UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Assessing Interpretability in Machine Translation Models for Low-Resource Languages

by

Tsholofelo Gomba
Supervisor: Vukosi Marivate.

Submitted in partial fulfillment of the requirements
for the degree of
Master of Science in Computer Science
in the Engineering, Built Environment and Information Technology
University of Pretoria, Pretoria

December, 2024

Publication data:
Miss.Tsholofelo Gomba. Springer Nature (Natural Language & Linguistic Theory), University of Pretoria, (MSc, Computer Science), Pretoria, December 2024

Abstract

In recent years, we have seen an increase in the adoption of Large Language Models (LLM) usage across many different applications. A practical example is OpenAI's ChatGPT, a tool based on InstructGPT that uses pre-training combined with questioning answering and guidance with reinforcement learning with human feedback.

A gap that still exists, the need for better coverage of low resource languages, has led to a substantial amount of research focused on multilingual LLMs in the Natural Language Processing (NLP) domain bringing about models such as NLLB-200, Glot500-m, and BLOOM. However, most of these black box multilingual LLMs fail at representing low resource languages, especially when applied to translation tasks, as their internal logic remain hidden from the user. This leaves one unable to account for or explain reasons for failures in real-life translations tasks.

This research investigates the performance and interpretability of two models, a LLM and a small-scale model, trained on low-resource language pairs Xhosa-Zulu and Tswana-Zulu. Both models make use of the transformer architecture. The research aims to evaluate the differences in translation quality and interpretability between the models, examining the role of attention mechanisms in capturing context and ensuring correct translations.

The research aims to evaluate the (1) differences in translation quality and interpretability between models of different scales, (2) the impact of training dataset sizes on translation quality, and (3) the effectiveness of post-model eXplainable AI (XAI) methods in evaluating generated translations and model efficiency in low-resource language settings. The post-model methods used are attention pattern analysis, BLEU scores, MMD scores and human evaluation methods.

We conclude that larger models handle linguistic complexities better, training on larger datasets generally improves translation quality, and diverse post-hoc evaluation methods are essential for a comprehensive assessment. This analysis contributes to a better understanding of the strengths and weaknesses of different model scales in machine translation, guiding future developments in XAI for machine translation of languages such as Swati, Tshiluba, Yoruba and other low-resource languages.

Keywords:

Interpretability, machine translation, transformers, low-resource languages

Contents

List of Figures

List of Tables

Chapter 1: Introduction

The adoption of LLMs in domains such as education, healthcare, and psychology has become a notable topic of research [5, 6, 7]. In the field of Neural Machine Translation (NMT), tools like Google Translate have become indispensable for communication across different languages [8, 9, 10]. Many of these translation applications utilize LLMs like BLOOM and ChatGPT [11, 12]. However, most LLMs function as black boxes, with their inner workings and logic hidden from users, leaving the rationale behind their predictions often unknown, even to experts [13]. This opacity can lead to errors in translation that are difficult to detect for non-native speakers, potentially resulting in miscommunication [8, 9]. These challenges highlight the need for interpretable models that are easier to understand and troubleshoot [14].

1.1 Problem statement

The rapid integration of LLMs, such as OpenAI's ChatGPT, into various daily applications has highlighted significant challenges in their use, particularly regarding interpretability and auditability [15]. As these black-box models become more ubiquitous, the need for transparency and interpretability in their decision-making processes is important for ensuring reliability and trustworthiness. This is especially critical in the field of NLP and Machine Translation (MT), where accurate translations are crucial for effective communication and understanding across different languages.

MT has revolutionized the way people communicate and learn new languages, with applications like Google Translate enabling users to easily translate text between various languages [8, 9, 10]. Despite their widespread use, LLMs often fall short in translation tasks, particularly for low-resource languages, leading to errors that can have significant impact. Native speakers may easily identify these inaccuracies, but for general users without a deep understanding of the target language, these errors can result in miscommunication or conveying incorrect messages to the recipient [8, 9].

The field of XAI seeks to address these challenges by enhancing the interpretability of machine learning models, including those used in multilingual translation [16]. XAI techniques provide insights into the functioning of models, which are typically made of billions of parameters and numerous layers, making them complex and difficult to understand. This research is motivated by the need to identify and mitigate failure points within multilingual LLMs used for translation tasks, particularly in low-resource language settings.

This research focuses on comparing the performance and interpretability of a large-scale LLM and a small-scale model in translating low-resource language pairs. By evaluating differences in translation quality and the impact of varying training dataset sizes, we aim to shed light on the strengths and weaknesses of models of different scales. Additionally, we explore the effectiveness of post-model XAI methods in assessing generated translations and improving model efficiency. This investigation will contribute to a deeper understanding of the

applicability of different models in low-resource language translation and guide future developments in XAI for enhancing model transparency and accountability.
The rest of the paper includes the background and literature study in Section 2, model considerations in Section 3 data considerations in Section 4, the methodology in Section 5, experiments carried out in Section 6, the results in Section 7 and the conclusion in Section 9. A list of references and the appendix conclude the paper.

1.2 Objectives

The primary goal of this research is to assess the interpretability of transformer models of different scales in low-resource language settings. Specifically, this research aims to answer the following questions:

- What is the correlation between attention patterns and translation quality?
- How accurate are post-model interpretability methods in evaluating low-resource language translation quality?
- What insights can be gained from:
  - comparing transformer NMT models of different scales?
  - applying model-agnostic post-hoc interpretability techniques to NMT models?
  - training NMT models on different training set sizes for the same translation task?

The secondary goal of this research is to establish a human feedback loop essential for validating model translations and the insights derived from interpretability techniques. By involving native speakers of the low-resource languages, we aim to confirm the validity of the research findings, thereby informing future decision-making processes and enhancing the reliability of XAI methodologies for low-resource language NMT translation.
All the reasons elaborated on above form some of the leading motivations for interpretable models, so that efforts are made towards AI models that are easier to understand and troubleshoot in practice[14].

1.3 Research hypotheses

This research investigates the performance and interpretability of transformer models of two varying scales, a LLM and small-scale model, for low-resource language pairs (Xhosa-Zulu and Tswana-Zulu). The study is guided by the following hypotheses:

- **H1**: LLMs achieve higher translation quality than small-scale models for low-resource languages.

- **H2**: Training dataset size has a significant impact on the translation quality of transformer models.
- **H3**: A range of complimentary post-model XAI methods, such as human evaluation and attention analysis, provide more accurate insights into translation quality than automated metrics alone (e.g., BLEU scores).

These hypotheses directly support the study's objectives and structure, prompting the investigation into model performance and interpretability.

1.4 Contributions

This research presents a comprehensive comparison of attention mechanisms across two neural network architectures, focusing on their ability to provide quality translations for low-resource language pairs. The key contributions of this study include:

- **Comparison of transformer architectures**: We compare the attention mechanisms of two transformer-based models—a small-scale model, OpenNMT, and a LLM, M2M100—to assess their efficacy in handling low-resource language translations for the chosen language pairs.
- **BLEU scores contribution**: At the time of writing, the only BLEU scores we could use as reference are those by Elmadani et. al (2022) for Xhosa-Zulu who benchmark a score of 8.5 on 1M aligned sentence pairs from the WMT22 dataset [17]. This research records BLEU scores of 29.59 on the same dataset fine-tuned on M2M100, of which the translations are evaluated as correct by human evaluators. The paper by Elmadani et. al (2022) does not record scores for Tswana-Zulu, which we provide in this paper.
- **Interpretability analysis**: We leverage XAI techniques to provide insights into how attention mechanisms contribute to model decisions, with a particular focus on the quality of translations for low-resource languages.
- **Recommendations for future research**: Based on the research findings, we provide recommendations for the development and deployment of explainable AI methods that are interpretable and capable of aiding evaluation towards high-quality machine translations for low-resource languages.

Through this work, we aim to advance the understanding of attention mechanisms in transformer-based neural network architectures and their implications for explainability in AI, particularly for low-resource languages. This research research contributes to future studies in this domain and offers practical insights for the development of more interpretable AI systems.

1.5 Limitations

1.5.1 Model complexity and performance

The research compares architectures implemented in a LLM and a small-scale model. These models inherently differ in complexity, which can affect their performance. The LLM, such as M2M100, is likely to capture more intricate language-specific linguistic features due to its size and training data. In contrast, the small-scale model, like OpenNMT, may be less capable in this regard but potentially more interpretable. This trade-off between complexity, performance, and explainability is a significant limitation, as it may influence the generalizability of the findings.

1.5.2 Training data constraints

Only data from the WMT22 dataset was used for training the models. The original training data for M2M100 is not publicly available; otherwise, it would have been utilized to potentially improve the model's accuracy. The reliance on a single dataset, particularly in the context of low-resource languages, may limit the robustness of the results. The accuracy and quality of the translations are directly influenced by the quality and comprehensiveness of the WMT22 dataset.

1.5.3 Impact of dataset on translation quality

The specific characteristics and quality of the WMT22 dataset significantly influence the performance of both models. Any biases or limitations within this dataset are likely to be reflected in the translation outputs. This limitation underscores the importance of dataset quality in training machine translation models, particularly for low-resource languages.

1.5.4 Generalizability to other language pairs

The findings of this research are specific to the language pairs studied, which are low-resource and non-English. While the insights gained are valuable, they may not be directly applicable to other other low resource language pairs, or those with abundant resources due to factors such as linguistic features. This limitation should be considered when extrapolating the results to other contexts.

1.6 Dissertation outline

The rest of this thesis will be organised as follows:

- Chapter 2 looks at the background and literature review around explainable AI, LLMs, machine translation and interpretability applied to multilingual LLMs

- Chapter 3 provides a deep dive into the models and corresponding model architectures for the models used in this research
- Chapter 4 provides a deep dive on the data used in this research
- Chapter 5 looks at the explainability methodologies applied to the models architectures and low resource languages
- Chapter 6 outlines the experiments conducted in this paper
- Chapter 7 discusses the results of the model training across various architectures and low resource languages
- Chapter 8 evaluates the research hypotheses based on previous chapters, contextualizes the research into the current research landscape and desribes limitations of the small-scale model
- Chapter 9 provides the conclusion and future work

Chapter 2: Background and literature study

2.1 Interpretable AI

eXplainable Artificial Intelligence (XAI), aims to makes sense of Machine Learning (ML) models using relationships learned by the model or those contained in data[18]. Interpretability in itself is defined by Kim et. al.[19] as, "the degree to which a human can consistently predict the model's result." Models with a higher degree of interpretability make it easier for humans to understand how a result came about[20, 21]. Oftentimes, the terms interpretability and explainability are used interchangeably, and this paper will follow this pattern.

When evaluating a model's interpretability, one can do so pre-model, in-model or post-hoc. Pre-model techniques apply to the data and not the model itself, thus explaratory data analysis tasks such as Principal Component Analysis (PCA) and MMD (Maximum Mean Discrepancy) may be applied[21, 13]. In-model applies to ML models which are intrinsically interpretable, providing enough detail into how the model works through constraints such as causality or monotonicity [22, 13]. Post-hoc techniques are applied to the learned model[23, 13]. The techniques applied may be model specific, if the technique only applies to that model; or model agnostic, if they can be applied to any model. Most post-hoc techniques are model agnostic while in-model techniques are model specific [13].
Model interpretability may also be evaluated globally, where we aim to comprehend the entire black box model at once, or locally, where we aim to understand how a single prediction was made[21].

2.2 Neural machine translation and sequence to sequence models

An application of LLMs is neural machine translation (NMT) or simply machine translation (MT). In MT, sequence to sequence models (seq2seq) are used. A sequence of words or characters $(x_1, ..., x_S)$, from the source language are mapped to the best translation $(y_1, ..., y_T)$ in the target language[24]. Thus many MT models are trained by maximizing the probability of deriving the target sentence given the source sentence and target language l_t :

$$P(y_1, \ldots, y_T \mid x_1, \ldots, x_S, l_t) \tag{2.2.0.1}$$

Seq2seq applications include speech recognition, dialog, tagging, MT [25], image captioning, pose prediction and syntactic parsing [26]. These models are made up of encoders and decoders which use Recurrent Neural Network (RNN) or transformer models based on LSTM (Long Short Term Memory) or GRU

(Gated Recurrent Unit), with LSTM being the more popular choice.
In MT tasks, it is imperative that the LM in use is context aware in order to provide the correct translation[24]. Thus, the encoder summarizes and converts the input sequence from the source language into hidden state and cell state vectors that represent the token and context of each token. The final hidden state is passed to the decoder as input. The decoder then determines the target language sentence by using the current hidden state vector, cell state vector and the previous token. Softmax is used to compute the probability vector for the words to append to the output sequence[27].

While seq2seq models have been foundational in MT, various advancements have been made over time. For example, attention mechanisms have been added to seq2seq models to improve translation quality by allowing the model to focus on specific parts of the input sentence when producing each part of the output sentence. In addition to traditional seq2seq models, more recent approaches such as transformer-based models including BERT[28], GPT[29], and others have gained prominence in MT due to their ability to capture complex dependencies and contextual relationships across long sequences.

2.3 Multilingual NMT interpretability

Existing research on multilingual neural machine translation (NMT) interpretability explores topics such as how to better understand and explain the internal workings of NMT models across different languages [30, 31].
In recent years, we have seen strides towards the interpretability of neural representations for non-English-only languages. For multilingual cross-lingual interpretability, Pires et al. (2019), focused on the ability to transfer representational knowledge across languages for zero-shot semantics [32]. This involves leveraging existing knowledge from one language and various probing experiments to understand semantics in another language without explicit training data or direct translation pairs. Further research has been catered towards research aims to shed light on the decision-making process of models, the impact of attention mechanisms [33, 34], and how models handle different linguistic features [35, 36]. Comparative linguistic interpretability, which involves assessing and understanding the ease with which linguistic structures or features can be compared across different languages has also been applied to models like mBERT.
Rama et al. (2020), probed mBERT representations to infer a phylogenetic language tree for 100 languages in an attempt to better interpret the model and explain its downstream task cross-lingual behavior [37].
Other notable areas of research include the introduction of attention mechanisms to understand how models focus on specific parts of input sentences when producing translations [38]. Researchers have investigated the role of linguistic features in translation quality and how different architectures handle multilingual data [39]. Additionally, there is growing interest in methods for visualizing and interpreting translation choices made by models, such as heatmaps and saliency maps [40]. This past literature significantly influenced the research direction by leading us to explore attention patterns, utilize visualizations, and investigate transformer architectures of different computational scales for interpreting low-resource language machine translation.

These efforts contribute to developing more transparent and explainable models, which are crucial for building trust and improving the reliability of multilingual NMT systems.

2.4 Large language models

LLMs are models trained on massive amounts of textual data, representing a significant advancement in the field of natural language processing (NLP), with practical applications such as question answering, text generation, translation and high accuracy text summarization [29].
These models, such as BERT[41], GPT-3 [42], and T5[43], are built on transformer architectures that enable them to capture long-range dependencies and contextual relationships within text. This has led to state-of-the-art performance across a wide range of NLP applications.
Multilingual LLMs aim to do the same tasks, with focus on non-English language datasets. Recent research into multilingual LLMs has focused on enhancing their performance across a wide range of languages, particularly low-resource languages [4].
These models, such as XLM-R and M2M-100, use transformer architectures to capture linguistic patterns across different languages, enabling state-of-the-art results in multilingual tasks. Research has also explored techniques for reducing bias and improving fairness in multilingual models[44]. Additionally, challenges such as scalability, efficiency, and ethical considerations in data usage continue to be addressed as these models evolve. Multilingual LLMs hold significant promise for the future of language processing and communication technologies.

2.5 BLEU scores for MT evaluation

MT research has traditionally relied on automated metrics such as BLEU (Bilingual Evaluation Understudy) scores to assess translation quality. BLEU, introduced by Papineni et al. in 2002 [45], is a precision-based metric that compares n-grams in the machine-generated translations to those in a set of reference translations. The simplicity and reproducibility of BLEU have made it the go to standard for MT evaluation in numerous studies [46, 47].

Despite its widespread use, BLEU has several limitations, especially in capturing the nuances of translation quality. Firstly, BLEU is known to correlate poorly with human judgment, particularly when evaluating translations of complex syntactic structures and idiomatic expressions [48]. It focuses mainly on surface-level matching and may not adequately reflect the semantic accuracy or fluency of translations. Research by Reiter (2018) highlights how BLEU can be misleading, especially for low-resource languages where reference translations might be scarce or varied [49].

Several studies advocate for complementing BLEU scores with human evaluations to provide a more comprehensive assessment of translation quality [50, 51]. Human evaluators can judge the adequacy and fluency of translations, offering insights into whether the translated text conveys the same meaning as the source text and whether it is grammatically and idiomatically correct in the

target language. Such evaluations are crucial for understanding the practical usability of MT systems in real-world scenarios.

However, the integration of human evaluation into the MT evaluation pipeline is often limited due to its time-consuming and costly nature. While recent studies have begun to recognize the importance of human input [52, 53], there is still a tendency to prioritize BLEU scores in literature, potentially overlooking critical aspects of translation quality that only human judges can reliably assess.

This research aims to address the gap in the current evaluation practices by incorporating human evaluators alongside BLEU scores to assess the correctness of translations in low-resource language settings. By doing so, we hope to provide a more balanced view of translation quality that considers both automated metrics and human judgments.

2.6 Human evaluation in machine translation

Human evaluation plays a fundamental role in assessing the quality of MT outputs. While automated metrics like BLEU [45] are widely used for their efficiency, they often fail to capture nuanced linguistic features, such as semantic accuracy, fluency, and cultural appropriateness. Human evaluation methods, on the other hand, provide valuable insights into translation quality through subjective analysis.

Human evaluation has been a critical component in benchmarking MT systems across large-scale shared tasks and studies. Bojar et al., 2018 [54] and Bojar et al., 2019 [55] incorporated direct assessment and error annotation methods for evaluating translations across diverse languages. Freitag et al., 2021[56] conducted a large-scale study combining Direct Assessment (DA) and Explanation Satisfaction Scale (ESS) to analyze translation errors, demonstrating the limitations of automated evaluation metrics.

2.6.1 Approaches to Human Evaluation

Human evaluation of translations generally involves evaluating outputs based on key criteria, including adequacy (how well the translation conveys the meaning of the source text) and fluency (how natural and grammatically correct the translation is in the target language). These evaluations are often performed using the following approaches:

- **DA**: Annotators rate translations on a continuous scale for adequacy and fluency. DA is widely adopted in shared tasks like WMT [54, 55] due to its reliability and simplicity.
- **Ranking-based evaluation**: Human evaluators are presented with translations generated by multiple MT models for the same source sentence and are asked to rank these outputs in order of quality, from best to worst. This approach simplifies the evaluation process compared to assigning absolute scores, reducing cognitive load for human annotators and leading to more consistent results. Studies like Bojar et al., 2016[57], Daybelge and Cicekli, 2011[58] and Freitag et al., 2021[56] have shown that ranking

is effective for comparing MT models. Ranking-based evaluation is particularly useful in scenarios where absolute scores or error annotations are not required, but relative MT model performance must be assessed.

- **Error annotation frameworks**: Methods such as the Multidimensional Quality Metrics (MQM) framework [59] and ESS focus on fine-grained error analysis. Annotators categorize errors (e.g., accuracy, fluency, or terminology) and assign severity scores. We expand on MQM and ESS in sections 2.6.2 and 2.6.3 as the selected human evaluation methods applied in this research.

