IsiXhosa Machine Translation in the Medical Domain

Elijah Sherman
shreli006@myuct.ac.za
Department of Computer Science,
University of Cape Town
Cape Town, South Africa

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

Malibongwe Makhonza
mkhmal024@myuct.ac.za
Department of Computer Science,
University of Cape Town
Cape Town, South Africa

ABSTRACT

Machine translation (MT) systems for low-resource languages like isiXhosa are still a big challenge because there aren't many parallel datasets available for training models. Current MT systems work well when translating between isiXhosa and English for more common words and sentences, but performance drops when faced with specialised medical terms. This project will examine whether using synthetic data created either through back-translation or via lexicon induction can improve overall translation quality, help fill vocabulary gaps, and make the model more robust when handling medical terms. To determine this we will create three MT models, each following a different approach, that will be evaluated using standard quality metrics (BLEU, chrF, and chrF++). The three models that we will be developing in this project are compared to each other as well as the Blocker evaluation dataset. These answers will provide new insights into domain-specific adaptation, the feasibility for using MT for isiXhosa in the medical domain, and improving MT for isiXhosa.

1 INTRODUCTION

South Africa is linguistically diverse. 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, most medical services operate 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 machine translation (MT) system for English–isiXhosa medical text would help reduce these risks and improve access to quality care.

Neural machine translation has improved significantly in recent years, especially after the introduction of attention-based models [2] and the Transformer architecture [26]. Large multilingual models like NLLB-200 now support more than 200 languages, including isiXhosa [6]. However, MT systems often perform poorly when the available training data is limited or when the text belongs to a specialised domain. Medical language includes technical terms, abbreviations, and culture-specific expressions that are not well covered in general-domain training data. As a result, domain-specific data is essential to improve translation performance [12].

While datasets for general English–isiXhosa translation have become more widely available in recent years (such as the WMT22 dataset [1]), domain-specific data for the medical field remains rare. Resources like the Blocker et al. dataset [3] have started to address this gap by creating specialised English–isiXhosa parallel corpora focused on medical communication. The Blocker et al. dataset, for example, contains 580 doctor–patient consultations, and will be used in this project to evaluate translation quality in the medical domain.

In this project, we will compare three approaches for building English–isiXhosa medical translation models. The first approach involves building two baseline models: one using a pre-trained multilingual model (NLLB-200) that we fine-tune using the general-domain WMT22 dataset, and another trained from scratch using the same dataset. The second approach uses back-translation to create synthetic medical-domain data from monolingual English medical text. The third approach applies DALI (Domain Adaptation by Lexicon Induction), which builds a bilingual lexicon and uses it to generate pseudo-parallel sentence pairs [10]. We will then evaluate the performance of each of these models on their ability to translate medical text between isiXhosa and English.

The rest of this proposal outlines the background and motivation for our project, presents our research questions, and explains our methodology, including how we train and evaluate each model. We also discuss ethical considerations, risks, and the resources required to carry out the project.

2 BACKGROUND

2.1 Neural Machine Translation for isiXhosa

Neural Machine Translation (NMT) has improved the quality of automatic translation systems by using deep learning and attention mechanisms [2]. The Transformer model is now the most widely used architecture, as it handles longer sentences and trains faster than earlier methods [26]. However, NMT models require large amounts of parallel data to perform well, and their performance often degrades sharply in low-resource settings if data is limited [11]. For high-resource languages like French or German, this data is usually available but for low-resource languages like isiXhosa, the lack of large parallel corpora remains a major barrier [16].

Some progress has been made through multilingual models such as NLLB-200 [6], which was trained on over 200 languages. It includes isiXhosa, but its performance is limited because the amount of isiXhosa data in the training set is still small. Similarly, large language models (LLMs) like ChatGPT and LLaMA have shown strong results for high-resource languages, but they struggle with African languages [17]. Robinson et al. (2023) found that these models often leave out or hallucinate isiXhosa terms [22]. This shows the importance of building or adapting models that are specialised for isiXhosa instead of relying on generic translation tools.

