural Machine Translation for Southern African Languages. arXiv:2104.00366 [cs.CL]

[19] OpenAI. 2025. ChatGPT. https://openai.com/chatgpt

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[21] Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*. 612–618.

[22] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT '18)*. Association for Computational Linguistics, Brussels, Belgium, 186–191. doi:10.18653/v1/W18-6319

[23] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. arXiv:2009.09025 [cs.CL]

[24] D. Saunders. 2021. Domain adaptation for neural machine translation. arXiv:2104.06951 [cs.CL]

[25] Statistics South Africa. 2022. *Census 2022: Statistical release p0301.4*. Technical Report. Statistics South Africa. https://www.statssa.gov.za/publications/P03014/P030142022.pdf

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5998–6008.

[27] Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, et al. 2024. AfriMTE and AfriCOMET: Enhancing COMET to Embrace Under-resourced African Languages. arXiv:2311.09828 [cs.CL]

[28] Y. Wang, Z. Chen, and K. Liu. 2022. Domain Adaptation in Neural Machine Translation: A Survey. *Comput. Surveys* 55, 3 (2022), 1–34.

[29] J. Zhang and C. Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1535–1545.

[30] Xuan Zhang and Kevin Duh. 2020. Reproducible and Efficient Benchmarks for Hyperparameter Optimization of Neural Machine Translation Systems. *Transactions of the Association for Computational Linguistics* 8 (2020), 393–408.