2.6.2 MQM

MQM is a detailed framework for assessing translation quality across multiple dimensions and error categories. Developed by the QTLaunchPad project and standardized by the European Union's QT21 project, MQM provides a robust, flexible, and detailed approach to evaluating the quality of translations[56, 59]. It incorporates a wide range of error types and severity levels, allowing for a nuanced assessment that goes beyond the capabilities of traditional metrics such as BLEU or METEOR. Below are detailed reasons for using MQM for low-resource language human evaluation:

(a) Comprehensive evaluation framework

MQM offers a more granular and comprehensive evaluation of translation quality compared to traditional metrics. It allows evaluators to classify errors into categories such as accuracy, fluency, terminology, and style, each with varying degrees of severity (minor, major, critical). This detailed categorization helps identify specific issues in translations, which is crucial for low-resource languages where linguistic nuances and context may be particularly challenging.

(b) Flexibility and customization

The flexibility of MQM makes it particularly suitable for low-resource language evaluation. Evaluators can tailor the framework to address the unique linguistic features and challenges of the target languages. This adaptability is vital for low-resource languages, which often lack extensive evaluation resources and tools. By customizing MQM, evaluators can focus on the most relevant aspects of translation quality for the specific language pair.

(c) Enhanced reliability and consistency

MQM's structured approach ensures a consistent and reliable evaluation process. The detailed guidelines and standardized error typology help reduce subjective biases among human evaluators, leading to more reliable and reproducible results. This consistency is essential when evaluating low-resource languages, where expert evaluators may be scarce, and consistent evaluation criteria are critical for obtaining valid results.

(d) Detailed insights for model improvement

Using MQM provides detailed insights into the specific strengths and weaknesses of the translation models. This information is invaluable for refining and improving the models. For low-resource languages, where data for training and evaluation is limited, understanding the exact nature of translation errors can guide targeted improvements, enhancing overall translation quality.

2.6.3 Explanation Satisfaction Metrics (ESM) and ESS

Explanation Satisfaction Metrics (ESM) are designed to evaluate how well explanations of model outputs satisfy the needs and expectations of human users. They focus on various aspects such as clarity, usefulness, and trust, providing a comprehensive measure of how effective explanations are in aiding understanding and decision-making [60]. One specific implementation of ESM is the Explanation Satisfaction Scale (ESS), which uses structured questionnaires to gather quantitative and qualitative feedback from human evaluators.
The ESS is a tool used to systematically collect human evaluators' feedback on the quality of explanations provided by machine learning models and their post-model evaluation techniques. ESS employs a Likert-scale questionnaire to rate various aspects of the explanations, such as their clarity, detail, helpfulness, and style [60, 61]. This structured approach allows researchers to quantitatively measure how well explanations meet users' needs and identify areas for improvement.

Example structure of ESS:

- Clarity: How clear and understandable is the explanation?
- Detail: How detailed and thorough is the explanation?
- Helpfulness: How helpful is the explanation in aiding understanding or decision-making?
- Trust: How much does the explanation increase your trust in the model's outputs?
- Style: How much does the translation keep to the source sentence style?

Given the description above, the text below outlines the motivation for using ESS for human evaluation:

(a) Comprehensive and detailed feedback

ESS provides detailed feedback across multiple dimensions of explanation quality. This comprehensive approach is particularly valuable for low-resource language translation, where understanding and improving model performance requires nuanced insights into how well explanations help users understand the model's behavior and outputs.

(b) Structured and quantifiable

The structured nature of ESS allows for quantifiable measurement of satisfaction across different dimensions. This quantifiability is crucial for systematic comparison and analysis, helping to identify specific strengths and weaknesses in the explanations provided by various post-model evaluation techniques.

(c) User-centered evaluation

ESS is inherently user-centered, focusing on the evaluators' perceptions and experiences. This focus ensures that the feedback collected is directly relevant to improving the usability and effectiveness of post-model evaluation tools in real-world scenarios, where human understanding and trust are paramount.

(d) Flexibility and adaptability

ESS can be easily adapted to different contexts and evaluation needs. For low-resource language translation, the scale can be customized to address specific challenges and linguistic features relevant to the languages being studied. This flexibility ensures that the evaluation is tailored to the unique aspects of low-resource languages.

2.6.4 Strengths and challenges of human evaluation

Human evaluation remains the gold standard for assessing machine translation quality. It provides a nuanced understanding of translation quality that automated metrics cannot fully capture. By incorporating subjective human judgments, methods like DA, MQM, and ESS identify subtle errors in accuracy, fluency, and style. However, human evaluation is time-consuming, costly, and often subject to annotator bias or variability [57, 59].

Despite these challenges, studies have demonstrated that combining human evaluation with automated metrics improves the reliability of MT system assessment, particularly for complex or low-resource languages.

Chapter 3: Models

The translation of low-resource languages poses unique challenges in the field of NLP. Selecting appropriate models for such tasks requires careful consideration of their capabilities and limitations. For this research, we will take a look at the models M2M100[4] and OpenNMT[62]. M2M100 and OpenNMT represent two contrasting approaches to machine translation: a state-of-the-art LLM and a versatile, modular, small-scale model, respectively. This choice is motivated by the need to explore and compare the performance and interpretability of different scales of models in low-resource settings.

3.1 M2M100

3.1.1 Comprehensive multilingual coverage

Multilingual Machine Translation (MMT) aims to build a single model that does translations between any language pairs. However, much research has been English-centric, in that data on which models are trained on was translated to and from English, thus creating an English-centric bias. This resulted in models that fail to appropriately represent real life translation applications and low performing non-English translation directions[4].
Given these limitations, M2M100[4] was developed in aim to create a true Many-to-Many multilingual translation model that can translate between any pair of 100 languages. Through large-scale mining, this research produced a dataset comprising of 7.5B training sentences thus allowing for training data across thousands of non-English translations. This state-of-the-art multilingual model was designed to handle translation tasks for a wide array of language pairs without relying on English as an intermediary. This is particularly important for low-resource languages, which may benefit from direct translations to and from other non-English languages.

In order to achieve translation across the 100 language pairs from different language families, M2M100 reduces complexity through automatic parallel corpora construction[63], using a novel data mining approach that exploits language similarity to avoid mining in all directions. The model also applies backtranslation - where the target language translated text is translated back to the source language text for quality and accuracy verification; to improve the translation quality on zero-shot and low resource language pairs.

3.1.2 Performance in low-resource settings

M2M100 has been trained on a large-scale multilingual corpus, allowing it to achieve high-quality translations even for languages with limited resources. Its ability to leverage shared representations across languages helps improve translation quality where data is sparse[4].

3.1.3 Transformer architecture

M2M100 makes use of a transformer architecture model whose encoder and decoder modules are based on the implementation proposed by Vaswani et al.,(2017)[64]. This architecture's attention mechanisms allow for better handling of contextual information, which is crucial for accurate translations in low-resource languages.
The subsequent headings describe the components that make up the transformer architecture as used by M2M100, detailing the decoder, encoder and attention layers.

(a) Encoder

The encoder takes as input a sequence of tokens $X = (x_1, ..., X_S)$ and source language l_s. It produces a sequence of embeddings $H = (h_1, ..., h_S)$ which are of the same length as the input sequence.

$$H = encoder(X, l_s) \tag{3.1.3.1}$$

(b) Decoder

The decoder takes as input the embeddings from the encoder along with the target language l_t, then token by token autoregressively produces the target sentence $Y = (y_1, ..., y_T)$.

$$\forall_i \in [1, ..., T], y_{i+1} = decoder(H, l_t, y_1, ..., y_i) \tag{3.1.3.2}$$

(c) Self-attention and feed-forward layers

Both the encoder and decoder transformer layers have a self-attention layer. This layer predicts/infers how strongly one sequence element is correlated to the other sequence elements, and updates the element accordingly.

$$A = norm(C + self - attention(C)) \tag{3.1.3.3}$$

Likewise, both the encoder and decoder have a feed forward layer. This layer passes each element sequentially through a multi-perceptron layer to produce the final output.

$$B = norm(A + feed - forward(A)) \tag{3.1.3.4}$$

Normalization is applied to both self-attention and feed-forward operations as shown above.

3.2 Open NMT

Open NMT (ONMT) is an open source NMT toolkit with two versions: OpenNMT-py and OpenNMT-tf. For this research, we made use of OpenNMT-py.
OpenNMT-py is the PyTorch version of the OpenNMT project with neural machine translation as one of its core use cases. It is designed to be research

friendly to try out new ideas in translation, language modeling, summarization, and many other NLP tasks[62].

This model, though simple, applies core NMT standard architectural norms including: transformer[1] or RNN architecures[62], different attention mechanisms, configurable encoder and decoder bridges, input feeding; standard learning techniques including: dropout, weight tying, learning rate scheduling and early stopping criteria.
When using ONMT, a source sentence of length l_x has each word represented by a sequence of one-hot encoded vectors $x_1, ..., x_{l_x}$. Likewise, a target sentence of length l_y has each word represented by a sequence of one-hot encoded vectors $y_1, ..., y_l$.

3.2.1 Flexibility and modularity

OpenNMT is an open-source framework designed to facilitate the development of neural machine translation systems. Its modular architecture allows for easy customization and extension, making it an ideal choice for researchers who need to experiment with different models and techniques. OpenNMT supports a wide range of configurations and models, including both transformer-based and traditional architectures, which can be fine-tuned to meet the specific requirements of low-resource language translation tasks [62].

This flexibility enables researchers to tailor the model to better handle the nuances of specific language pairs, providing a more adaptable solution compared to more rigid, pre-trained models. OpenNMT's modular design also allows for the integration of various pre- and post-processing techniques, which are essential for achieving high-quality translations in diverse language settings.

3.2.2 Ease of use and rapid prototyping

OpenNMT is known for its user-friendly interface and comprehensive documentation, which facilitate rapid prototyping and experimentation. This ease of use is crucial for research in low-resource language translation, where quick iteration and testing of different approaches can significantly impact the development of effective models. The framework's support for multiple programming languages and platforms further enhances its accessibility and usability for a wide range of researchers [62].

Additionally, OpenNMT provides tools for data preprocessing and filtering, model training, and evaluation, allowing researchers to quickly set up and test various configurations. This capability is particularly important in low-resource settings, where time and computational resources may be limited.

3.2.3 Community support

OpenNMT benefits from a robust community and a rich ecosystem of tools and resources. This active support network, through the ONMT community webpages, provides access to a wealth of knowledge, and implementation examples,

[1] OpenNMT-py Docs https://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-train-the-transformer-model

making it easier for researchers to get started and build upon existing work. The community's collaborative nature fosters the sharing of techniques and improvements, which can accelerate the development of high-quality translation systems for low-resource languages. This extensive support makes OpenNMT a powerful and versatile choice for low-resource language translation research.

3.2.4 Transformer architecture

ONMT implements the decoder encoder transformer architecture proposed by Vaswani et. al.(2017)[64] with code based on The Annotated Transformer by Rush et. al.(2018)[62, 65].
The subsequent headings describe the components that make up the transformer architecture as used by ONMT, detailing the decoder and encoder layers, whose architecture differs from that of M2M100 in Section 3.1.3.

(a) Encoder

The encoder takes as input the sequence $x_1, ..., x_{l_x}$. We use $E_{src}x_I$ to look up the word embedding for each input word then add a position encoding to it. We then stack the resulting sequence of word embeddings to form matrix $\mathbb{X} \in \mathbb{R}^{l_x \times d}$, with l_x being the sentence length and d the dimensionality of the embeddings. Given d_a as the attention space dimensionality and d_o as the output dimensionality, we have the following learnable parameters:

$$A \in \mathbb{R}^{d \times d_a} \tag{3.2.4.1}$$

$$B \in \mathbb{R}^{d \times d_a} \tag{3.2.4.2}$$

$$C \in \mathbb{R}^{d \times d_o} \tag{3.2.4.3}$$

We use the matrices above to transform the input matrix into new word representations H. The new representation includes the attention of all other source words, with $softmax(XAB^\top X^\top)$ computing self attention:

$$H = softmax(XAB^\top X^\top)XC \tag{3.2.4.4}$$

This encoder implementation uses multi-head attention whereby, for each head, we compute the transformation k times with different parameters A, B and C. Once this parallel computation is done, we concatenate all Hs, apply layer normalization and finally, a feed forward layer as below:

$$H = [H^{(1)}; \ldots; H^{(k)}] \tag{3.2.4.5}$$

$$H' = norm(H) + X \tag{3.2.4.6}$$

$$H^{(enc)} = feedforward(H') + H' \tag{3.2.4.7}$$

We can stack multiple encoding layers by setting $X = H^{enc}$ and repeating the above computation. We have $d_o = d/k$ to that $H \in \mathbb{R}^{l_x} \times d$.

(b) Decoder

The decoder is similar to the encoder, except, the decoder takes as input the stacked target embeddings $Y \in \mathbb{R}^{l_y} \times d$. The decoder computes masked self attention using $softmax(YAB^\top Y^\top)$ as part of the computation below:

$$H = softmax(YAB^\top Y^\top)YC \tag{3.2.4.8}$$

Once we obtain $H' = H + Y$, we compute multi-headed attention between the decoder representations H' and $H^{(enc)}$. We use $softmax(H'AB^\top H^{(enc)\top})$ to compute the source target attention:

$$Z = softmax(H'AB^\top H^{(enc)\top})H^{(enc)}C \tag{3.2.4.9}$$

We then apply the feed forward and normalization layers as follows:

$$H^{(dec)} = feed - forward(norm(H' + Z)) \tag{3.2.4.10}$$

Finally, to predict the target words, we use:

$$H^{(enc)}W_{out} \tag{3.2.4.11}$$

Choosing M2M100 and OpenNMT for this research provides a robust comparative framework for evaluating machine translation in low-resource languages. M2M100 represents the cutting edge of multilingual, large-scale models with sophisticated mechanisms for handling diverse languages, while OpenNMT offers a highly customizable, resource-efficient alternative. This comparison will shed light on the trade-offs between translation quality, interpretability, and efficiency, guiding future developments in machine translation and explainable AI for low-resource languages.

Chapter 4: Data

4.1 Language selection

Having selected M2M100 as the LLM and given ONMT's freedom to train on any language pairs, we selected language pairs on which M2M100 was initially trained on for which we could attain bilingual human evaluators to verify translation results. As such, we selected the language pairs Xhosa-Zulu and Tswana-Zulu.
We fine-tuned M2M100 only on those languages for optimal training on these language pairs and trained ONMT on the same language pairs. All the language pairs are from the Niger-Congo subset[4] as per table 4.1

ISO	**Language**	**Family**	**Script**
am	Amharic	Ethiopian	Ge'ez
ff	Fulah	Niger-Congo	Latin
ha	Hausa	Afro-Asiatic	Latin
ig	Igbo	Niger-Congo	Latin
ln	Lingala	Niger-Congo	Latin
lg	Luganda	Niger-Congo	Latin
nso	Northern Sotho	Niger-Congo	Latin
so	Somali	Cushitic	Latin
sw	Swahili	Niger-Congo	Latin
ss	Swati	Niger-Congo	Latin
tn	**Tswana**	**Niger-Congo**	**Latin**
wo	Wolof	Niger-Congo	Latin
xh	**Xhosa**	**Niger-Congo**	**Latin**
yo	Yoruba	Niger-Congo	Latin
zu	**Zulu**	**Niger-Congo**	**Latin**

Table 4.1: Low resource languages selected from original M2M100 languages list[4].

4.2 Datasets used

4.2.1 WMT22

An older WMT dataset was used for training some of the languages in M2M100. We confirmed that it was not for the Xhosa-Zulu and Tswana-Zulu language

pairs listed in table 4.2 so an to ensure we are not fine-tuning with the same dataset that the model was trained on. As such, we used the WMT22[1] dataset which was released in 2022. It provides a large corpus with millions of records across many languages including Tswana-Zulu and Xhosa, so we considered this for the training, testing and evaluation sets. We list the number of parallel sentences for the language pairs in table 4.3.

ISO 639-1 Code (ONMT)	**ISO 639-2/3 Code (WMT22)**	**Language**
tn	tsn	Tswana
xh	xho	Xhosa
zu	zul	Zulu

Table 4.2: ISO 639-1 code, ISO 639-2/3 code and language name for Tswana, Xhosa and Zulu

Language Pairs	**Size**
Tswana - Zulu	341 119
Xhosa - Zulu	1 066 327

Table 4.3: WMT22 dataset language pairs lengths (in ascending order)

4.2.2 The Vuk'uzenzele South African multilingual corpus

In order to evaluate the correctness of the translations generated by M2M100 and ONMT, we sampled four sentences for each language pair from the Vukuzenzele South African multilingual corpus dataset[2]. This corpus contains correct evaluated parralel translations for both Tswana-Zulu and Xhosa-Zulu. We confirmed that there is no overlap between the WMT22 and Vuk'zenzele dataset.

[1] WMT22 `https://www.statmt.org/wmt22/large-scale-multilingual-translation-task.html`

[2] Vukuzenzele corpus `https://github.com/dsfsi/vukuzenzele-nlp/blob/master/README.md`

Chapter 5: Methodology

This chapter describes the methodologies applied in this research in order to evaluate the translations generated by the LLM and small-scale model. Figure 5.1 provides a high level summary of the methods applied.

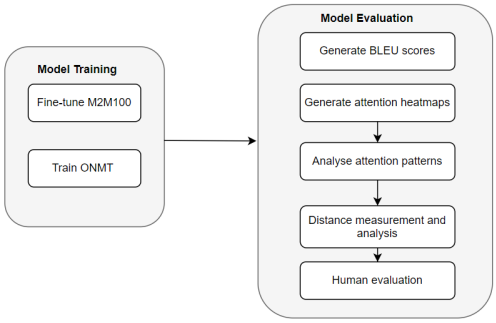

Figure 5.1: A high level summary of the methodology applied

5.1 Model training

The first step in the practical application of this research is training the models on the low resource language pairs under consideration. The fine-grain training procedures are outlined in section 6

5.2 Model evaluation

This research applies post model interpretability techniques to explain the performance of the trained models in translating the low resource language pairs. Post model (post-hoc) interpretability refers to techniques applied to the model after training in order to help us answer the question: *what else can the model tell us?*[21]. Specific methodologies are defined in the following subsections.

5.2.1 BLEU scores

To evaluate translation quality, we apply BLEU (Bilingual Evaluation Understudy) scores, which quantify the correspondence between the machine-generated translation and a reference translation. BLEU scores are computed based on the n-gram overlap between the reference and the hypothesis translations, penalizing sentences that are either too short or too long relative to the reference. This scoring mechanism provides a robust measure of translation accuracy by assessing both precision and length consistency [45].

In this research, BLEU scores will be used to evaluate the performance of both the LLM and the small-scale model across various training datasets, including full, two-thirds, and one-third subsets. The consistent application of BLEU scoring across these datasets enables us to objectively compare translation quality and identify performance trends relative to the amount of training data used.

BLEU scores are chosen for their effectiveness in capturing essential aspects of translation quality, such as fluency and accuracy, without requiring extensive human intervention. This makes BLEU an ideal metric for evaluating machine translation, particularly in scenarios where rapid and consistent assessment is necessary.

By utilizing BLEU scores, we aim to ensure that the translation outputs of the models are evaluated with a reliable and standardized metric, facilitating a clear understanding of each model's capability in producing accurate translations.