Recent shared tasks like the WMT22 have helped expand the amount of parallel data available for African languages, including isiXhosa [1]. WMT (Workshop on Machine Translation) is an annual competition where research groups from around the world build and evaluate machine translation systems using the same public datasets and evaluation methods. The 2022 edition included a focus on low-resource African languages, with English–isiXhosa being one of the supported language pairs. Adelani et al. (2022) built a multilingual model for African languages and improved performance using synthetic data and back-translation techniques. These datasets and shared benchmarks are especially valuable because they provide a standard way to compare models and build stronger baselines.

2.2 Domain Adaptation for Translation

Even when a model is trained on a large general-domain dataset, like WMT22, it may not perform well when translating text from a specialised domain like medicine. This is because medical language uses technical terms, abbreviations, and formal phrasing that are very different from everyday language. Domain adaptation refers to techniques that help a translation model adjust to the style and vocabulary of a specific domain.

One of the most common techniques is fine-tuning a pre-existing model on in-domain data [12]. Another widely used method is back-translation, where a monolingual in-domain corpus in the source language (English), such as PriMock57 [13], is translated into the target language (isiXhosa) using an existing model. The resulting pseudo-parallel sentence pairs are then added to the training set to help the model learn domain-specific vocabulary. This method is especially helpful when there is not enough real bilingual data available.

Another useful method is DALI (Domain Adaptation by Lexicon Induction) [10]. It uses a bilingual dictionary to replace words in monolingual domain text with translations from the source language, creating pseudo-parallel sentences. This approach is useful when there is not enough real medical translation data available.

A recent study by Marashian et al. (2025) evaluated domain adaptation methods for medical translation in truly low-resource settings. Their experiments showed that DALI outperformed other approaches like continual pre-training and dictionary-guided decoding [14]. Even though it is one of the simplest methods, DALI achieved the best performance across languages like Croatian and Maltese. This confirms that low-complexity, data-centric methods like DALI are still very effective for domain adaptation in low-resource settings, especially in healthcare.

Our project builds on these approaches by applying back-translation and DALI to adapt general English–isiXhosa models to the medical domain, where accuracy and correct terminology are especially important.

3 RESEARCH QUESTIONS

Clear communication is essential for good medical care. English is the main language of medical practice in South Africa, yet many patients communicate more easily in isiXhosa [25]. Current MT systems handle everyday language well, but their performance worsens when dealing with specialised medical terms, abbreviations, or culturally specific expressions [12]. Errors in translation can lead to patients being misdiagnosed and poor adherence to treatment due to confusion [5].

Until recently, the medical domain lacked parallel data for English–isiXhosa. Blocker et al. (2025) have started to fill this gap, providing a small but high-quality corpus of doctor–patient consultations [3]. Below are the specific research questions:

1. **If we only have general-domain data available, does adapting a massively multilingual model for English–isiXhosa translation outperform training from scratch on isiXhosa medical MT?**
   This question examines whether building a model from scratch can lead to better translation accuracy for isiXhosa medical terminology compared to adapting an already existing multilingual model.
   Hypothesis: The model based on a pre-trained massively multilingual model that is fine-tuned for isiXhosa translation will perform better than a model trained from scratch.

2. **Does back-translation of medical domain sentences improve medical English-isiXhosa translation?**
   This question tests whether creating synthetic training data through back-translation can compensate for the scarcity of authentic English–isiXhosa medical domain data, and if this approach creates a model that can effectively and correctly handle translation of medical terminology.
   Hypothesis: Models trained on synthetic data generated using back-translation will perform better than models trained on scarce, real English–isiXhosa medical domain pairs for medical translation.

3. **Does DALI-generated synthetic data of medical domain sentences improve medical English–isiXhosa translation?**
   This question tests whether lexicon-based domain adaptation methods can make up for the lack of medical-domain bilingual data, and whether they can help the model learn and correctly translate specialised medical vocabulary.
   Hypothesis: Models trained on synthetic data generated using DALI will perform better than models trained on scarce, real English–isiXhosa medical domain pairs for medical translation.

4 PROCEDURES AND METHODS

4.1 Data

This section outlines the datasets used for training and evaluation of our models, including how the datasets were created and which ones will be used for each of our models.

We will be using the dataset created by Blocker et al. (2025) to evaluate our models, referring to it as the Blocker evaluation dataset. This dataset was adapted from the PriMock57 dataset [13] which contains mock doctor-patient consultations. Blocker et al. (2025) adapted this dataset by translating and recording selected consultations in both languages using South African actors.. The result of this is 580 texts in both languages [3].

The baseline model will be trained on a large general domain English-isiXhosa dataset, such as the WMT22 dataset [15]. Both the synthetic data model and the DALI model will be trained using the rest of the data from the PriMock57 dataset [13], i.e. unused PriMock57 consultations. The synthetic model uses NLLB [9] to translate the consultations to isiXhosa. The DALI model will use word-to-word mapping to build the lexicon, drawing information from several sources: the remaining PriMock57 data, a general bilingual isiXhosa-English dictionary, and a bilingual medical dictionary.

4.2 Models

4.2.1 *Baseline Model.*

We will use two different approaches to build our baseline models. The first approach uses a pre-trained multilingual translation model called NLLB-200 (No Language Left Behind) [6]. This model, developed by Meta and available through Hugging Face, supports over 200 languages including isiXhosa and has shown strong performance even for low-resource languages. We will use the NLLB-200 model for English–isiXhosa translation and fine-tune it using the WMT22 general-domain English–isiXhosa dataset, which contains an extensive amount of parallel sentences from diverse sources including government websites, religious texts, and news articles [1]. Fine-tuning on this general-domain corpus allows us to evaluate how well a high-resource, multilingual model can handle isiXhosa translation before any domain-specific adaptation. The resulting model will be able to perform translation in both directions: English ↔ isiXhosa. This approach will then be evaluated and compared to our scratch baseline approach in order to address our research question of whether adapting a massively multilingual model fine-tuned only on general-domain data can outperform models trained from scratch when applied to specialised isiXhosa medical translation tasks.

The second approach involves training a machine translation model entirely from scratch using the same WMT22 dataset. Unlike the pre-trained NLLB model, this model will be initialised with random weights and trained end-to-end only on the English–isiXhosa data from WMT22. This will allow us to observe the impact of multilingual pre-training versus training a dedicated model from a large, high-quality parallel corpus. As with the first baseline, this model will support translation in both directions (English ↔ isiXhosa), enabling it to handle a range of communication scenarios in the medical domain.

4.2.2 *Synthetic Data Generation Model.*

Modern neural MT systems require large, domain-specific parallel corpora, but English–isiXhosa medical data are scarce [1]. This limited amount of data leads to poor translations that omit content or leave English text untranslated [21]. Recent work has shown that while MT performance is improving, low-resource languages like isiXhosa lag behind high-resourced languages [1].

To bridge this gap, synthetic data generation methods were developed. Synthetic data generation creates synthetic parallel data which can be used to train and fine-tune models [23]. It is an effective way to adapt models to a specific domain [8]. This is done by using monolingual in-domain data - in our case, English medical data from the PriMock57 dataset - to produce parallel examples for training [8]. Instead of relying on in-domain parallel corpora, which is extremely rare, we train or fine-tune models on synthetically generated data [8]. To create this synthetic data for fine-tuning, we will be creating two machine-translation models: isiXhosa → English, trained using back-translation (BT), and English → isiXhosa, trained with forward translation (FT).

Back-Translation (isiXhosa → English). Back-translation is a widely used method for data augmentation in low-resource machine translation [24]. It uses monolingual target language data and translates it "back" into the source language using a reverse model [23].