5.2.2 Attention heatmaps

We generate attention heatmaps for source and target translations for both the LLM and the small-scale model. Both the LLM and the small-scale model utilize transformer architectures which incorporate self-attention mechanisms. Attention heatmaps visualize the attention weights, indicating how much focus each word in the source sentence receives from each word in the target sentence[66]. Examining these heatmaps helps us understand how each model handles the translation process, particularly for low-resource languages.

5.2.3 Attention patterns analysis

Attention pattern analysis helps us describe the inner workings of NMT models by analysing the attention heatmaps generated prior in a human-understandable manner. Researchers can gain insights into how the model aligns and attends to different parts of the input sequence during the translation process[67]. We analyze the attention heatmaps to identify common patterns such as diagonal (direct correspondence) and off-diagonal patterns (indirect or contextual dependencies). By comparing the attention patterns between the LLM and the small-scale model, we aim to understand differences in how each model processes translations and captures linguistic features.

5.2.4 Distance measurement

In order to support and/or further evaluate the correctness of the translations provided by the previous three methods, we apply a distance measurement metric to measure the similarity between the model translations and the target sentences obtained from the Vuk'zenzele dataset[2] for the language pair under consideration.

The Maximum Mean Discrepancy (MMD) is a statistical measure used to quantify the discrepancy between two probability distributions by comparing their respective feature spaces[68, 69]. MMD is used in various domains, includ-

ing machine learning, domain adaptation, and statistical hypothesis testing for measuring the similarity or dissimilarity of distributions.

For this research, the MMD score provides a measure of how similar or dissimilar the embeddings of the model translation and the target sentences are. A lower MMD score indicates that the distributions of embeddings are more similar, suggesting that the sentences are semantically closer or have similar representations. A higher MMD score indicates greater dissimilarity between the embeddings, suggesting differences in semantics or representation. We explain the MMD formula we are using below:

$$\text{MMD Score} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{m}k(\mathbf{x}_i,\mathbf{x}_j)+\frac{1}{m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}k(\mathbf{y}_i,\mathbf{y}_j)-\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k(\mathbf{x}_i,\mathbf{y}_j) \tag{5.2.4.1}$$

where:

- i and j are indices representing the sentence embeddings
- $k(x_i, x_j)$ represents the kernel function for the model translation embeddings
- $k(y_i, y_j)$ represents the kernel function for the target sentence embeddings
- $k(x, y)$ represents the kernel function for both embeddings

Kernel function: we use the Radial Basis Function (RBF) kernel to which measures the similarity between pairs of embeddings. The model translation sentence embeddings x and target sentence embeddings y are transformed as follows:

$$k(x,y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right) \tag{5.2.4.2}$$

Gram matrix: We apply the kernel function to all pairs of the embeddings to compute the Gram matrix. k_{xy} represents the model translation sentence embeddings Gram matrix, and k_{yy} represents the target sentence embeddings Gram matrix, with k_{xy} representing the Gram matrix of both embeddings.
MMD score: we compute the MMD score by calculating the difference between the mean similarity within each distribution (the mean of k_{xx} and k_{yy}) and twice the mean similarity between the two distributions (the mean of k_{xy}). This quantifies the discrepancy between the distributions of embeddings from the two sentences.

5.2.5 Human evaluation

The final step in the evaluation involves introducing human feedback in two parts: (1) evaluating the correctness of the model translations using MQM, (2) evaluating the correctness of the results and conclusions derived from the post model XAI techniques applied using ESS.
MQM, introduced in Section 2.6.2, is employed to verify the correctness of translations produced by both the LLM (M2M100) and the small-scale model (OpenNMT). Bilingual evaluators will use the MQM framework to systematically assess translations, categorizing errors into predefined types and severity

levels. This detailed evaluation will provide a comprehensive measure of translation quality, highlighting specific areas for improvement and ensuring a thorough verification process.
While MQM provides a detailed framework for assessing translation quality, it primarily focuses on categorizing translation errors and their severity. It does not explicitly measure the satisfaction of users with the explanations provided by the models or post-model evaluation techniques. ESS, described in Section 2.6.3, complements MQM by adding a layer of user-centered evaluation that focuses on the interpretability and utility of explanations. Traditional metrics like BLEU and METEOR focus on the accuracy of translations but do not provide insights into the interpretability or usefulness of model explanations. They lack the user-centered perspective that ESS offers, which is crucial for understanding how well explanations aid human evaluators in interpreting model outputs.

5.2.6 Summary of methods

The methodologies outlined above provide a comprehensive approach to evaluating and comparing the performance and interpretability of large-scale and small-scale transformer models in translating low-resource languages. Through attention heatmaps, pattern analysis, distance measurement, and human evaluation, this research aims to advance the understanding of how different model scales handle the complexities of low-resource language translation.

Chapter 6: Experiments

6.1 Train/Test/Dev dataset splits

To ensure uniformity in the results across all models, we divided the dataset to maintain consistency between the fine-tuning of M2M100 and the training of ONMT, using identical WMT22 dataset splits. Each of the testing and evaluation sets comprised 2,000 aligned pairs. The remaining data constituted the full training set, which was further divided into three subsets: one-third, two-thirds, and the complete set.

Dataset sizes and benefits

- **Benefits of a small dataset:**
  - Allows for efficient evaluation of how each model utilizes the available data, especially for low-resource languages.
  - Enables preliminary evaluations and adjustments with smaller datasets, saving computational resources.
- **Benefits of a large dataset:**
  - Allows for observation of the performance scalability of the models with increasing training data.
  - Helps identify incremental performance improvements as more training data becomes available.
  - Highlights improvements in translation quality with larger datasets.
  - Facilitates examination of how attention heatmaps and patterns change with different dataset sizes.
  - Provides a basis to investigate consistency and accuracy in translations with larger training sets.

We ensured no overlap of training data across the one-third and two-thirds splits, as well as between the testing and evaluation sets. During the data cleaning process, some records, including empty and duplicate entries, were removed, resulting in the final dataset sizes presented in Table 6.1.

Motivation for splitting the training dataset

Splitting the training dataset into subsets of one-third, two-thirds, and the full set allows us to reap the benefits associated with both small and large datasets. This approach provides valuable insights into various aspects of model performance and behavior:

(1) Evaluating data efficiency and model scalability

Language	Training	Testing/Eval
Xhosa-Zulu (full)	1 062 338	1995
Xhosa-Zulu (1/3)	361 194	1995
Xhosa-Zulu (2/3)	701 143	1995
Tswana-Zulu (full)	337 119	1999
Tswana-Zulu (1/3)	112 373	1999
Tswana-Zulu (2/3)	224 746	1999

Table 6.1: Training and test/eval dataset split sizes per language

(i) Efficiency:

By training models on different fractions of the dataset, we can assess how efficiently each model utilizes the available data. This is crucial for low-resource languages, where obtaining large datasets is challenging.

(ii) Scalability:

Observing the performance of the LLM and the small-scale model as the amount of training data increases helps us understand the data requirements for achieving satisfactory performance, shedding light on the trade-offs between model complexity and training data size.

(2) Assessing performance improvements

(i) Incremental performance gains:

Using dataset splits allows us to identify incremental performance improvements with additional training data, helping to pinpoint the point of diminishing returns where more data does not significantly enhance performance.

(ii) Translation quality:

Evaluating models on various data subsets helps us understand how translation quality improves with more data, setting benchmarks for the minimum data required to achieve acceptable translation quality.

(3) Investigating interpretability and model behavior

(i) Attention mechanisms:

By examining attention heatmaps and patterns for models trained on different dataset sizes, we can analyze how the models' focus changes with more data. This helps in understanding the interpretability of the models and whether additional data leads to more interpretable and reliable attention patterns.

(ii) Consistency in translation:

Evaluating models with different data subsets allows us to investigate if larger training sets lead to more consistent and accurate translations, which is especially important for low-resource languages where linguistic nuances are crucial.

(4) Robustness and generalization

(i) Generalization capability:

Training on varying amounts of data reveals the generalization capabilities of the models. A model that performs well even with one-third of the data might be more robust and adaptable to different datasets or low-resource scenarios.

(ii) Robustness:

Analyzing performance across different dataset sizes helps assess the robustness of the models to fluctuations in training data availability, which is critical for low-resource languages where training data might be limited or unevenly distributed.

(5) Resource and computational considerations

(i) Computational efficiency:

Smaller datasets require less computational power and time to train. By starting with one-third and two-thirds subsets, we can perform preliminary evaluations and adjustments before committing resources to training on the full dataset.

(ii) Feasibility for low-resource settings:

Understanding model performance with limited data and computational resources is crucial for practical applications in low-resource environments, guiding decisions on model deployment and data collection strategies.

Conclusion on dataset splits

Splitting the training dataset into one-third, two-thirds, and the full set provides a comprehensive understanding of how data size impacts model performance and interpretability. This detailed analysis of data efficiency, performance scalability, interpretability, robustness, and computational considerations is essential for the effective application of machine translation models, especially in low-resource language settings.

6.2 Language pairs

Table 6.2 provides the source to target language mappings based on the languages subset from Table 4.2. The actual mappings, e.g. xho-zul, where influenced by the structure of the WMT22 training dataset's parallel translations. For example, the dataset has xho-zul language pair translations, so instead of doing a zul-xho translation, which would simply be back translation, we used the as is state of the dataset.
We specified the number of sentences available for each language pair as per WMT22 dataset in Table 4.2. We applied the same dataset split sizes across both models for training and fine-tuning as documented in Table 6.1.

Source Language	**Target Languages**
Tswana (tsn)	Zulu (zul)
Xhosa (xho)	Zulu (zul)

Table 6.2: Source and target language pair mappings used for model training

6.3 Segmentation with SentencePiece

Both M2M100 and ONMT use SentencePiece[1] for the segmentation of text into subword units. This approach is particularly advantageous for languages like Tswana, Xhosa, and Zulu, which often present challenges for traditional word-based tokenization methods due to their agglutinative nature and complex morphology [70, 71]. These languages frequently create new words through affixation and compounding, resulting in a large number of unique word forms that can overwhelm a model with an overly large vocabulary if whole words are used as tokens.

By breaking down words into smaller subword units, SentencePiece helps to generate more manageable and efficient vocabularies that capture the morphemic structure of these languages more effectively [72]. This results in better representation of linguistic elements and improves the model's ability to handle the nuances of translation for low-resource languages like Tswana, Xhosa, and Zulu.

Furthermore, subword segmentation ensures that rare and unseen words can be processed by recombining known subword units, which enhances the generalization capabilities of the translation models [41]. This approach not only reduces the vocabulary size but also improves the model's ability to translate languages with rich morphology and complex word formation processes, common in Tswana, Xhosa, and Zulu .

[1] https://github.com/google/sentencepiece

6.4 M2M100 fine-tuning parameters

We fine-tuned the M2M100_418M model which has 418 million parameters. The model has 12 layers in both the encoder and decoder and 16 attention heads in the multi-head attention mechanism.
We conducted the fine-tuning on a single GPU NVIDIA RTX A5000.

6.5 OpenNMT training parameters

As mentioned in Section 3, we used the Pytorch implementation of ONMT: OpenNMT-py.
The encoder and decoder stacks are made of 6 layers. There are 8 attention multiheads. The feed-forward neural network dimension is 2048. The experiments were conducted using a single GPU with the specifications: NVIDIA RTX A5000. We performed validation on every 1000 steps. If the validation accuracy does not improve for 4 consecutive checking steps, we stop training using the early stopping strategy.
To mitigate any potential bias that might result from terminating ONMT training runs prematurely and to ensure that the model's performance is not overly dependent on early stopping parameters, we also conducted experiments without specifying early stopping criteria. Instead, we trained the model for a total of 50,000 steps. The results of these training sessions are presented in Section 7 along with other training results.

Chapter 7: Results

This chapter presents the results of training the LLM and small-scale model for the translation tasks for the language pairs Xhosa-Zulu and Tswana-Zulu. We applied the XAI techniques described in Chapter 5 and provide explanations for each methodology in each section, starting with BLEU score evaluation, MMD scores, attention pattern analysis and finally human evaluation.

7.1 BLEU scores

For each language pair (Xhosa-Zulu and Tswana-Zulu), we recorded the BLEU scores generated post training for both the small-scale mode and LLM. We provide an analysis of the scores in the subsequents subsections

7.1.1 Analysis for small-scale model (ONMT)

For the Xhosa-Zulu language pair, the ONMT model's BLEU scores (refer to Figure 7.1) show a significant improvement as the training data increases. Starting from a score of 20.09 with 1/3 of the data, the score rises to 25.56 with 2/3 of the data, and finally reaches 29.59 with the full dataset. Similarly, for the Tswana-Zulu language pair, the BLEU scores improve from 14.38 (1/3 data) to 18.96 (2/3 data), and 22.19 with the full dataset. These results indicate that the ONMT model benefits considerably from larger training sets, suggesting that data volume is a critical factor in enhancing translation quality for small-scale models in low-resource settings.

7.1.2 Analysis for LLLM (M2M100)

The LLM shows an improvement in BLEU scores (refer to Figure 7.1) with increased training data, though the pattern is slightly different. For Xhosa-Zulu, the BLEU scores progress from 25.72 (1/3 data) to 27.08 (2/3 data), and 28.25 with the full dataset. For Tswana-Zulu, the scores increase from 15.26 (1/3 data) to 17.26 (2/3 data), and 18.93 with the full dataset. Although M2M100 shows improvements with more data, the increments are less pronounced compared to ONMT, potentially due to the large-scale model's ability to generalize better even with smaller datasets.

7.1.3 Influence on model interpretability

The BLEU scores (refer to Figure 7.1) reflect not only the translation quality but also influence the interpretability of the models. High BLEU scores generally correlate with translations that are more aligned with reference translations, which aids in easier interpretation of model behavior. For the LLM (M2M100), the relatively stable and high BLEU scores across different data splits suggest a consistent and reliable translation performance, which aligns with the even

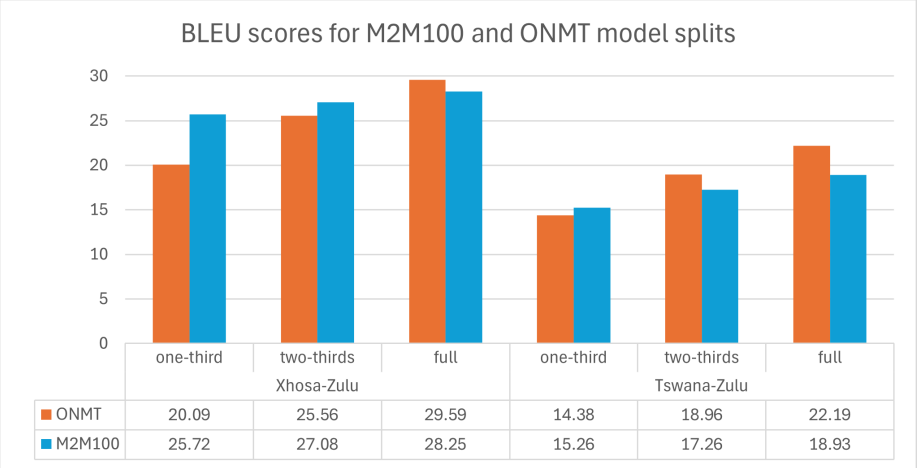

Figure 7.1: The image presents the BLEU scores for Xhosa-Zulu and Tswana-Zulu language pairs, evaluated using two models: M2M100 and OpenNMT (ONMT). These scores are analyzed across different dataset splits (one-third, two-thirds, and full training set split). The BLEU scores offer insight into the translation quality of both LLM and small-scale models for the low-resource language pairs.

attention patterns observed in the model's outputs. These patterns indicate a more balanced focus on source and target sentences, making the model's decision-making process easier to interpret.

In contrast, the small-scale model (ONMT) exhibits more variability in BLEU scores and attention patterns. The uneven and sometimes linear attention patterns seen in ONMT's translations, combined with the significant improvements in BLEU scores with more data, suggest that the model's interpretability is more sensitive to the amount of training data. The ONMT model's reliance on larger datasets to achieve higher BLEU scores indicates potential overfitting issues when data is limited, leading to less reliable and more challenging-to-interpret translations in low-resource scenarios.

In conclusion, the analysis of BLEU scores for both small-scale and large-scale models highlights the importance of data volume in low-resource language translation. While LLMs like M2M100 maintain consistent BLEU scores across different dataset sizes, small-scale models like ONMT require substantial training data to achieve comparable performance. This difference highlights the need for diverse evaluation methods to fully understand translation quality and reliability of models in low-resource language settings.

7.1.4 Training ONMT without early stopping

In attempt to ensure that we allocated enough resources and training iterations to training ONMT so it generates better quality translations for our low-resource language pairs, we carried out experiments using the Xhosa-Zulu language pairs where we trained the model with and without early stopping. We noted the results which are expanded on in the subsequent subsections.

- Training set size: full training set
- Training steps: 50 000
- BLEU score: 30.22

We observed that training ONMT for 50,000 iterations, without specifying early stopping criteria, resulted in a BLEU score increase of 0.63, from 29.59 to 30.22. Training with early stopping specified took 3 hours, 44 minutes, and 2 seconds, whereas training for 50,000 iterations without early stopping took 10 hours, 18 minutes, and 30 seconds. Although we could have extended the training time to determine if the model's performance would improve with more iterations, the minimal increase in translation quality led us to decide against further resource allocation. This could be explored in future research.

We plotted two line graphs to illustrate the learning rate captured by the model after every 1,000 iterations, to assess the model's learning progress as training steps increased. When using the full training set with early stopping, ONMT stopped training after 18,000 steps, with a learning rate of 0.00099 at 1,000 steps and 0.00066 at 18,000 steps, as shown in Figure 7.2. Early stopping ensured efficient training by halting once improvements diminished. In contrast, training without early stopping continued for 50,000 steps, with the learning rate decreasing from 0.00099 at the start to 0.0004 at 50,000 steps, as shown in Figure 7.3. We also translated four Xhosa source sentences to Zulu post training with and without early stopping with the results in Table 7.1. Despite the increase in training iterations, there is barely any improvement in the machine translation quality. Given this analysis, this research trains ONMT with early stopping across all training set sizes as allocating more resources to training would be wasteful.

In the following sections, we take a look at model attention using attention heatmaps and attention pattern analysis.

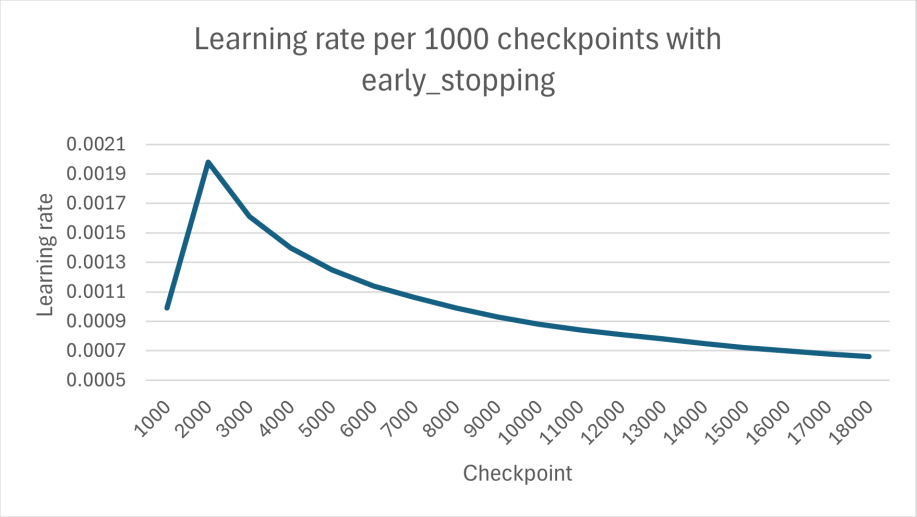

Figure 7.2: The learning rate peaks early at 2,000 checkpoints (0.0021) and declines steadily over 18,000 checkpoints. Early stopping ensures efficient training by stopping if there is no improvements after three iterations, preventing overfitting and optimizing resource use.