For BT (isiXhosa → English), we start with English (the target language) and translate it into isiXhosa (the source language), using the NLLB-200 multilingual model [6], specifying English → isiXhosa as the translation direction. This model generates an isiXhosa translation of the real English data from the PriMock57 dataset. The resulting isiXhosa translation is synthetic and referred to as isiXhosa'.

Once we have these synthetic translations, we create an (isiXhosa' - real English) medical domain pair. This pair is used as parallel data to fine-tune an isiXhosa → English model. We will be fine-tuning NLLB-200, specifically the isiXhosa → English direction.

Forward Translation (English → isiXhosa). Forward translation reverses the approach: we take monolingual source language text and translate it into the target language to create synthetic pairs [27].

For forward translation, we will again use a pre-trained model for English → isiXhosa translation. We will use the NLLB-200 multilingual model [6], specifying English → isiXhosa as the translation direction. This model translates real English medical data from PriMock57 into synthetic isiXhosa (isiXhosa').

This creates a (real English - isiXhosa') medical domain pair, which we then use as parallel data to fine-tune our original forward translation English → isiXhosa model.

In both BT and FT cases, we make use of the PriMock57 English dataset to create (isiXhosa' - English) and (English - isiXhosa') pairs respectively. However, the difference in pairing order is significant and leads to differences in performance. Research has shown that BT usually has better performance than FT for low-resource languages [4]. This is because in BT, the model is trained to produce real, correct English sentences because the target side (English) contains correct English examples. While FT uses output that is potentially flawed, as the target side (isiXhosa') is synthetic and dependent on the correctness of our pre-trained English → isiXhosa model [4]. While BT has its advantages, we will be implementing both methods to assess their effectiveness and performance for English-isiXhosa medical domain translation, which can be compared to our baseline models.

4.2.3 *Model Trained using the DALI Approach.*

Another synthetic data generation technique is Domain Adaptation by Lexicon Induction (DALI). DALI uses in-domain monolingual source data with out-of-domain parallel data to create an in-domain lexicon [10].

The lexicon is then used to generate a pseudo-parallel corpus for training the MT model [10]. In terms of this project, we will use English source data from the medical domain and out-of-domain isiXhosa-English data (such as a bilingual dictionary extracted using fast-align) to create the lexicon. There are various methods which can be used to build the bilingual dictionary such as fast-align [7]which was used in the original DALI paper [10]. To ensure lexicon quality, we will manually review a sample of the most frequent mappings and remove any that are incorrect or not relevant to the medical domain. We will use the lexicon to translate English medical sentences into isiXhosa through word-to-word mapping, creating a pseudo-parallel medical domain corpus [14]. Using this corpus, we will train a bilingual MT model capable of translating in both directions: English ↔ isiXhosa.

This approach was chosen as it has been shown to perform well when adapting to specific domains, specifically for low-resource languages [14]. DALI has the potential to outperform models that use back-translation in terms of accuracy as you can precisely control which words are mapped to each other.

4.3 Evaluation

The two baseline models will be compared to each other; the back-translation and DALI models will each be compared to the baseline models, and all models will be evaluated against the Blocker dataset. The models will be compared using three metrics: BLEU, chrF, and chrF++.

BLEU automatically evaluates MT quality by comparing it to professional human translations [18]. The similarity of the translations is done by calculating the n-gram matches between the candidate and reference translations [18]. BLEU scores often correlate well with human evaluations.

chrF, based on character n-grams, often outperforms BLEU in correlating with human judgements [19]. This improvement in performance is likely due to the fact that this evaluation technique is done at the character level and is therefore more sensitive than the BLEU metric. chrF balances precision and recall over character n-grams in both the reference translation and the MT [19].

chrF++ is an adaptation of chrF that includes word n-grams as well as character n-grams which are averaged to get the chrF++ score [20]. It has been shown that for sentences receiving higher scores from human evaluators, chrF scores correlate more strongly with human judgements than wordF scores (the word n-gram F-scores) [20]. This contrast—where wordF tends to be more pessimistic—led to the idea that combining chrF and wordF could improve the overall correlation with human evaluations, which led to chrF++. chrF and chrF++ are better for isiXhosa, where translations often involve sub-words.