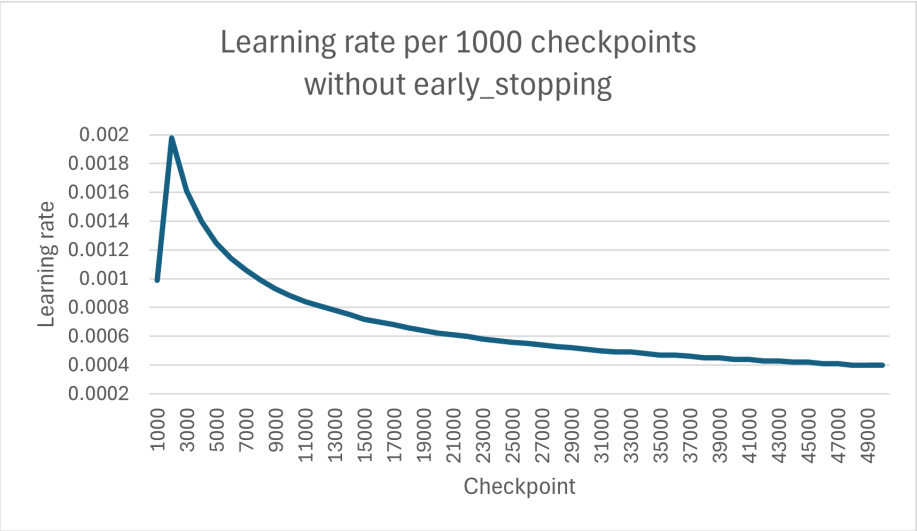

Figure 7.3: The learning rate peaks early but declines gradually over 49,000 checkpoints, shown by the declining curve. The absence of early stopping here highlights wasted computational resources with minimal improvement on translation quality in comparison to Figure 7.2.

Source Xhosa sentences	**Zulu translations with early stopping**	**Zulu translations without early stopping**
uharrison ubalisa ibali lesinye sezigulana zasemosa maria.	Izithandani Zokubhangqa Ngababili	Ukushintshanisa okungcono kakhulu kwe- Compounding
akukho nkxwaleko inokuba ngaphezulu kweyomzali onabantwana abakhala kuye befuna ukutya abe yena engazi nokuba uza kukufumana phi na oko kutya.	Akukho nkathazo ekuphenyeni nasekuphenyeni	Ayikho imiphumela ebuhlungu kunazo zonke emhlabeni wonke!
sisenzo esibaluleke kakhulu kuso nasiphi na isizwe esisekelwe kumba wokuhlonipha amalungelo oluntu.	Ukuhambelana Okubanzi kokulingana	Kunzima okulingene nesifuna ukuhlala
singurhulumente asikhange simeme ukuba kudanjiswe okanye kuthuliswe oko kugxekwa.	Isibikezelo seinthanethi noma sokubikezela	Ilayisi abilisiwe

Table 7.1: A comparison of the machine translations generated from training ONMT with and without the early stopping criteria on the full training set for the Xhosa-Zulu language pairs. There is minimal translation quality improvement whether early stopping is specified or not.

7.2 Attention heatmaps and patterns

In this section, we sampled one source sentence and its target sentence from the Vuk'zenzele dataset[2]. The source sentence is translated using both the LLM and small-scale model and the attention heatmap of the translation generated the final decoder layer. For each language pair, we present the attention pattern plots (Figures 7.4 to 7.9), the source and target sentence, the MT sentence by M2M100 and ONMT, and the attention pattern analysis for each of the one-third, two-third and full set data splits as defined in Section 6.1.

7.2.1 Xhosa-Zulu attention heatmaps for the one-third training set

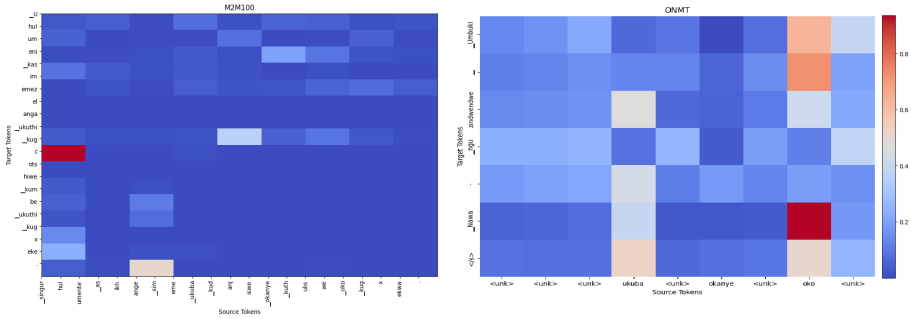

Figure 7.4: Xhosa-Zulu attention heatmaps for source and target using one-third of the training set.

Source Xhosa sentence: singurhulumente asikhange simeme ukuba kudanjiswe okanye kuthuliswe oko kugxekwa.

Target Zulu sentence: Uhulumeni akazange ameme ukuba kugcotshwe noma kuthululwe lokho kugxekwa.

M2M100: Uhulumeni kasimemezelanga ukuthi kugcotshiwe kumbe ukuthi kugxeke. *(left)*

ONMT: Umbuki zindwendwe ngu- Nawa *(right)*

Attention patterns: An evenly distributed attention heatmap with mostly the same color and a few bright spots suggests the LLM (M2M100) broadly spreads its focus across the sentence, occasionally emphasizing keywords. In contrast, the small-scale model's heatmap, with two bright columns and shades of blue, indicates a concentrated focus on specific parts of the sentence, suggesting a more deterministic attention mechanism. This difference reflects the LLM's ability to capture broader context versus the small-scale model's reliance on specific words for translation.

7.2.2 Xhosa-Zulu attention heatmaps for the two-thirds training set

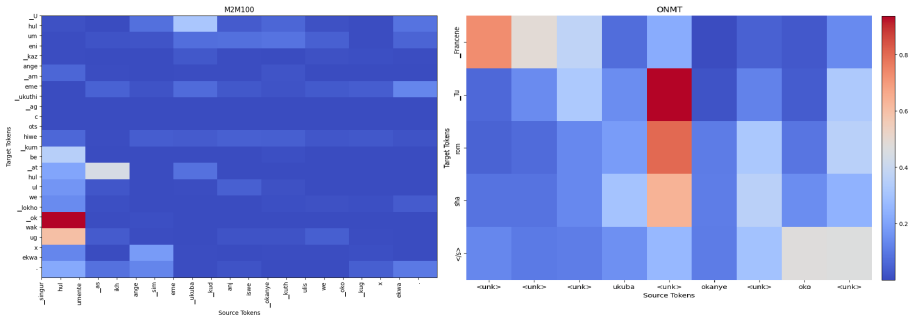

Figure 7.5: Xhosa-Zulu attention heatmaps for source and target using two-thirds of the training set.

Source Xhosa sentence: singurhulumente asikhange simeme ukuba kudanjiswe okanye kuthuliswe oko kugxekwa.

Target Zulu sentence: Uhulumeni akazange ameme ukuba kugcotshwe noma kuthululwe lokho kugxekwa.

M2M100: Uhulumeni kazange ameme ukuthi agcotshiwe kumbe athululwe lokho okwakugxekwa. *(left)*

ONMT: Francene Turomsha *(right)*

Attention patterns: The LLM heatmap shows evenly distributed attention with minor outliers, indicating a balanced focus across the entire sentence while occasionally emphasizing specific words, suggesting a comprehensive approach to translation. In contrast, the small-scale model's heatmap displays partially diagonal attention for the first three-quarters of the sentence, indicating a focus on sequentially related words. The last quarter of the sentence shifts to a linear pattern, highlighting a different processing strategy that may prioritize individual words or phrases independently.

7.2.3 Xhosa-Zulu attention heatmaps for the full training set

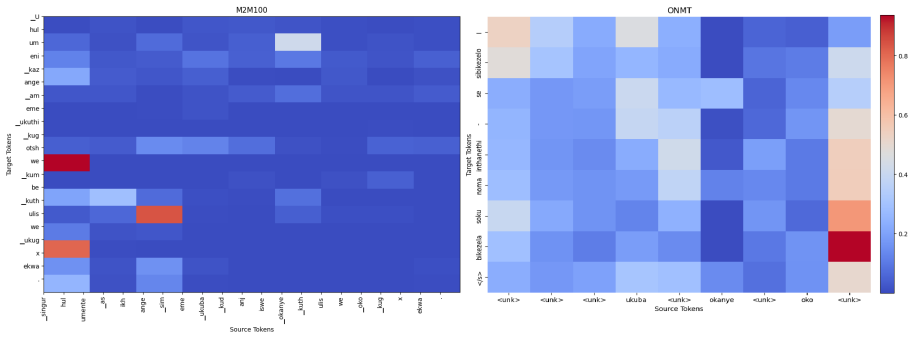

Figure 7.6: Xhosa-Zulu attention heatmaps for source and target using the full training set.

Source Xhosa sentence: singurhulumente asikhange simeme ukuba kudanjiswe okanye kuthuliswe oko kugxekwa.

Target Zulu sentence: Uhulumeni akazange ameme ukuba kugcotshwe noma kuthululwe lokho kugxekwa.

M2M100: Uhulumeni kazange ameme ukuthi kugotshwe kumbe kuthuliswe ukugxekwa. *(left)*

ONMT: Isibikezelo se-inthanethi noma sokubikezela *(right)*

Attention patterns: The LLM heatmap shows evenly distributed attention with minor outliers, indicating a balanced focus across the entire sentence while occasionally emphasizing specific words, suggesting a comprehensive approach to translation. The small-scale model's heatmap displays sparsely medium-distributed attention for the first half of the sentence, low attention for two-thirds of the remaining half, and higher attention towards the end of the source sentence. This pattern reveals that the small-scale model allocates its focus unevenly, concentrating more on specific segments of the sentence. This suggests that the small-scale model may be prioritizing certain parts of the input more heavily, potentially due to limited capacity or differing translation strategies, highlighting areas where the model might struggle or excel in processing linguistic features.

7.2.4 Tswana-Zulu attention heatmaps for the one-third training set

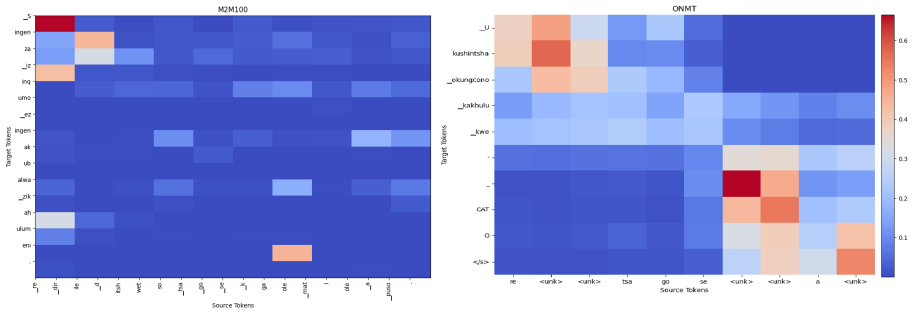

Figure 7.7: Tswana-Zulu attention heatmaps for source and target using one-third of the training set.

Source Tswana sentence: re dirile ditshwetso tsa go se kgaole matlole a puso.

Target Zulu sentence: sesithathe isinqumo soku ngaqhubeki nokunciphisa isabelomali sikahulumeni ngokukhuphula intela.

M2M100: Singenza izinqumo ezingenakubalwa zikahulumeni. *(left)*

ONMT: Ukushintshanisa okungcono kakhulu kwe- CATO *(right)*

Attention patterns: The LLM heatmap shows somewhat distributed attention with some outliers, meaning the model generally maintains a balanced focus across most of the sentence, but there are certain words or segments that receive significantly more attention. This suggests that while the LLM is broadly consistent in its attention, it also identifies and emphasizes specific words or phrases that it deems particularly important for the translation task. The small-scale model's heatmap displays a diagonal attention pattern, with higher attention in the top-left and bottom-right quadrants and lower attention in the top-right and bottom-left quadrants. This pattern suggests that the small-scale model is focusing more intensely on certain parts of the sentence at the beginning and end while paying less attention to the middle parts. This indicates that the small-scale model might be emphasizing the initial and final segments of the sentence, potentially capturing key elements at these positions but possibly overlooking important contextual information in the middle.

7.2.5 Tswana-Zulu attention heatmaps for the two thirds training set

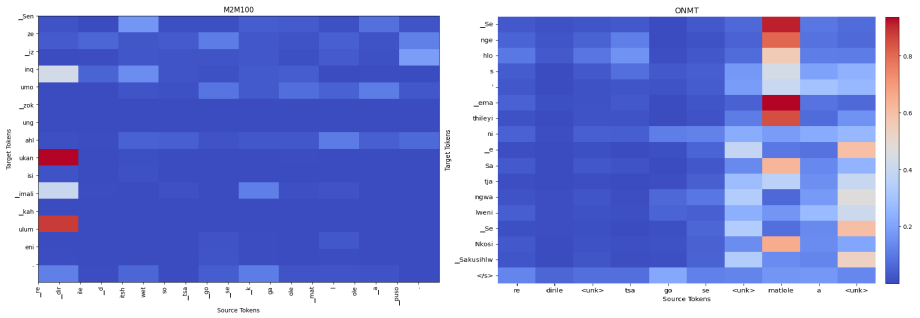

Figure 7.8: Tswana-Zulu attention heatmaps for source and target using two thirds of the training set.
Source Tswana sentence: re dirile ditshwetso tsa go se kgaole matlole a puso.
Target Zulu sentence: sesithathe isinqumo soku ngaqhubeki nokunciphisa isabelomali sikahulumeni ngokukhuphula intela.
M2M100: Senze izinqumo zokungahlukanisi imali kahulumeni. *(left)*
ONMT: Sengehlos' emathileyini eSatjangwalweni SeNkosi Sakusihlwa *(right)*
Attention patterns: The LLM displays constant attention with minor outliers, indicating a balanced focus across most of the sentence with occasional emphasis on specific words. In contrast, the small-scale model shows consistently lower attention for the left half of the sentence and relatively medium to high attention for the right half, suggesting it places more importance on the latter part of the sentence. This pattern implies that the small-scale model might prioritize later contextual information, potentially impacting translation accuracy and coherence. This focus results in a completely incorrect translation for the small-scale model.

7.2.6 Tswana-Zulu attention heatmaps for the full training set

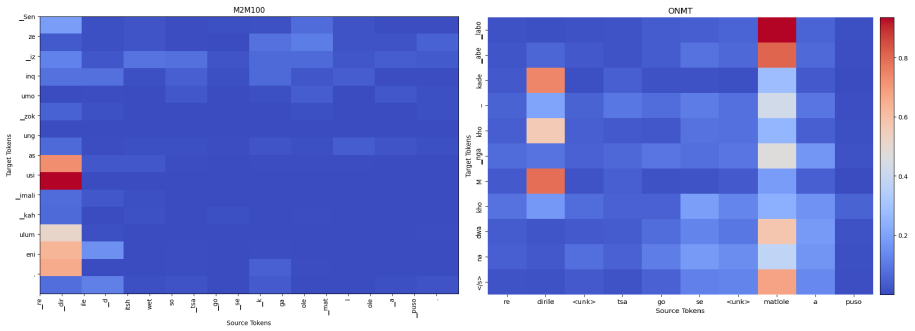

Figure 7.9: Tswana-Zulu attention heatmaps for source and target using full training set.

Source Tswana sentence: re dirile ditshwetso tsa go se kgaole matlole a puso.

Target Zulu sentence: sesithathe isinqumo soku ngaqhubeki nokunciphisa isabelomali sikahulumeni ngokukhuphula intela.

M2M100: Senze izinqumo zokungasusi imali kahulumeni. *(left)*

ONMT: labo abekade kho ngaMkhodwana *(right)*

Attention patterns: The LLM exhibits constant attention with minor outliers, indicating a balanced focus across the sentence with occasional emphasis on specific words. The small-scale model focuses heavily on specific positions within the sentence (second and second-last columns), indicating that these positions are critical for understanding or generating the translation. However, the rest of the sentence receives much less attention. This lack of balanced focus could lead to translation issues in capturing the overall context hence the ONMT translation greatly differs from the target Zulu sentence.

The following section takes a closer look at the MMD scores computed between the target Vuk'zenzele sentence and generated MT sentence from both the LLM and small-scale model.

7.3 MMD distance metrics

As per methodology discussion in Section 5.2.4, we will take a closer look at the MMD scores derived from model translations to target sentence similarity computations. Higher MMD scores (close to 1) indicate that the model translations are very similar to the target sentences while lower MMD scores (closer to 0) indicate that the model translation fail to align with the target sentences.

7.3.1 Xhosa-Zulu MMD scores

The scores in Figure 7.2 from the Xhosa-Zulu language pair suggest that the translations produced by the model trained with two-thirds of the data are generally more accurate, exhibiting lower MMD scores compared to those trained on the full and one-third datasets. This trend implies that the model achieves a balance between the amount of training data and translation quality, with the two-thirds training set providing the optimal data size for effective learning.
Moreover, the overall low MMD scores indicate that the LLM (M2M100) produces translations that are very similar to the target sentences, reflecting the model's robustness and capability in handling Xhosa-Zulu translations. The lower MMD scores also highlight the model's ability to maintain consistency across different training set sizes, with translations remaining close to the desired output. These findings highlight the importance of selecting an appropriate amount of training data to enhance the performance of NMT models for low-resource language pairs.
Figure 7.3 shows that ONMT consistently produces higher MMD scores whose values are greater than 0.05 thus much closer to 1. This indicates that its translations are generally less similar to the target sentences.
Despite the higher MMD scores, there are instances where the ONMT models generate translations containing words that appear in the target sentences. However, these occurrences are not frequent enough to offset the overall trend of less accurate translations.
These MMD results align with attention pattern analysis results, which highlighted that the ONMT models may have difficulty properly assigning attention during the translation process. This observation highlights the challenges faced by the small-scale models in achieving high-quality translations for low-resource language pairs, prompting the need for further refinement and optimization of such models

Sentence	**MMD Score**
Target: uharrison uxoxa udaba ngesinye seziguli zasemosa maria.	
1/3 : U-Harrison ulandisa indaba yesinye seziguli zasemoyeni uMariya.	0.00367
2/3 : UHarrison ulandisa indaba yesinye seziguli zikaMaria.	0.00205
Full: UHarrison ulandisa ngendaba yesinye seziguli zasemosa maria.	0.00250
Target: alukho usizi olwedlula olomzali okhalelwa abantwana befuna ukudla, kodwa benge nalutho abangabanika lona.	
1/3 : Akukho ukuhlupheka okungaba ngaphezu kwabazali abanezingane ezikhala kuye befuna ukudla abe yena engazi ukuthi uzokutholaphi lokho kudla.	0.01277
2/3 : Akukho ukuhlupheka okungaba ngaphezu komzali onabantwana abakhala kuye befuna ukudla futhi yena engazi ukuthi uzokuthola kuphi lokho kudla.	0.01585
Full: Akukho ukuhlupheka okungaba ngaphezu komzali onabantwana abakhala kuye befuna ukudla futhi yena engazi nokuthi uzokuthola kuphi lokho kudla.	0.01478
Target: kubalulekile kunoma umuphi umphakathi ophila ngokuhlonipha amalungelo abantu.	
1/3 : yisenzo esibaluleke kunazo zonke kunoma yisiphi isizwe esisekelwe ekuhlonipheni amalungelo abantu.	0.0080
2/3 : yisenzo esibaluleke kakhulu kunoma yiliphi isizwe esisekelwe endabeni yokuhlonipha amalungelo abantu.	0.0057
Full: Kuyisenzo esibaluleke kakhulu kunoma yisiphi isizwe esisekelwe endabeni yokuhlonipha amalungelo abantu.	0.0061
Target: njengohulumeni asifuni ukuthi lokhu kugxeka kube nentukuthelo noma kuthuliswe.	
1/3 : Uhulumeni kasimemezelanga ukuthi kugcotshiwe kumbe ukuthi kugxeke.	0.00287
2/3 : Uhulumeni kazange ameme ukuthi agcotshiwe kumbe athululwe lokho okwakugxekwa.	0.00247
Full: Uhulumeni kazange ameme ukuthi kugotshwe kumbe kuthuliswe ukugxekwa.	0.00380

Table 7.2: This table presents the MMD scores for the LLM (M2M100) translations against the Vuk'zenzele target sentences for Xhosa-Zulu translations using different the different training data splits: one-third, two-thirds, and the full training set.