We will calculate all three metrics and compare them to each other and to Blocker et al. (2025). Blocker et al. evaluate seven varied models; we will only compare those relevant to ours. For example, we will compare NLLB to our baseline model, but not necessarily to DALI due to architectural differences.

Since we are focusing on the medical domain we will also be evaluating how well each of our models translates medical terminology in particular. We follow the evaluation method used by Blocker et al. (2025). This will be done by using an error rate. The error rate will be calculated as: Error rate = 100 minus the percentage of correctly translated medical terms (output/reference) [3].

5 ETHICAL, PROFESSIONAL AND LEGAL ISSUES

The Blocker evaluation dataset [3] is open source meaning we can use it for our purposes without having to obtain the rights to use it. The dataset was created from scripted, synthetic doctor-patient consultations written by researchers [13]. These consultations were performed by actors, recorded and manually transcribed. These consultations do not represent any real patient or any sensitive information.

The models we have proposed will not be made publicly available for real world use when this project is complete. The purpose of this study is only to explore different approaches, not to develop a commercially available software.

6 PROJECT PLAN

6.1 Research Contributions

After this project has been completed we will have provided answers to our research questions. These answers will provide new insights into domain-specific adaptation, the feasibility for using MT in the medical domain and improving MT for isiXhosa. The research community will be able to take our findings and use them as a stepping stone. They will see what parts of our research work and what parts do not which gives them guidance on how to move forward when doing further research.

6.2 Expected Impact

This project will advance the research done on MT for isiXhosa in the medical domain which will lead to an improvement in the abilities of doctors and patients in South Africa to communicate more effectively. We expect to see promising results that improve the current state of MT for isiXhosa in the medical domain as it has been shown that when models are fine-tuned on a specific domain they perform better than models that are not trained on any specific domain [10].

After the project is completed it creates the potential for it to be adapted for commercial use so that doctors can use it as a tool to allow them to communicate with patients who only speak isiXhosa or are more comfortable speaking isiXhosa than English.

6.3 Risk and Risk Management Strategies

See Appendix A.

6.4 Timeline

See Appendix B.

6.5 Resources Required

To execute this project, we need specific computational, data, and human resources, detailed below.

Hardware and Computing. Access to the Centre for High Performance Computing (CHPC) GPU cluster is needed for training and fine-tuning. Computers are required for data preparation and cleaning, implementing the models, and evaluating the results.

Software and Tools. PyTorch or TensorFlow will be used with Hugging Face Transformers to build, fine-tune, and experiment with our models. The Meta NLLB-200 MT model [9] will be used as a pre-trained model that we will be fine-tuning. Version control via GitHub will track code changes and allow team contribution to a shared code-base. BLEU, chrF, and chrF++ will be used for evaluation.

Datasets. The Blocker evaluation dataset [3] is needed for evaluating our models. The WMT22 dataset, specifically the English–isiXhosa parallel subset, and the remainder of the PriMock57 dataset will be used for training and fine-tuning our models.

People Required. The team consists of three members, each focusing on an independent model (baseline, back-translation, and DALI). Work allocation is detailed in Section 6.8. Our supervisor will guide us on methodology and project direction.

6.6 Deliverables

Literature Review Deliverables. The **Literature Review Scaffold** will be used to plan and break down past work. The **Final Literature Review** will analyse prior work on isiXhosa MT in the medical domain, including domain adaptation methods. Gaps and opportunities for improvement in isiXhosa MT are discussed.

Project Proposal Deliverables. The **Project Proposal Draft** outlines the project aims, presents background research, and states research questions and hypotheses. It specifies the methodology, models, and datasets, and addresses ethical and legal considerations along with the project plan. The **Project Proposal Presentation** will be delivered, and feedback incorporated to produce the **Final Project Proposal**.