Sentence	**MMD Score**
Target: uharrison uxoxa udaba ngesinye seziguli zasemosa maria.	
1/3 : Kuxoxwa ngakho.	0.05838
2/3 : Francene Turomsha	0.11612
Full: Izithandani Zokubhangqa Ngababili	0.11493
Target: alukho usizi olwedlula olomzali okhalelwa abantwana befuna ukudla, kodwa benge nalutho abangabanika lona.	
1/3 : Akukho lutho olubi kokubhekana nosizi	0.04677
2/3 : Akunankinga nokuthinta ngokweqile; akubekezeleleki - Kungakhathaliseki ukuthi yikuphi ukudla okungaphumelela.	0.03612
Full: Akukho nkathazo ekuphenyeni nasekuphenyeni	0.05978
Target: kubalulekile kunoma umuphi umphakathi ophila ngokuhlonipha amalungelo abantu.	
1/3 : Okubaluleke kakhulu ku ??	
2/3 : Okuningi ngokuthinta ngokujulile Umnenke ??	0.05223
Full: Ukuhambelana Okubanzi kokulingana	0.05372
Target: njengohulumeni asifuni ukuthi lokhu kugxeka kube nentukuthelo noma kuthuliswe.	
1/3 : Umbuki zindwendwe ngu- Nawa	0.05821
2/3 : Francene Turomsha	0.13680
Full: Isibikezelo se-inthanethi noma sokubikezela.	0.04665

Table 7.3: This table displays the MMD scores for the ONMT translations against the Vuk'zenzele target sentences for Xhosa-Zulu.

7.3.2 Tswana-Zulu MMD scores

The MMD scores for the LLM for Tswana-Zulu MT outputs in Table 7.4 indicate that the model trained with two-thirds of the training set generally achieves the lowest MMD scores, suggesting higher translation accuracy.
This trend highlights the effectiveness of the M2M100 model in handling Tswana-Zulu translations, particularly when provided with an optimal amount of training data. The findings emphasize the importance of the quantity and quality of training data in enhancing the model's translation performance, highlighting that a balanced training dataset can significantly improve the accuracy and reliability of translations for low-resource language pairs.
Similar to the results observed for the Xhosa-Zulu models in 7.3, the MMD scores for the small-scale model Tswana-Zulu translations are relatively higher, indicating poor translation quality. The higher MMD scores, as shown in Table 7.5, suggest that the ONMT model's translations are less similar to the target sentences, reflecting significant discrepancies in translation accuracy. These results are further supported by human evaluations, which consistently rated the translations as incorrect.
Overall, these findings demonstrate that small-scale models struggles to achieve the same level of translation quality as LLMs emphasizing the importance of robust training data and model architecture in enhancing translation accuracy.

Sentence	**MMD Score**
Target: sesithathe isinqumo soku ngaqhubeki nokunciphisa isabelomali sikahulumeni ngokukhuphula intela.	
1/3: Singenza izinqumo ezingenakubalwa zikahulumeni.	0.09379
2/3: Senze izinqumo zokungahlukanisi imali kahulumeni.	0.00641
Full: Senze izinqumo zokungasusi imali kahulumeni	0.00710
Target: ngenxa yalokhu, silindele ukuncipha kwengenisomali ebanjelwe ngezigidigidi ezi ngamar156 ekusetshenzi sweni kwemali okungena nzalo esikhathini esiphakathi.	
1/3: Ngenxa yalokho, sithemba ukuhlukunyezwa okungenani izigidi ezingama- r156 zamarandi phakathi konyaka.	0.00948
2/3: Ngenxa yalokhu, sinethemba lokuba nesilinganiso sokungabhadali izigidigidi zamarandi ezingu-1156 kulo nyaka.	0.00533
Full: Ngenxa yalokhu, sinomuzwa wokuthi asinakekeli izigidigidi zamarandi eziyi-1156 zamarandi kuleli sonto.	0.00842
Target: lokhu kuzosiza ukunci phisa ukuntuleka kwemali nokunciphisa izidingo zoku thi siboleke.	
1/3: Lokhu kuzosinceda ukuthi sinciphise izikweletu futhi sinciphise imali.	0.02820
2/3: Lokhu kuzosisiza sinciphise izikweletu futhi sinciphise izindleko zemali.	0.00774
Full: Lokhu kuzosisiza sinciphise izikweletu futhi sinciphise izindleko zezimali.	0.00783
Target: ukuqeda indlala akusona isenzo sokupha.	
1/3: Ukukhipha isinyathelo akuyona isinyathelo.	0.00367
2/3: Ukuqeda indlala akusona isinyathelo sokuba nomusa.	0.00246
Full: Ukuphela kwendlala akusona isinyathelo sobubele.	0.00281

Table 7.4: This table presents the MMD scores for the M2M100 models' translations against the Vuk'zenzele target sentences for Tswana-Zulu.

Sentence	**MMD Score**
Target: sesithathe isinqumo soku ngaqhubeki nokunciphisa isabelomali sikahulumeni ngokukhuphula intela.	
1/3: Ukushintshanisa okungcono kakhulu kwe- CATO	0.06409
2/3: Sengehlos' emathileyini eSatjangwalweni SeNkosi Sakusihlwa	0.08753
Full: labo abekade kho ngaMkhodwana	0.08983
Target: ngenxa yalokhu, silindele ukuncipha kwengenisomali ebanjelwe ngezigidigidi ezi ngamar156 ekusetshenzi sweni kwemali okungena nzalo esikhathini esiphakathi.	0.52728
1/3: UDe Lille uyalelwe ukuba athathe amasakana akhe aphume ehhovisi lomkhandlu	0.52728
2/3: komkhawulo wengeza	0.53836
Full: Bekushiyana ukwenza ama-megas	0.50508
Target: lokhu kuzosiza ukunciphisa ukuntuleka kwemali nokunciphisa izidingo zoku thi siboleke.	
1/3: Izikhangibavakashi for Kids e-Italy	0.10220
2/3: Hlekhi Msiza	0.13309
Full: Madelene Mayeshiba	0.22532
Target: ukuqeda indlala akusona isenzo sokupha.	
1/3: Ifulethi lonke e-Hebden Bridge	0.07630
2/3: Dala bahlanza nsuku zonke	0.06815
Full: -qeda qugwala	0.05588

Table 7.5: This table presents the MMD scores for the ONMT model translations.

7.4 Human evaluations

We utilized two distinct groups of evaluators, each group consisting of three individuals. The evaluators were carefully selected based on their linguistic proficiency and professional expertise to ensure comprehensive and unbiased evaluation results.

Group one: native speakers

The first group comprised native speakers of Tswana, Xhosa, and Zulu. These three evaluators were tasked with assessing the correctness of the machine translations generated by the small-scale models (ONMT) and the LLM (M2M100) models. Their primary responsibility was to determine whether the translations were accurate and contextually appropriate, as discussed in Section 7.4.1. To ensure controlled and consistent evaluation outcomes, these native speakers were also asked to complete the MQM questionnaire, with detailed results provided in subsection 7.4.2. This dual evaluation approach helped to capture both subjective assessments of translation quality and objective scoring metrics.

Group two: machine translation and software development Experts

The second group consisted of fellow researchers with backgrounds in machine translation and software development. These evaluators were selected to evaluate the effectiveness and relevance of the XAI methods applied in this research. Given their expertise, they were able to provide informed feedback on the applicability of these methods in evaluating translation correctness, as reported by the native speaker evaluators.

We sought their input to determine how satisfied they were with the XAI evaluation techniques, including attention heatmaps and BLEU scores, which might not have been fully understood by the native speakers due to their technical nature. Involving researchers in this evaluation process is crucial, as they are likely to use such methods in future research or practical applications where human evaluators may not always be available due to time or cost constraints. The insights from this group are crucial for understanding the broader implications and usability of these XAI methods in low-resource language contexts. The results and discussions on their feedback are elaborated on in subsection 7.4.3.

The involvement of these two distinct groups allowed us to triangulate the research findings, combining linguistic expertise with technical understanding to provide a well-rounded evaluation of the translation models and their outputs.

7.4.1 Translation correctness

Xhosa-Zulu translation evaluations

From Table 7.6, we observe that the LLM model trained on the full dataset produced the most accurate translations for Xhosa-Zulu translations, closely aligning with the target sentences. Despite the MMD scores suggesting otherwise (refer to Tables 7.2 and 7.4), the full dataset model achieved correct translations more consistently. Errors in the one-third and two-thirds datasets were typically due to a few mistranslated words (red colored Table 7.6), which significantly altered the meaning of the sentences. This highlights the model's sensitivity to training data size and the importance of accurate word-level translations for maintaining overall sentence integrity.
The small-scale model, as shown in Table 7.7, produced incorrect translations across all dataset split models. The highlighted table portion illustrates the models' attempts to generate partial translations, which still failed to convey the correct meaning. This highlights the challenges faced by the ONMT models in accurately translating low-resource language pairs.

Sentence	Evaluation
Source: uharrison ubalisa ibali lesinye sezigulana zasemosa maria.	
Target: UHarrison ulandisa indaba yesinye seziguli zasemosa maria.	
1/3 : U-Harrison ulandisa indaba yesinye seziguli zasemoyeni uMariya.	Incorrect
2/3 : UHarrison ulandisa indaba yesinye seziguli zikaMaria.	Incorrect
Full: UHarrison ulandisa ngendaba yesinye seziguli zasemosa maria.	Correct
Source: akukho nkxwaleko inokuba ngaphezulu kweyomzali onabantwana abakhala kuye befuna ukutya abe yena engazi nokuba uza kukufumana phi na oko kutya.	
Target: Akukho ukuhlupheka okungaba ngaphezu komzali onabantwana abakhala kuye befuna ukudla futhi engazi ukuthi uzokuthola kuphi lokho kudla.	
1/3 : Akukho ukuhlupheka okungaba ngaphezu kwabazali abanezingane ezikhala kuye befuna ukudla abe yena engazi ukuthi uzokutholaphi lokho kudla.	Incorrect
2/3 : Akukho ukuhlupheka okungaba ngaphezu komzali onabantwana abakhala kuye befuna ukudla futhi yena engazi nokuthi uzokuthola kuphi lokho kudla.	Incorrect
Full: Akukho ukuhlupheka okungaba ngaphezu komzali onabantwana abakhala kuye befuna ukudla futhi yena engazi ukuthi uzokuthola kuphi lokho kudla.	Correct
Source: sisenzo esibaluleke kakhulu kuso nasiphi na isizwe esisekelwe kumba wokuhlonipha amalungelo oluntu.	
Target: Kuyisenzo esibaluleke kakhulu kunoma yisiphi isizwe esisekelwe endabeni yokuhlonipha amalungelo abantu.	
1/3 : yisenzo esibaluleke kunazo zonke kunoma yisiphi isizwe esisekelwe ekuhlonipheni amalungelo abantu.	Incorrect
2/3 : yisenzo esibaluleke kakhulu kunoma yiliphi isizwe esisekelwe endabeni yokuhlonipha amalungelo abantu.	Incorrect
Full: Kuyisenzo esibaluleke kakhulu kunoma yisiphi isizwe esisekelwe endabeni yokuhlonipha amalungelo abantu.	Correct
Source: singurhulumente asikhange simeme ukuba kudanjiswe okanye kuthuliswe oko kugxekwa.	
Target: Uhulumeni akazange ameme ukuba kugcotshwe noma kuthululwe lokho kugxekwa.	
1/3 : Uhulumeni kasimemezelanga ukuthi kugcotshiwe kumbe ukuthi kugxeke.	Incorrect
2/3 : Uhulumeni kazange ameme ukuthi agcotshiwe kumbe athulwe lokho okwakugxekwa.	Incorrect
Full: Uhulumeni kazange ameme ukuthi kugotshwe kumbe kuthuliswe ukugxekwa.	Correct

Sentence	**Evaluation**
Source: uharrison ubalisa ibali lesinye sezigulana zasemosa maria.	
Target: UHarrison ulandisa indaba yesinye seziguli zasemosa maria.	
1/3: Kuxoxwa ngakho.	Incorrect
2/3: Francene Turomsha	Incorrect
Full: Izithandani Zokubhangqa Ngababili	Incorrect
Source: akukho nkxwaleko inokuba ngaphezulu kweyomzali onabantwana abakhala kuye befuna ukutya abe yena engazi nokuba uza kukufumana phi na oko kutya.	
Target: Akukho ukuhlupheka okungaba ngaphezu komzali onabantwana abakhala kuye befuna ukudla futhi engazi ukuthi uzokuthola kuphi lokho kudla.	
1/3: Akukho lutho olubi kokubhekana nosizi	Incorrect
2/3: Akunankinga nokuthinta ngokweqile; akubekezeleleki - Kungakhathaliseki ukuthi yikuphi ukudla okungaphumelela.	Incorrect
Full: Akukho nkathazo ekuphenyeni nasekuphenyeni	Incorrect
Source: sisenzo esibaluleke kakhulu kuso nasiphi na isizwe esisekelwe kumba wokuhlonipha amalungelo oluntu.	
Target: Kuyisenzo esibaluleke kakhulu kunoma yisiphi isizwe esisekelwe endabeni yokuhlonipha amalungelo abantu.	
1/3: Okubaluleke kakhulu ku ??	Incorrect
2/3: Okuningi ngokuthinta ngokujulile Umnenke ??	Incorrect
Full: Ukuhambelana Okubanzi kokulingana	Incorrect
Source: singurhulumente asikhange simeme ukuba kudanjiswe okanye kuthuliswe oko kugxekwa.	
Target: Uhulumeni akazange ameme ukuba kugcotshwe noma kuthululwe lokho kugxekwa.	
1/3: Umbuki zindwendwe ngu- Nawa	Incorrect
2/3: Francene Turomsha	Incorrect
Full: Isibikezelo se-inthanethi noma sokubikezela	Incorrect

Table 7.7: Translations by ONMT for each of the one-third, two-thirds, and full training set splits for Xhosa-Zulu.

Tswana-Zulu translation evaluations

The two-thirds and full training set models produced correct translations more often than the one-third training set model for the LLM as shown in Table 7.8. However, for the second sentence, both the two-thirds and full training sets fail to translate the currency representation from r156 to 1156, highlighting a weakness in the LLMs handling of numeric translations for the language pairs, making otherwise correct translations incorrect.
The first and third source sentences have translations from the full set that are a mixture of Xhosa and Zulu. For example, *senze izinqumo zokungasusi imali kahulumeni* would be correct in informal use between natives, but since 'susi' is Xhosa. Classifying this as a correct Zulu translation would be incorrect. Mistranslations for the Tswana-Zulu language pairs by the LLM are represented in red in Table 7.8.
Table 7.9 shows the Tswana-Zulu translations for the small-scale model across the different dataset splits. All translations were evaluated as incorrect. The Tswana-Zulu models, regardless of the training set size, fail to produce correct translations. Compared to the Xhosa-Zulu models, the Tswana-Zulu language pair had significantly smaller training set sizes resulting in the models being unable to form even partial translations, as was the case in the Xhosa-Zulu models in Table 7.7. This highlights the importance of increasing the size and quality of datasets for low-resource languages to ensure better translations.

Sentence	**Evaluation**
Source: re dirile ditshwetso tsa go se kgaole matlole a puso.	
Target: sesithathe isinqumo soku ngaqhubeki nokunciphisa isabelomali sikahulumeni ngokukhuphula intela.	
1/3: Singenza izinqumo ezingenakubalwa zikahulumeni	Incorrect
2/3: Senze izinqumo zokungahlukanisi imali kahulumeni.	Correct
Full: Senze izinqumo zokungasusi imali kahulumeni.	Incorrect
Source: ka ntlha ya seno, re solofela go nna le phokotso ya go se duelele dinamane ya bokanaka r156 bilione mo pakagareng ya monongwaga.	
Target: ngenxa yalokhu, silindele ukuncipha kwengenisomali ebanjelwe ngezigidigidi ezi ngamar156 ekusetshenzi sweni kwemali okungena nzalo esikhathini esiphakathi.	
1/3: Ngenxa yalokho, sithemba ukuhlukunyezwa okungenani izigidi ezingama- r156 zamarandi phakathi konyaka.	Incorrect
2/3: Ngenxa yalokhu, sinethemba lokuba nesilinganiso sokungabhadali izigidigidi zamarandi ezingu-1156 kulo nyaka.	Incorrect
Full: Ngenxa yalokhu, sinomuzwa wokuthi asinakekeli izigidigidi zamarandi eziyi-1156 zamarandi kuleli sonto.	Incorrect
Source: seno se tla re thusa go fokotsa dikoloto le go fokotsa dikadimo tsa madi.	
Target: lokhu kuzosiza ukunci phisa ukuntuleka kwemali nokunciphisa izidingo zoku thi siboleke.	
1/3: Lokhu kuzosinceda ukuthi sinciphise izikweletu futhi sinciphise imali.	Incorrect
2/3: Lokhu kuzosisiza sinciphise izikweletu futhi sinciphise izindleko zemali.	Correct
Full: Lokhu kuzosisiza sinciphise izikweletu futhi sinciphise izindleko zezimali.	Correct
Source: go fedisa tlala ga se kgato ya kutlwelobotlhoko.	
Target: ukuqeda indlala akusona isenzo sokupha.	
1/3: Ukukhipha isinyathelo akuyona isinyathelo.	Incorrect
2/3: Ukuqeda indlala akusona isinyathelo sokuba nomusa.	Correct
Full: Ukuphela kwendlala akusona isinyathelo sobubele.	Correct

Table 7.8: This table shows the translations by M2M100 for the one-third, two-thirds, and full training set splits for Tswana-Zulu.

Sentence	**Evaluation**
Source: re dirile ditshwetso tsa go se kgaole matlole a puso.	
Target: sesithathe isinqumo soku ngaqhubeki nokunciphisa isabelomali sikahulumeni ngokukhuphula intela.	
1/3: Ukushintshanisa okungcono kakhulu kwe- CATO	Incorrect
2/3: Sengehlos' emathileyini eSatjangwalweni SeNkosi Sakusihlwa	Incorrect
Full: labo abekade kho ngaMkhodwana	Incorrect
Source: ka ntlha ya seno, re solofela go nna le phokotso ya go se duelele dinamane ya bokanaka r156 bilione mo pakagareng ya monongwaga.	
Target: ngenxa yalokhu, silindele ukuncipha kwengenisomali ebanjelwe ngezigidigidi ezi ngamar156 ekusetshenzi sweni kwemali okungena nzalo esikhathini esiphakathi.	
1/3: UDe Lille uyalelwe ukuba athathe amasakana akhe aphume ehhovisi lomkhandlu	Incorrect
2/3: komkhawulo wengeza	Incorrect
Full: Bekushiyana ukwenza ama-megas	Incorrect
Source: seno se tla re thusa go fokotsa dikoloto le go fokotsa dikadimo tsa madi.	
Target: lokhu kuzosiza ukunciphisa ukuntuleka kwemali nokunciphisa izidingo zoku thi siboleke.	
1/3: Izikhangibavakashi for Kids e-Italy	Incorrect
2/3: Hlekhi Msiza	Incorrect
Full: Madelene Mayeshiba	Incorrect
Source: go fedisa tlala ga se kgato ya kutlwelobotlhoko.	
Target: ukuqeda indlala akusona isenzo sokupha.	
1/3: Ifulethi lonke e-Hebden Bridge	Incorrect
2/3: Dala bahlanza nsuku zonke	Incorrect
Full: -qeda qugwala	Incorrect

Table 7.9: Translations by ONMT for each of the one-third, two-thirds, and full training set splits for Tswana-Zulu.

7.4.2 MQM

Section 2.6.2 introduced us to MQM as a human evaluation method for nuanced translation quality with the exact MQM questionnaire used provided in Appendix B.1. This section takes a closer look at the LLM (M2M100) and small-scale model (ONMT) across the one-third two-thirds and full dataset training splits, providing a summarized view of the human feedback. For each dataset split, we have an of the user rankings explanation followed by a plot thereof starting with M2M100 then ONMT.

M2M100 translation quality evaluation

Xhosa-Zulu evaluation

The translations from Tswana to Zulu were evaluated across different training set sizes: full set, two-thirds set, and one-third set. The metrics assessed were accuracy, fluency, terminology, locale convention, and style. The results are detailed as follows:

Full training set - Figure 7.10

- **Accuracy - 5**: The translations were highly accurate, closely aligning with the intended meaning in the target language.
- **Fluency - 5**: Translations were very fluent, demonstrating natural and grammatically correct language usage.
- **Terminology - 5**: The terminology used in the translations was appropriate and correctly applied, maintaining consistency with domain-specific vocabulary.
- **Locale convention - 5**: The translations adhered strictly to locale-specific conventions, such as cultural and regional nuances.
- **Style - 5**: The style of the translations was very good, preserving the tone and formality of the original content.

Two-thirds training set - Figure 7.11

- **Accuracy - 3**: The translations were neutral in accuracy, capturing the general meaning but missing some nuances.
- **Fluency - 4**: Translations were good in terms of fluency, mostly natural and grammatically correct, with occasional awkward phrasing.
- **Terminology - 4**: Terminology usage was good, generally appropriate and mostly consistent with domain-specific vocabulary.
- **Locale convention - 3**: The adherence to locale-specific conventions was neutral, with some translations reflecting cultural nuances while others did not.
- **Style - 3**: The style was neutral, adequately reflecting the original content's tone and formality in most cases.

One-third training set - Figure 7.12

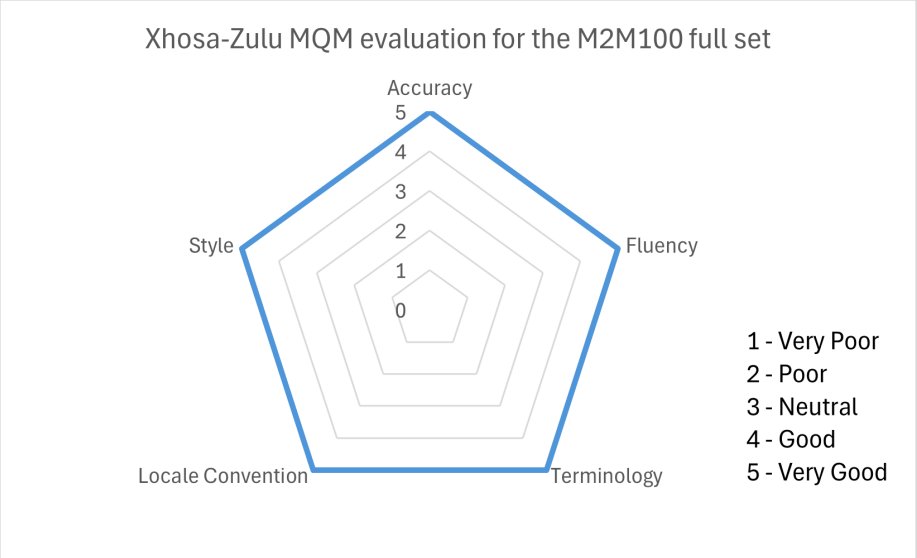

Figure 7.10: MQM evaluation results averages when M2M100 was fine-tuned on the full training set split for Xhosa-Zulu

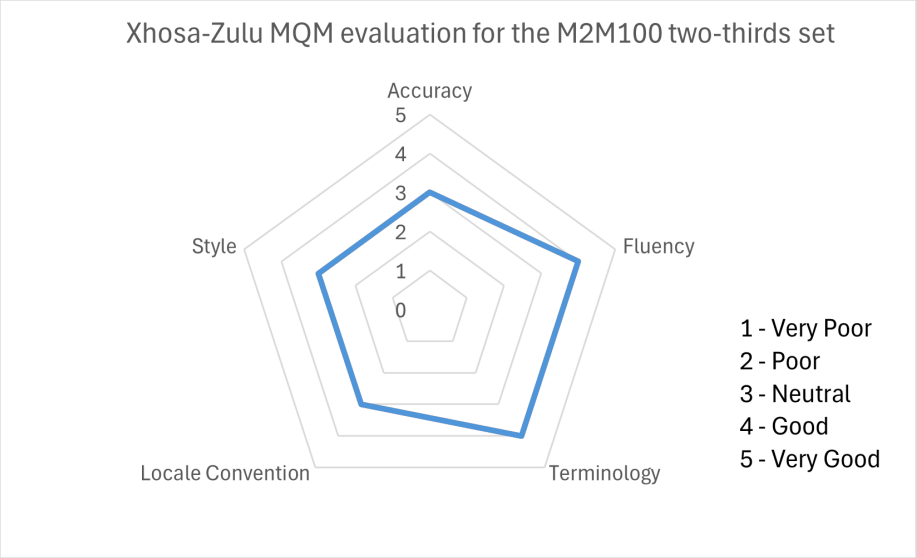

Figure 7.11: MQM evaluation results averages when M2M100 was fine-tuned on the two-thirds training set split for Xhosa-Zulu

- **Accuracy - 2**: The translations were poor in accuracy, often deviating from the intended meaning and missing key nuances.
- **Fluency - 3**: Fluency was neutral, with translations being somewhat natural but containing grammatical errors and unnatural phrasing.
- **Terminology - 3**: Terminology usage was neutral, sometimes appropriate but lacking consistency with domain-specific vocabulary.
- **Locale convention - 3**: The adherence to locale-specific conventions was neutral, with translations occasionally reflecting cultural nuances but often missing them.
- **Style - 2**: The style was poor, frequently failing to preserve the tone and formality of the original content.

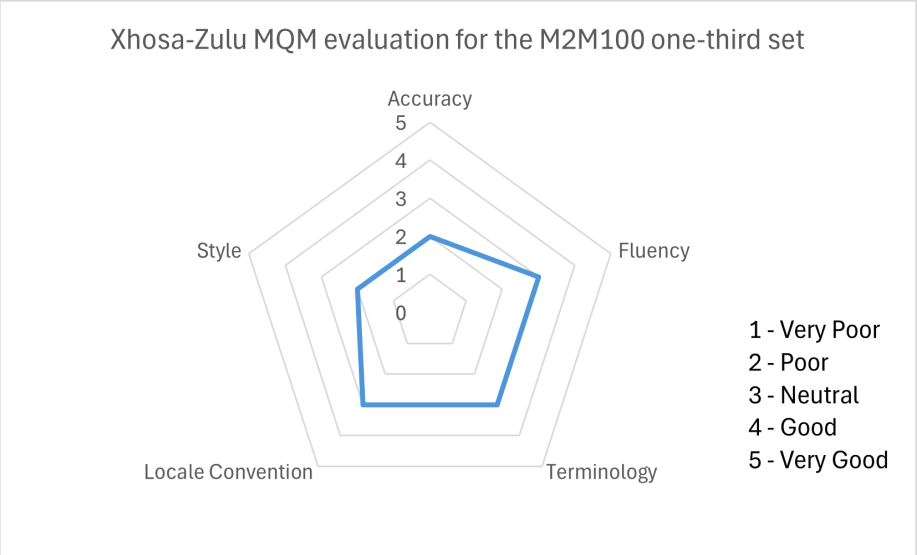

Figure 7.12: MQM evaluation results averages when M2M100 was fine-tuned on the one-third training set split for Xhosa-Zulu

Tswana-Zulu translation evaluation

The translations from Tswana to Zulu were evaluated across different training set sizes: full set, two-thirds set, and one-third set. The metrics assessed were accuracy, fluency, terminology, locale convention, and style. The results are detailed as follows:

Full training set - Figure 7.13

- **Accuracy - 3**: The translations provided a neutral level of accuracy, capturing the general meaning but often missing finer nuances of the original text.
- **Fluency - 3**: Fluency was also neutral, with translations being somewhat natural but exhibiting occasional awkward or unnatural phrasing. We also note translations overlapping to Xhosa when Zulu translation are desired.
- **Terminology - 4**: The use of terminology was good, generally appropriate and consistent with domain-specific vocabulary.
- **Locale convention - 3**: Adherence to locale-specific conventions was neutral, with translations inconsistently capturing currency information.
- **Style - 4**: The style was good, maintaining the tone and formality of the original text in most cases. There were outliers as demonstrated in table 7.8 where one word changed the tone of the sentence completely.

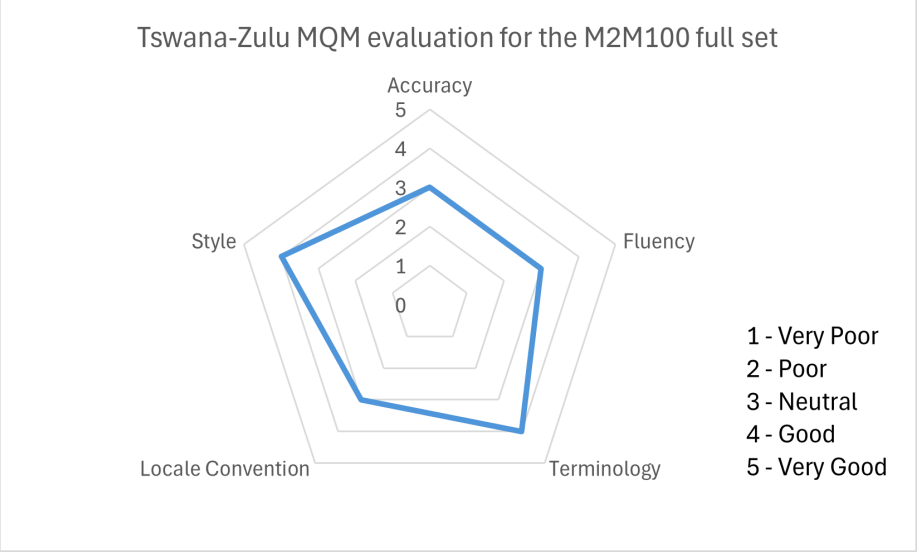

Figure 7.13: MQM evaluation results averages when M2M100 was fine-tuned on the full training set split for Tswana-Zulu

Two-thirds training set - Figure 7.14

- **Accuracy - 5**: The translations achieved a very high level of accuracy, effectively conveying the meaning and subtleties of the original text.
- **Fluency - 5**: Fluency was rated as very good, with translations being natural and grammatically correct.
- **Terminology - 4**: The terminology usage was good, appropriate, and consistent with domain-specific terms, though there is room for improvement where numeric data is concerned.
- **Locale convention - 4**: Adherence to locale-specific conventions was good overall, with the minor failure to perform currency translation correctly.
- **Style - 5**: The style was very good, effectively preserving the original text's tone and formality.

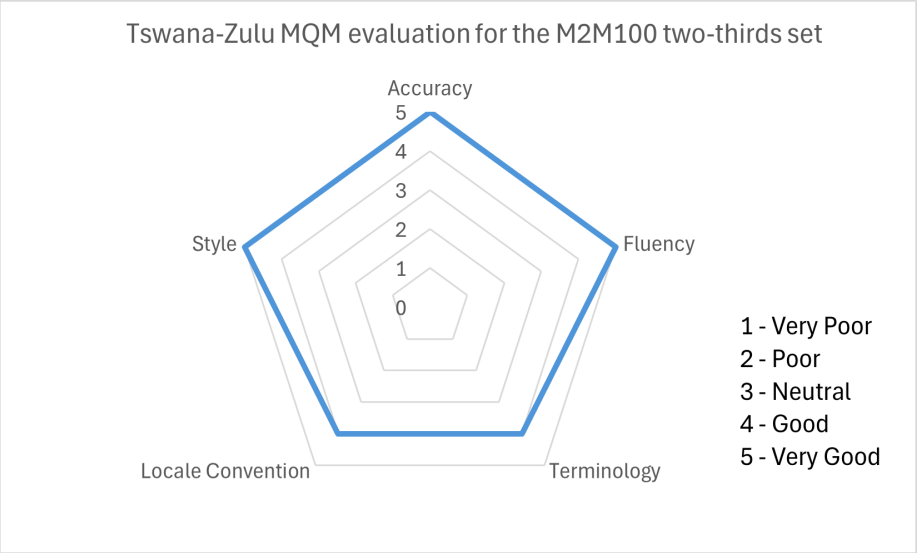

Figure 7.14: MQM evaluation results averages when M2M100 was fine-tuned on the two-thirds training set split for Tswana-Zulu

One-third training set - Figure 7.15

- **Accuracy - 2**: Accuracy was poor, with translations frequently missing key nuances and often not conveying the intended meaning accurately.
- **Fluency - 3**: Fluency was neutral, with translations showing some naturalness but containing several grammatical errors and awkward phrasing.
- **Terminology - 3**: Terminology usage was neutral, sometimes appropriate but lacking consistency and precision with domain-specific vocabulary.
- **Locale convention - 4**: Locale-specific conventions were good, with translations generally reflecting correct currency translation.

- **Style - 3**: The style was neutral, sometimes preserving the original tone and formality but often failing to do so consistently.

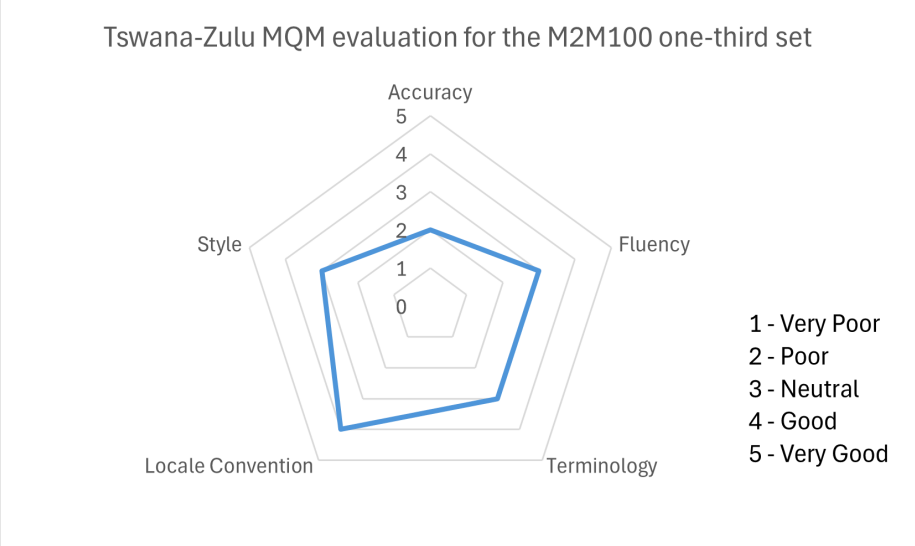

Figure 7.15: MQM evaluation results averages when M2M100 was fine-tuned on the one-third training set split for Tswana-Zulu

ONMT translation quality evaluation

The ONMT translations were assessed across different training set sizes: full set, two-thirds set, and one-third set. The evaluations were performed on key metrics including accuracy, fluency, terminology, locale convention, and style. Regardless of the language pair, the ONMT translations were rated very poorly across all metrics and training set sizes. We do not have separate images for Xhosa-Zulu and Tswana-Zulu translations as the scores we same regardless of language pair. These results complement the translation correctness evaluation results from Tables 7.7 and 7.9

Full training set - Figure 7.16

- **Accuracy - 1**: Translations failed to accurately convey the meaning of the source text, with frequent mistranslations and loss of key information.
- **Fluency - 1**: The translations were highly unnatural and grammatically incorrect, leading to incomprehensible and sometimes senseless outputs.
- **Terminology - 1**: Terminology was consistently incorrect and inconsistent, failing to use appropriate domain-specific vocabulary.
- **Locale Convention - 1**: The translations did not adhere to locale-specific conventions at all.
- **Style - 1**: The translations did not maintain the tone and formality of the original text, resulting in stylistically inappropriate outputs.

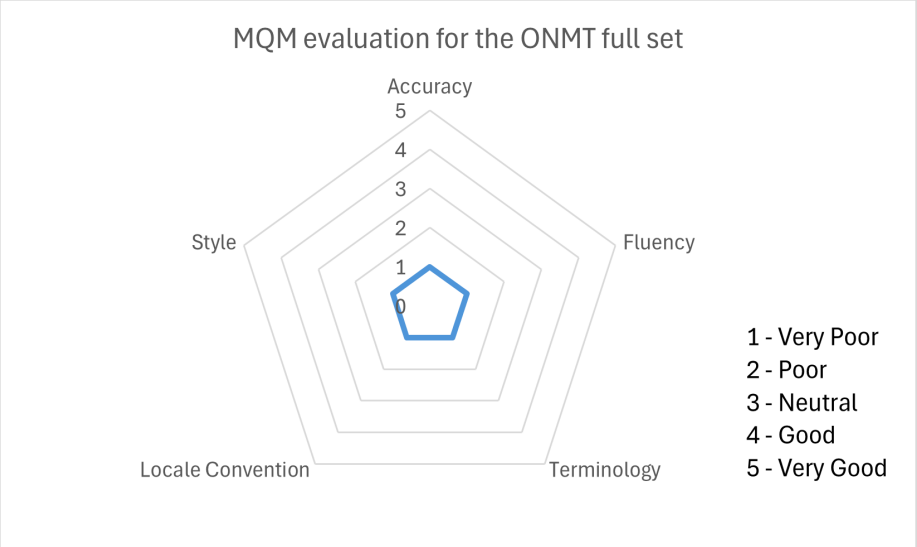

Figure 7.16: MQM evaluation results averages when ONMT was trained on the full training set split for both Xhosa-Zulu and Tswana-Zulu

Two-thirds training set - Figure 7.17

- **Accuracy - 1**: Similar to the full training set, the translations were highly inaccurate, losing the core meaning of the source text.
- **Fluency - 1**: Fluency was very poor, with outputs that were unnatural and grammatically incorrect.
- **Terminology - 1**: The incorrect translations failed to use correct terminology and failed to apply the correct vocabulary.
- **Locale convention - 1**: The translations did not reflect any locale-specific norms. No numeric data was ever translated by the model to demonstrate this
- **Style - 1**: The style was very poor, with translations not matching the original text's tone, formality or structure.