Final Project Deliverables. The **Final Paper Draft** will provide background, analysis and evaluation of model results. Feedback will be incorporated to produce the **Final Paper Submission**. The **Final Code Submission** will include code for all three models. The **Final Project Demonstration** will present and analyse results and discuss the overall project. A **Poster** and **Website** will present research findings and describe methodology and results.

Machine Translation Systems. As part of the deliverables, three English–isiXhosa NMT models will be produced: the **baseline model**, the **back-translation model**, and the **DALI model**.

6.7 Milestones

(See Appendix B for the full Gantt chart.)

- Literature review completion: 28 Mar 2025
- Project proposal final submission: 06 May 2025
- Methods implementation completion: 10 Aug 2025
- Evaluation completion: 21 Aug 2025
- Honours project completion: 27 Oct 2025

6.8 Work Allocation

6.8.1 Malibongwe Makhonza – Baseline Methods

Approach 1: MT model from scratch

Phase 1: Define a Transformer architecture suitable for low-resource training. Ensure that the WMT22 dataset is cleaned and formatted. This entails removing duplicate or misaligned sentence pairs, and making sure punctuation is correct and consistent. Thereafter, set up training scripts and book compute time on the CHPC cluster.

Phase 2: Train the model from scratch on the WMT22 dataset. Then, save intermediate and final models.

Phase 3: Experiment and test the models with different hyper-parameters while collecting results. Compare the results with the NLLB-baseline model and record findings.

Approach 2: Fine-tuned NLLB

Phase 1: Load the NLLB-200 model from Hugging Face and confirm it runs correctly. Ensure that the WMT22 dataset is cleaned and formatted. This entails removing duplicate or misaligned sentence pairs, and making sure punctuation is correct and consistent where necessary. Thereafter, set up training scripts and book compute time on the CHPC cluster.

Phase 2: Fine-tune the NLLB model on the WMT22 dataset. Save intermediate and final models.

Phase 3: Experiment and test the models with different hyper-parameters while collecting results. Compare the results with the scratch-baseline model and record findings.

6.8.2 Nick Matzopoulos – Back-translation Implementation

Approach 1: Back-translation (isiXhosa → English)

Phase 1: Load the NLLB isiXhosa → English model from Hugging Face and make sure it works as expected. Clean and format the PriMock57 dataset to remove speaker tags (e.g., “Doctor:” or “Patient:”) and incorrect formatting. This makes sure that every line is a complete sentence, providing high-quality English input to the English→isiXhosa model, which generates the synthetic isiXhosa used to fine-tune our isiXhosa→English model.

Phase 2: Input the high-quality English data into the NLLB English → isiXhosa model. This generates synthetic isiXhosa text, creating (isiXhosa, English) pairs.

Phase 3: Use the synthetic (isiXhosa, English) data to fine-tune the isiXhosa → English NLLB model. Monitor performance using metrics. Test the model with different hyper-parameters and save the intermediate and final trained model. Evaluate performance against the baseline approaches.

Approach 2: Forward-translation (English → isiXhosa)

Phase 1: Load the NLLB English → isiXhosa model from Hugging Face and confirm it runs correctly. Use the PriMock57 English corpus cleaned and formatted in Phase 1 of the back-translation pipeline.

Phase 2: Input the cleaned English data into the NLLB English → isiXhosa model to generate synthetic isiXhosa text, creating (English, isiXhosa) pairs.

Phase 3: Use the synthetic (English, isiXhosa) dataset to fine-tune the English → isiXhosa NLLB model. Monitor performance using metrics. Test the model with different hyper-parameters and save the intermediate and final trained model. Evaluate performance against the baseline approaches.

6.8.3 Elijah Sherman – DALI Implementation

Phase 1: Extract a bilingual dictionary from the out-of-domain parallel English–isiXhosa data using fast-align [7]. Collect and clean English monolingual data from the medical domain (PriMock57). Set up the DALI pipeline as described in [10]. Book compute time on the CHPC cluster.

Phase 2: Use the bilingual dictionary to create word-to-word mappings between English medical terms and their isiXhosa equivalents. Apply these mappings to back-translate the English medical sentences into isiXhosa, generating a pseudo-parallel corpus of (English, isiXhosa) pairs. Then, refine the lexicon as needed.

Phase 3: Use the synthetic pseudo-parallel corpus to train or fine-tune a bilingual MT model capable of translating in both directions: English ↔ isiXhosa. Experiment with different hyper-parameters, save intermediate and final models, and evaluate performance against the baseline approaches.

REFERENCES

[1] Adelani, D. I., Whitenack, D., Neubig, G., et al. Findings of the wmt'22 shared task on african language translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT 2022)* (2022), pp. 773–800.

[2] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2015).

[3] Blocker, A., Meyer, F., Biyabani, A., Mwangama, J., Datay, M. I., and Malila, B. Benchmarking isixhosa automatic speech recognition and machine translation for digital health provision. In *Proceedings of the Workshop on Patient-oriented Language Processing* (2025).

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

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

[6] Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. No language left behind: Scaling human-centered machine translation. urlhttps://arxiv.org/abs/2207.04672, 2022. arXiv:2207.04672.

[7] Dyer, C., Chahuneau, V., and Smith, N. A. 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* (Atlanta, Georgia, June 2013), L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds., Association for Computational Linguistics, pp. 644–648.

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

[9] Face, H. facebook/nllb-200-distilled-600m. https://huggingface.co/facebook/nllb-200-distilled-600M, 2025.

[10] Hu, J., Xia, M., Neubig, G., and Carbonell, J. Domain adaptation of neural machine translation by lexicon induction, 2019.

[11] Huang, J., Li, J., and Xue, H. A comparative study of rbmt, smt, and nmt for low-resource domain adaptation. *Information 11*, 5 (2020), 259.

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

[13] Korfiatis, A. P., Moramarco, F., Sarac, R., and Savkov, A. Primock57: A dataset of primary care mock consultations. *arXiv preprint arXiv:2204.00333* (2022).

[14] Marashian, A., Rice, E., Gessler, L., Palmer, A., and von der Wense, K. From priest to doctor: Domain adaptaion for low-resource neural machine translation. *arXiv preprint arXiv:2412.00966* (2024).

[15] NLLB Team, et al. Wmt22 african languages parallel corpus. https://huggingface.co/datasets/allenai/wmt22_african, 2022. Dataset hosted on Hugging Face Datasets.

[16] Nyoni, E., and Bassett, B. A. Low-resource neural machine translation for southern african languages. *arXiv preprint arXiv:2104.05579* (2021).

[17] Ojo, J., and Ogueji, K. How good are commercial large language models on african languages? *arXiv preprint arXiv:2305.06530* (2023).

[18] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (2002), pp. 311–318.

[19] Popović, M. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation* (2015), pp. 392–395.

[20] Popović, M. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation* (2017), pp. 612–618.

[21] Restack, A. Quantitative analysis of english–isixhosa parallel datasets. *Journal of Low-Resource MT 12*, 2 (2023), 1–12.

[22] Robinson, N., Ogayo, P., Mortensen, D. R., and Neubig, G. Chatgpt mt: Competitive for high- (but not low-) resource languages. *arXiv preprint arXiv:2309.07423* (2023).

[23] Saunders, D. Domain adaptation for neural machine translation. *arXiv preprint arXiv:2104.06951* (2021).

[24] Sennrich, R., Haddow, B., and Birch, A. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2016), pp. 86–96.

[25] Statistics South Africa. Census 2022: Statistical release p0301.4. https://www.statssa.gov.za/publications/P03014/P030142022.pdf, 2022. Accessed 18 April 2025.

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), vol. 30, pp. 5998–6008.