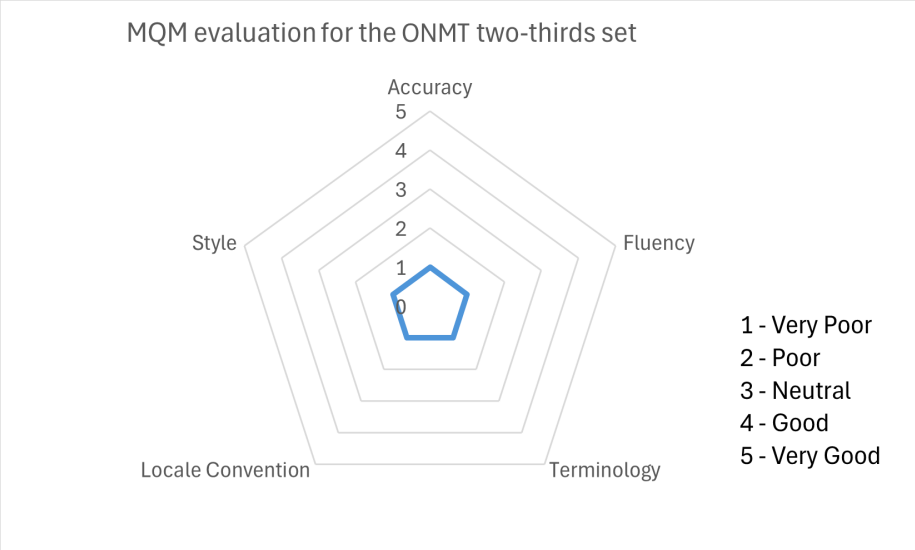

Figure 7.17: MQM evaluation results averages when ONMT was trained on the two-thirds training set split for both Xhosa-Zulu and Tswana-Zulu

One-third training set - Figure 7.18

- **Accuracy - 1**: Translations were extremely inaccurate, often deviating from the meaning of the source text.
- **Fluency - 1**: The translations were unnatural and grammatically flawed, making them difficult to understand.
- **Terminology - 1**: Terminology usage was very poor, lacking consistency and relevance to the subject matter.
- **Locale convention - 1**: Locale-specific conventions were not adhered to. The model failed to translate locale related data such as time and currency from the source sentences as shown in table 7.9.

- **Style - 1**: The style was consistently inappropriate, not preserving the original text's tone or formality.

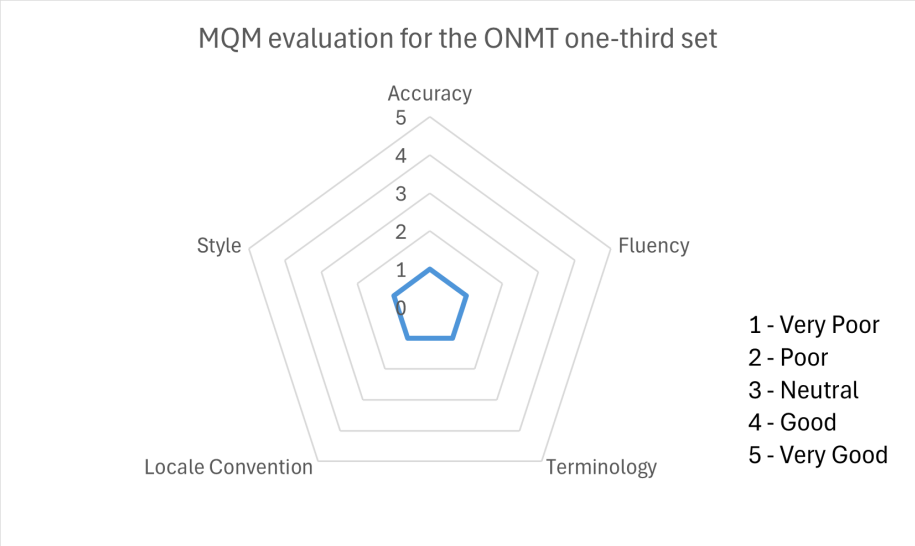

Figure 7.18: MQM evaluation results averages when ONMT was trained on the one-third training set split for both Xhosa-Zulu and Tswana-Zulu

Conclusions from MQM evaluations

MQM evaluation summary

The MQM evaluation results highlight the stark contrast in translation quality between the M2M100 and ONMT models for Xhosa-Zulu and Tswana-Zulu translations. The human evaluators assessed the translations based on five key metrics: accuracy, fluency, terminology, locale convention, and style.

M2M100 model performance

Xhosa-Zulu translation: The M2M100 model demonstrated outstanding performance in translating Xhosa to Zulu. With the full training set, it achieved the highest ratings across all metrics: accuracy, fluency, terminology, locale convention, and style, each receiving a perfect score of 5. This indicates that the translations were not only accurate but also fluent, used correct terminology, adhered to locale-specific norms, and maintained the appropriate style. This is reflected by the correctness of the translations post human evaluation as evident in table 7.6

However, when the training data was reduced to two-thirds, there was a notable decline in performance, particularly in accuracy and locale convention, which dropped to 3. This suggests that while the model can still produce acceptable translations with reduced data, its ability to maintain high standards

across all metrics is compromised. Further reduction to one-third of the training data led to even lower ratings, with accuracy dropping to 2 and style to 2, indicating poor translation quality with insufficient training data.

Tswana-Zulu translation: For Tswana to Zulu translations, the M2M100 model's performance was less impressive when the full training set was used. Accuracy, fluency, and locale convention ratings were rated at 3, reflecting the findings from table 7.8 where the model mistranslated numeric/currency data. The model performed slightly better in terminology and style, each rated at 4. This suggests that while the translations were somewhat correct and stylistically acceptable, they lacked fluency and failed to fully capture the locale-specific nuances.

When training data was reduced to two-thirds, the model's performance improved remarkably in accuracy, fluency, and style, each scoring a perfect 5. This unexpected result indicates that the model might have generalized better with a smaller but still substantial amount of data.
However, further reduction to one-third of the training set resulted in poor ratings in accuracy and terminology, suggesting that a minimal amount of data significantly impairs the model's ability to produce high-quality translations.

ONMT model performance

The ONMT model, in contrast, performed consistently poorly across all metrics and training set sizes, regardless of the language pair. The human evaluators rated it with the lowest possible score of 1 for accuracy, fluency, terminology, locale convention, and style in every scenario. This indicates a significant inability of the ONMT model to generate reliable translations, failing to maintain even basic translation quality standards.

Conclusion

In conclusion, the MQM evaluations reveal that the M2M100 model is significantly superior to the ONMT model in translating both Xhosa to Zulu and Tswana to Zulu, especially with adequate training data. The M2M100 model's performance varies with the amount of training data, excelling with a full set and even improving in certain metrics with a two-thirds set for Tswana-Zulu. The ONMT model's uniformly poor performance across all metrics and data sizes highlights its inadequacy for these translation tasks.

Future work should focus on enhancing the robustness of translation models like M2M100 when handling reduced data sizes and exploring why models like ONMT fail to meet the basic translation requirements. In addition, more research is needed to understand the model behaviors and performance inconsistencies observed with varying amounts of training data.

7.4.3 ESS

The ESS survey aims to evaluate the clarity, detail, helpfulness, trust, and overall satisfaction of the method under consideration. We provide below the average weights obtained when the evaluators evaluated the applicability of each XAI method. The questionnaire also has a few qualitative questions that allow evaluators to provide feedback outside the structured questionnaire setup. We provide detailed feedback from the survey process for each model below. The results are on the overall model, that is the application of the XAI methods on the LLM and on the small-scale model instead of the one-third, two-thirds and full training set split. It is unnecessary to do this fine-grain approach as we are evaluating XAI methods on the models not the models themselves.

M2M100 ESS evaluation

Figure 7.19 shows the human evaluation rating for the XAI methods BLEU scores, MMD, attention heatmaps and attention analysis for evaluation M2M100 translations. Each subsequent bullet point details the overall average ratings in support of Figure 7.19.

- **BLEU scores:**
  - **Clarity:** BLEU scores are a well-known measure of translation quality. Evaluators found them clear and straightforward to interpret.
  - **Detail:** Evaluators were satisfied with the detailed explanations accompanying the BLEU scores.
  - **Helpfulness:** The evaluators found BLEU scores to be a helpful indicator of translation quality, providing some insights into model performance.
  - **Trust:** The evaluators trusted the BLEU scores as a reliable metric for assessing translation accuracy and consistency.
  - **Overall satisfaction:** The overall satisfaction with BLEU scores was high, reflecting preference of application of this metric in evaluating translation quality.
- **Attention heatmaps:**
  - **Clarity:** Evaluators found attention heatmaps reasonably clear.
  - **Detail:** The detail provided by attention heatmaps was appreciated, as they offered insights into how translations were generated.
  - **Helpfulness:** Attention heatmaps were evaluated as helpful in visualizing the model's focus during translation, aiding in understanding translation decisions.
  - **Trust:** Evaluators trusted attention heatmaps to an extent but acknowledged that their interpretation might require more technical expertise.
  - **Overall satisfaction:** the evaluators expressed positive satisfaction to attention heatmaps, highlighting their usefulness as a visualization tool.

- **Attention patterns:**
  - **Clarity:** Attention patterns were clear to evaluators, providing an understandable representation of the model's focus on specific text segments.
  - **Detail:** Evaluators appreciated the detailed insights provided by attention patterns, which helped explain the model's behavior during translation.
  - **Helpfulness:** The evaluators found attention patterns particularly helpful in understanding the nuances of the translation process.
  - **Trust:** Evaluators trusted attention patterns as a valuable tool for analyzing the consistency and accuracy of the translations.
  - **Overall satisfaction:** Overall satisfaction with attention patterns was very high, indicating their usefulness in translation evaluation.
- **MMD:**
  - **Clarity:** MMD was clear to evaluators, effectively measuring the divergence between translations and reference texts.
  - **Detail:** The evaluators valued the detailed analysis provided by MMD, which helped their understanding of translation quality.
  - **Helpfulness:** MMD was found to be very helpful in assessing translation quality.
  - **Trust:** Evaluators trusted MMD as a reliable metric for evaluating the alignment between translated and reference texts.
  - **Overall satisfaction:** Overall satisfaction with MMD was very high, reflecting its effectiveness as an evaluation of translations.

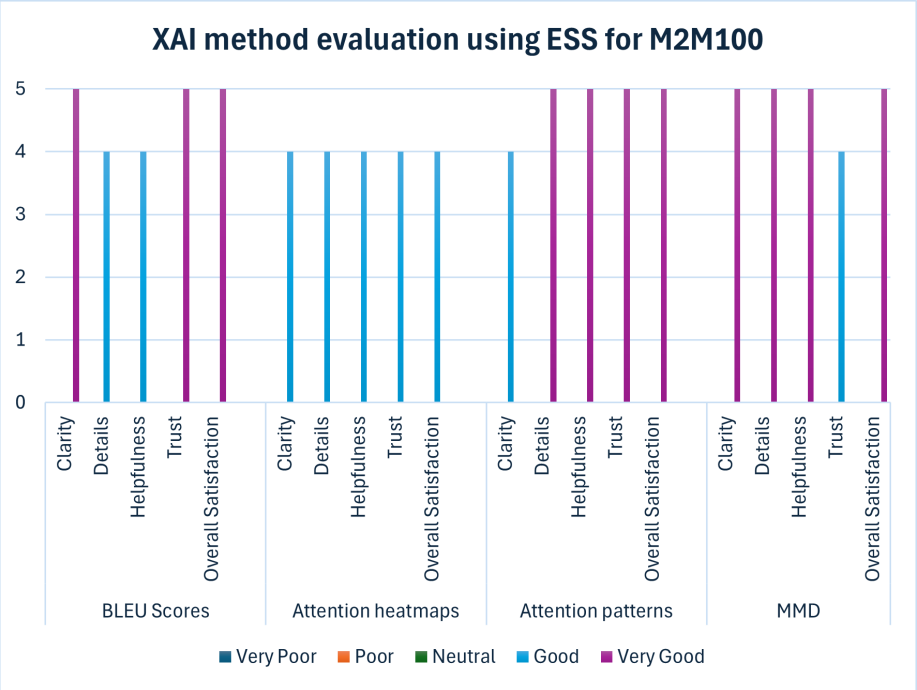

Figure 7.19: Average ratings from the human evaluators on the satisfaction with the relevancy and applicability of the XAI methods across the M2M100 models. Attention pattern analysis, MMD and BLEU scores were considered to have provided satisfactory explanations for the inner workings of the model while attention heatmaps were found least satisfactory

ONMT ESS evaluation

Figure 7.20 shows the human evaluation rating for the XAI methods BLEU scores, MMD, attention heatmaps and attention analysis for evaluation ONMT translations. Each subsequent bullet point details the overall average ratings in support of Figure 7.20.

- **BLEU scores**
  - **Clarity:** Evaluators found BLEU scores to be unclear and challenging to interpret, giving a rating of 1 given how incorrect translations were across all model variations.
  - **Detail:** The evaluators gave a rating of 2 as the scores did not offer adequate detail on translation quality.
  - **Helpfulness:** Evaluators rated the helpfulness of BLEU scores as poor, with a score of 1, indicating low usefulness in indicating translation quality.
  - **Trust:** BLEU scores received a low trust rating of 1, reflecting a lack of confidence in their accuracy.
  - **Overall satisfaction:** Overall satisfaction with BLEU scores was rated as very low, with a score of 1, indicating significant dissatisfaction for translation evaluation for ONMT translations for Xhosa-Zulu and Tswana-Zulu.
- **Attention heatmaps**
  - **Clarity:** Attention heatmaps were found to be clear and comprehensible by evaluators, earning a clarity score of 4.
  - **Detail:** Evaluators rated the level of detail in attention heatmaps with a score of 4, indicating satisfactory granularity.
  - **Helpfulness:** The helpfulness of attention heatmaps in evaluating model performance was rated positively, with a score of 4.
  - **Trust:** Attention heatmaps received a trust score of 4, reflecting some confidence in their ability to evaluate model and translation accuracy.
  - **Overall satisfaction:** Overall satisfaction with attention heatmaps was good, with a score of 4, showing general approval.
- **Attention patterns**
  - **Clarity:** Attention patterns were rated very clear, with a score of 5 for clarity, indicating strong comprehensibility.
  - **Detail:** The level of detail in attention patterns was rated with a score of 4, showing satisfaction with the provided insights.
  - **Helpfulness:** Evaluators found attention patterns to be very helpful for assessing translation quality, with a score of 5.
  - **Trust:** A high trust score of 5 was given to attention patterns, indicating strong confidence in their reliability.

  - **Overall satisfaction:** Overall satisfaction with attention patterns was very high, with a score of 5, reflecting broad approval.

- **MMD**
  - **Clarity:** Evaluators rated the clarity of MMD as very good, with a perfect score of 5, indicating clear and understandable results.
  - **Detail:** The level of detail provided by MMD was highly appreciated, earning a score of 5.
  - **Helpfulness:** MMD was seen as very helpful in evaluating translation quality, with a score of 5.
  - **Trust:** A trust score of 4 was given to MMD, reflecting good confidence in its accuracy.
  - **Overall satisfaction:** Overall satisfaction with MMD was high, with a score of 5, indicating strong approval from the evaluators as an evaluation metric for model and translation accuracy.

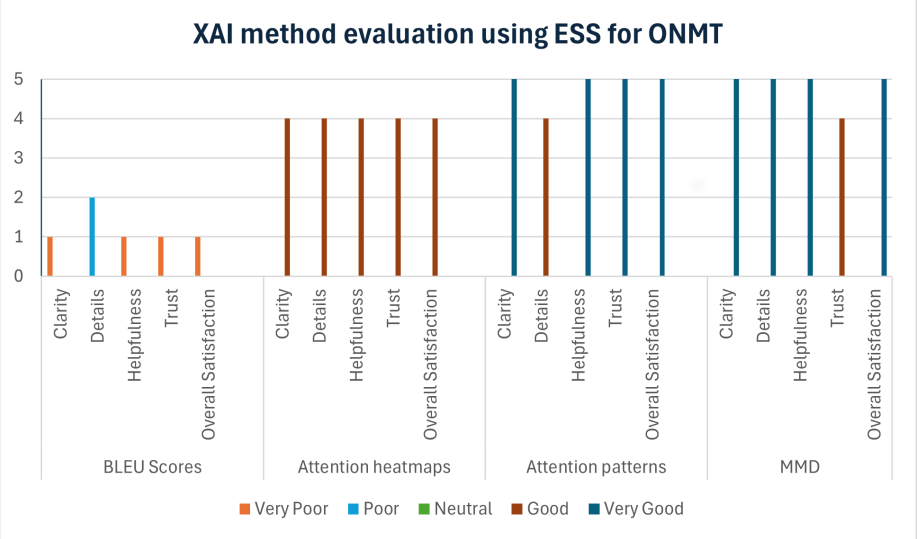

Figure 7.20: Average ratings from the human evaluators on the satisfaction with the relevancy and applicability of the XAI methods across the ONMT models. Attention pattern analysis and MMD scores were considered to have provided satisfactory explanations for explaining the model's translations, aligning with the poor translations evaluations from Section 7.4. Attention heatmaps were somewhat helpful while BLEU scores were least helpful. This aligns with Section 7.1 where ONMT had high BLEU scores and poor translation quality

Conclusions from ESS evaluations

As indicated in the qualitative metrics for both the evaluation on M2M100 and ONMT in Appendix B.3, all the evaluators have knowledge of machine translation so they have some prior knowledge of the methods applied for evaluating translation quality in this research. From their responses, we see a greater inclination towards MMD as an ideal method for evaluating translation correctness when references are available regardless of model size with high ratings for clarity, helpfulness, detail, trust and overall satisfaction.
Attention pattern analysis was the next recommendation for explaining model behaviour. It is important to keep in mind that this method is tightly coupled to the availability of attention heatmaps or attention visualizations of some sort. The evaluators have some confidence in the BLEU scores, especially when focus is on M2M100, however loose confidence in the metric when ONMT is under evaluation. They recommend exploration of various evaluation metrics such as COMET for a more accurate representation as the high BLEU scores fail to represent translation failures.

In conclusion, the feedback indicates a clear hierarchy of preferred evaluation methods, with MMD and attention pattern analysis being highly valued for their ability to provide detailed and trustworthy assessments of translation quality independent of the model scale under evaluation. The insights also highlight the limitations of traditional metrics like BLEU. BLEU might be relevant for some models and languages, but when certain low resources are under investigation, it might be ideal to explore other quantitative metrics. This survey highlights the need for a diverse set of evaluation tools to accurately gauge the effectiveness of machine translation models across different contexts and datasets.