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

A Risk Assessment and Management Plan

Risk	Probability	Impact	Consequence	Mitigation	Monitoring	Management
Delays in Getting Access to CHPC Resources	Medium	Medium	Model training might be delayed if CHPC or GPU cluster is busy, which can slow down the project.	• Book CHPC slots early • Run smaller tests locally	Check CHPC queue times regularly	If unavailable, use cloud GPUs or smaller training jobs
Model training time exceeds expectation	Medium	Medium	We fall behind schedule and might miss crucial project deadlines.	• Overestimate training time requirements • Consult someone knowledgeable on training timing	Compare actual vs. planned progress regularly	Cut non-critical features and focus on core essentials
Group member cannot continue	Low	High	Redistribute workload among remaining members and adjust project scope accordingly	Plan tasks such that they are modular and independent such that removing one doesn't affect the rest	Regular progress check-ins to identify early signs of member disengagement or overload	Can remove a research question from the project or redistribute workload

Table 1: Risk assessment and management plan for the project

B Project Timeline

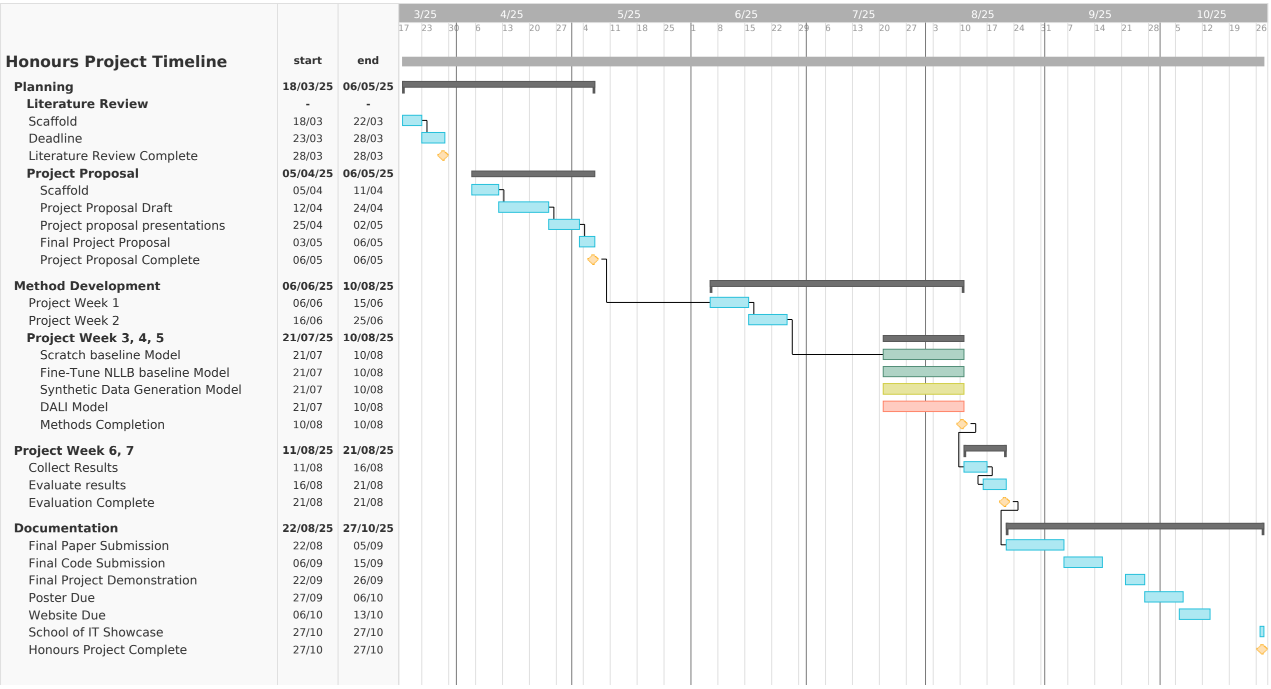

Figure 1: Gantt chart showing the project timeline and milestones