Chapter 8: Discussion

This study compared the performance and interpretability of transformer-based NMT models—specifically, the small-scale OpenNMT and the large-scale LLM M2M100—on low-resource language pairs Xhosa-Zulu and Tswana-Zulu. Our findings shed light on translation quality, the impact of dataset size, and the efficacy of post-model XAI techniques. Given the chapters preceding this, we take a look at the hypotheses outlined in Section 1.3 and discuss other points highlighting the relevance of this research in regards to the current XAI research landscape.

8.1 Key findings

Our key findings align with our research hypotheses as follows:

8.1.1 Translation quality and model scale (H1):

The M2M100 model outperformed the small-scale OpenNMT model, achieving BLEU scores of 29.59 for Xhosa-Zulu translations—substantially higher than the 8.5 reported by Elmadani et al. (2022) [17]. This confirms H1, demonstrating that large-scale LLMs deliver superior translation quality due to their extensive training on multilingual corpora. This also encourages towards the development and use of LLMs for low resource language translation.

8.1.2 Impact of dataset size (H2):

The performance of the small-scale model improved as training dataset size increased, though it was insufficient to produce good quality translation as evident from the human evaluations (refer to Section 7). Furthermore, the LLM produced better quality translations when the full training set was used in comparison to when the one-third or two-third splits were used (refer to Tables 7.6 and 7.8) This supports H2 and highlights the challenges of achieving high-quality translations when working with limited data in low-resource settings.

8.1.3 Insights from post-model XAI methods (H3):

Automated metrics like BLEU scores captured overall performance trends but failed to identify nuanced translation errors (e.g., fluency and semantic mismatches). The human evaluation methods MQM and ESS proved more effective in identifying subtle errors missed by BLEU. They proved complimentary to the attention pattern analysis, heatmaps and MMD scores in providing comprehensive insight to evaluating translation correctness, thus validating H3.

8.2 Relevance in existing research landscape

While there are known XAI methods like LIME[73] and SHAP[74], the application of attention pattern analysis, BLEU scores, MMD scores, and human evaluation methods as XAI methods of choice for the research was done after careful consideration as per points below:

8.2.1 Task-specific insights:

Attention analysis is tailored for models like transformers, where attention mechanisms are part of the model's explainability. BLEU and MMD directly measure performance and distribution alignment, making them better suited for evaluating tasks like translation or generative outputs.

8.2.2 Global vs. local explanations:

LIME/SHAP provide explanations at the local level (single predictions). BLEU, MMD, and attention analysis provide global performance measures that can evaluate a model's behavior across datasets.

8.2.3 Human evaluation:

Unlike LIME/SHAP, which are algorithmic, human evaluation introduces human judgment and contextual understanding to assess the quality of outputs.

The following points highlight the specific motivation for each XAI method over LIME/SHAP for this particular study.

- **BLEU Scores**
  - BLEU measures n-gram overlap between machine-generated and reference translations, providing a numerical quality score for translation outputs.
  - LIME/SHAP explain individual predictions, while BLEU evaluates overall model performance, particularly useful for evaluating MT.
- **Attention Pattern Analysis**
  - It visualizes where the model "focuses" during predictions, providing insights into its decision-making. The attention weights can be analyzed statistically or visually to quantify their alignment with human expectations.
  - LIME and SHAP provide feature importance but are agnostic to attention mechanisms. Attention analysis, on the other hand, is model-intrinsic and directly interpretable for transformer-based models.
- **MMD Scores**
  - MMD compares distributions (e.g., output embeddings) to assess model behavior.

  - LIME/SHAP focus on local explanations (single outputs), whereas MMD provides global insights into the behavior and consistency of model predictions.

- **Human Evaluation**
  - Human evaluation assesses how well outputs align with human expectations. Results are often aggregated into numerical scores as with MQM and ESS.
  - Human evaluation directly measures the real-world utility and coherence of model outputs, while LIME/SHAP give technical feature importance values.

While LIME and SHAP are widely used for model-agnostic explanations, methods like attention analysis, BLEU, MMD, and human evaluation provide domain-specific, quantitative insights that align better with MT tasks by evaluating both the global performance and alignment of outputs with human expectations, thus offering another perspective to XAI.

Our results align with prior studies that emphasize the limitations of BLEU scores in low-resource language translation [56, 57]. Our use of human evaluation methods highlighted translation errors that automated metrics overlooked.

This study further reinforces the importance of dataset size, echoing research by Elmadani et al. (2022) [17], who identified significant performance challenges for low-resource Southern African language MT. However, our results go a step further by providing new BLEU benchmarks for Tswana-Zulu, addressing a gap in the literature.

8.3 Challenges and limitations of the small-scale model

The performance of the small-scale model (OpenNMT) in this study revealed several challenges and limitations that impacted its ability to deliver high-quality translations for the low-resource language pairs Xhosa-Zulu and Tswana-Zulu. These challenges highlight the broader difficulties faced by small-scale transformer models in handling low-resource language tasks, which might be transferrable to other low-resource language pairs:

8.3.1 Data sparsity and generalization issues

The small-scale model struggled to generalize well when trained on smaller datasets. Unlike LLMs, which leverage extensive pre-training on multilingual corpora, the small-scale model exhibited poor performance when the training data was limited (i.e. one-third and two-thirds dataset splits). This resulted in reduced translation accuracy, particularly when handling complex sentence structures and uncommon linguistic patterns such as ?? in place of actual words (refer to Table 7.7).

8.3.2 Difficulty with longer sentences

The small-scale model demonstrated significant challenges in accurately translating longer sentences (refer to MTs in Tables 7.7 and 7.9. The model's limited capacity to handle long-range dependencies within the input text caused errors in word alignment and sentence fluency. This was particularly evident in cases where the source sentences included multiple clauses or nuanced grammatical constructs.

8.3.3 Alignment and attention issues

Analysis of the attention patterns in Section 7.2 revealed that the small-scale model frequently misaligned words between the source and target languages. These alignment errors often led to missing or incorrect translations, especially for low-frequency terms or domain-specific vocabulary, which are common in low-resource language pairs. This limitation highlights the model's inability to adequately learn relationships between source and target tokens due to insufficient training data.

8.3.4 Linguistic complexity

Low-resource languages like Xhosa, Zulu and Tswana exhibit unique linguistic features, such as agglutination and rich morphology, which the small-scale model struggled to capture. This limitation resulted in grammatical errors and inconsistencies in translations, further reducing translation quality compared to the LLM.

8.3.5 Sensitivity to dataset size

The small-scale model's performance was highly sensitive to the size of the training dataset. While increasing the dataset size improved its translation quality to some extent, it still fell short of the performance achieved by the LLM. This sensitivity highlights the limited capacity of small-scale models to handle low-resource tasks effectively without significant data augmentation or transfer learning techniques.

Chapter 9: Conclusion and future work

With so many black box models in use in the real word, for example OpenAI's ChatGPT, it is imperative we make efforts towards the explainability of the models inner workings for auditability purposes. Given the experiments carried out during this research in efforts to achieve this, we draw the conclusions outlined below in detail to addres our research objectives from Section 1.2.

9.1 Summary on XAI methods applied

9.1.1 BLEU scores

While both the small-scale model and the LLM exhibit similar BLEU scores, their translations diverge significantly in terms of quality. Despite achieving comparable or even higher BLEU scores, the small-scale model's translations often lack semantic fidelity and naturalness, indicating potential issues with overfitting or inadequate capturing of linguistic nuances. This highlights the limitations of solely relying on quantitative metrics like BLEU scores for evaluating translation models. Therefore, future research should incorporate qualitative analyses, including human evaluation and linguistic diagnostics, to provide a more holistic assessment of translation quality. By doing so, we can better understand the strengths and weaknesses of different models and refine their capabilities for practical applications in real-world scenarios.

9.1.2 Correlation between attention patterns and translation quality

Models of different scales tend to capture attention differently, affecting translation quality. While both the small-scale model and the LLM exhibit similar BLEU scores, their translations diverge significantly in terms of quality. The small-scale model's translations often lack semantic fidelity and naturalness, despite achieving comparable or even higher BLEU scores. This indicates potential issues with overfitting or inadequate capturing of linguistic nuances. The even attention patterns of the LLM, with few outliers, suggest better attention distributions that lean more towards target sentences, while the uneven and sometimes diagonal attention patterns of the small-scale model reveal its struggle to maintain consistent translation quality. This underscores the importance of visualizing attention patterns to explain and understand model performance beyond quantitative metrics like BLEU scores.

9.1.3 Accuracy of post-model interpretability methods in evaluating low-resource language translation quality

Post-model interpretability methods such as attention pattern analysis and distance measurement metrics like MMD can provide valuable insights into translation quality. While BLEU scores offer a quantitative evaluation, they do not fully capture the nuances required for explaining model behavior in low-resource language translation tasks. Attention heatmaps, pattern analysis, and MMD distance measurements help identify translation errors and understand how models handle different linguistic structures and contextual dependencies. These methods, particularly when combined with human evaluations, offer a more comprehensive assessment of translation quality. MMD distance analysis, in particular, has been highlighted for its critical role in measuring alignment between model-generated translations and target sentences, thus reflecting its effectiveness in identifying discrepancies and ensuring translation accuracy.

9.1.4 Insights from comparative analysis and model-agnostic techniques

(a) Comparing transformer NMT models of different scales

Comparing transformer NMT models of different scales reveals significant differences in how they handle translation tasks. The LLM's even attention patterns indicate a better overall handling of linguistic structures, leading to translations that are closer to the target sentences. In contrast, the small-scale model's uneven and sometimes diagonal attention patterns suggest difficulties in maintaining consistent translation quality. This comparison highlights the need for larger models or more sophisticated architectures to better capture the complexities of low-resource languages.

(b) Applying model-agnostic post-hoc interpretability techniques to NMT models

Model-agnostic post-hoc interpretability techniques, such as attention pattern analysis and MMD distance measurements, are crucial for evaluating and improving NMT models. These techniques provide detailed insights into model behavior, allowing researchers to identify and address specific issues. For instance, attention pattern analysis can reveal which parts of the sentence the model focuses on, while MMD distance measurements can quantify the similarity between model outputs and reference translations. These insights are invaluable for refining model performance and ensuring that translations meet the required quality standards.

(c) Training NMT models on different training set sizes for the same translation task

Training NMT models on different training set sizes for the same translation task shows varying impacts on translation quality. Larger training sets generally

lead to better translation accuracy, fluency, and overall quality, as evidenced by higher human evaluation scores. However, even with smaller training sets, some models can achieve high BLEU scores, though this does not necessarily correlate with high-quality translations. This observation underscores the importance of using diverse evaluation methods to obtain a non-biased and comprehensive assessment of model performance.

Conclusion

In conclusion, we recommend that research on the interpretability of low-resource languages incorporate qualitative analyses, including human evaluation and linguistic diagnostics, to provide a more holistic assessment of translation quality. By doing so, we can better understand the strengths and weaknesses of different models and refine their capabilities for practical applications in real-world scenarios. We hope that this work encourages the need to create large-scale quality datasets for low-resource languages, so we have more training data to produce better performing models.
We note that it is ideal to use a multilingual LLM like M2M100 for the translation of Xhosa-Zulu and Tswana to Zulu, with more data needed to produce correct Xhosa-Zulu translations while less data is needed for correct Tswana to Zulu translations. small-scale models like ONMT must be used with caution for creating translation applications in the real world should resources not be available to use larger models, especially where there are no native human evaluators of the languages to evaluate translation correctness. Human feedback is important so we do not rely on one or more seemingly correct sentence to assume model performs well. We recommend rather using small-scale models as surrogate to explain the inner workings of large-scale models instead as they can be fine-tuned to emulate LLMs.
Models of different scale tend to capture attention differently. We recommend visualizing attention for research where performance benchmarks are weighted in order to help explain model performance.
Distance measurement metrics like MMD can be used to evaluate translation correctness if references exist. To some extent, they complement human evaluations well. When evaluating translation models on low-resource language pairs, we recommend using various evaluation methods that evaluate results from the previous stage in order to produce non-biased evaluations

9.2 Future work

9.2.1 Investigating model behavior with misaligned sentence pairs

One of the intriguing findings of this research is the ability of NMT models to generate correct translations even when trained on datasets containing misaligned sentence pairs. This observation raises several questions about the underlying mechanisms that enable these models to compensate for or overlook misalignments in the training data. Future research should focus on the following aspects:

- **Characterizing misalignments:** A comprehensive analysis of the nature and frequency of misalignments within training datasets is necessary. This includes identifying common patterns of misalignment and understanding how these errors propagate through the training process.
- **Model robustness:** Investigating the robustness of various NMT models to training data misalignment. This involves testing different model architectures and training paradigms to assess their sensitivity to and recovery from misaligned input.
- **Learning dynamics:** Understanding how models learn from misaligned pairs by analyzing the learning dynamics at different stages of training. This can provide insights into whether models are discarding noisy data or learning to generalize despite inconsistencies.
- **Synthetic data experiments:** Conducting controlled experiments with synthetic datasets where the degree and type of misalignment can be precisely controlled to study the impact on model performance.

9.2.2 Developing an XAI framework for low-resource languages

The evaluation of machine translation systems, especially for low-resource languages, presents significant challenges due to the scarcity of human evaluators and the diverse linguistic properties across languages. To address these issues, future research should aim to develop a framework for XAI that facilitates the interpretability and evaluation of translation models for low-resource languages. The key components of this framework include:

- **Cost-effective evaluation:** Recognizing the high cost and time constraints associated with hiring human evaluators, the framework should leverage automated and semi-automated methods to assess translation quality efficiently. This includes integrating various metrics and tools to minimize the reliance on human evaluation.
- **Flexible evaluation methods:** Although a one-size-fits-all evaluation method may not be feasible due to the unique characteristics of different languages, the framework should offer a flexible set of steps and methods. This would enable researchers to tailor their evaluation approach to specific languages while adhering to a consistent overarching methodology.
- **Incorporating different analysis methods:** The framework should advocate for the use of pre-model, in-model and post-model interpretability methods, such as MMD and attention pattern analysis. These methods can provide valuable insights into model behavior and translation correctness, particularly for languages with limited resources.
- **Scalable and reproducible:** Ensuring that the framework is scalable to accommodate large datasets and complex models, and that it promotes reproducibility across different research settings and language pairs.

By addressing these areas, future research can advance the understanding of NMT model behavior and improve the interpretability and evaluation of translations, particularly for low-resource languages. This will ultimately contribute to more robust and reliable machine translation systems.

Acknowledgments

We would like to express our sincere gratitude to the human evaluators who contributed their time and expertise to assess the model translations and complete both MQM and ESS evaluations for this study. Their invaluable insights and meticulous evaluations were essential in validating the quality and accuracy of the models' outputs. Their dedication to this research is deeply appreciated.

Bibliography

[1] D. Adelani, M. M. I. Alam, A. Anastasopoulos, A. Bhagia, M. R. Costa-jussà, J. Dodge, F. Faisal, C. Federmann, N. Fedorova, F. Guzmán, S. Koshelev, J. Maillard, V. Marivate, J. Mbuya, A. Mourachko, S. Saleem, H. Schwenk, and G. Wenzek, "Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, (Abu Dhabi, United Arab Emirates (Hybrid)), pp. 773–800, Association for Computational Linguistics, Dec. 2022.

[2] R. Lastrucci, I. Dzingirai, J. Rajab, A. Madodonga, M. Shingange, D. Njini, and V. Marivate, "Preparing the vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora," in *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, (Dubrovnik, Croatia), pp. 18–25, Association for Computational Linguistics, May 2023.

[3] V. Marivate, D. Njini, A. Madodonga, R. Lastrucci, and J. Dzingirai, Isheanesu Rajab, "The vuk'uzenzele south african multilingual corpus," Feb. 2023.

[4] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," *J. Mach. Learn. Res.*, vol. 22, jan 2021.

[5] A. Korinek, "Language models and cognitive automation for economic research," Working Paper 30957, National Bureau of Economic Research, February 2023.

[6] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, "Risks and benefits of large language models for the environment," *Environmental Science & Technology*, vol. 57, no. 9, pp. 3464–3466, 2023. PMID: 36821477.

[7] M. Ashoori and J. D. Weisz, "In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes," *arXiv*, 2019.

[8] J. Ebrahimi, D. Lowd, and D. Dou, "On adversarial examples for character-level neural machine translation," in *Proceedings of the 27th International Conference on Computational Linguistics* (E. M. Bender, L. Derczynski, and P. Isabelle, eds.), (Santa Fe, New Mexico, USA), pp. 653–663, Association for Computational Linguistics, Aug. 2018.

[9] L. N. Vieira, M. O'Hagan, and C. O'Sullivan, "Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases," *Information, Communication & Society*, vol. 24, no. 11, pp. 1515–1532, 2021.

[10] A. Way, "Emerging use-cases for machine translation," in *Proceedings of Translating and the Computer 35*, (London, UK), Aslib, Nov. 28-29 2013.

[11] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. T. Martins, "Hallucinations in large multilingual translation models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1500–1517, 2023.

[12] R. Bawden and F. Yvon, "Investigating the translation performance of a large multilingual language model: the case of BLOOM," in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz, eds.), (Tampere, Finland), pp. 157–170, European Association for Machine Translation, June 2023.

[13] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, 2019.

[14] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, pp. 1 – 85, 2022.

[15] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[16] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019.

[17] K. Elmadani, F. Meyer, and J. Buys, "University of cape town's WMT22 system: Multilingual machine translation for Southern African languages," in *Proceedings of the Seventh Conference on Machine Translation (WMT)* (P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, eds.), (Abu Dhabi, United Arab Emirates (Hybrid)), pp. 1039–1048, Association for Computational Linguistics, Dec. 2022.

[18] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019.

[19] B. Kim, R. Khanna, and O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (Red Hook, NY, USA), p. 2288–2296, Curran Associates Inc., 2016.

[20] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[21] C. Molnar, *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Independently published (February 28, 2022), 2023. `https://christophm.github.io/interpretable-ml-book/`(visited 2024-09-20).

[22] C. Rudin, "Please stop explaining black box models for high stakes decisions," *Stat*, vol. 1050, p. 26, 2018.

[23] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[24] H. Schwenk, A. Rousseau, and M. Attik, "Large, pruned or continuous space language models on a gpu for statistical machine translation," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pp. 11–19, 2012.

[25] G. Neubig, "Neural machine translation and sequence-to-sequence models: A tutorial," *arXiv preprint arXiv:1703.01619*, 2017.

[26] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," *arXiv preprint arXiv:1703.08581*, 2017.

[27] P. Singh, "A simple introduction to sequence to sequence models." `https://www.analyticsvidhya.com/blog/2020/08/a-simple-introduction-to-sequence-to-sequence-models/`. Accessed: 2024-07-17.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[29] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[30] S. Kudugunta, A. Bapna, I. Caswell, and O. Firat, "Investigating multilingual NMT representations at scale," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 1565–1575, Association for Computational Linguistics, Nov. 2019.

[31] T. Purason and A. Tättar, "Multilingual neural machine translation with the right amount of sharing," in *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (H. Moniz,

L. Macken, A. Rufener, L. Barrault, M. R. Costa-jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy, and M. Fonteyne, eds.), (Ghent, Belgium), pp. 91–100, European Association for Machine Translation, June 2022.

[32] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 4996–5001, Association for Computational Linguistics, July 2019.

[33] A. Abbas, "Unraveling the threads: A thorough examination of attention mechanisms in deep learning models," *ResearchGate*, 02 2024.

[34] S. Kardakis, I. Perikos, F. Grivokostopoulou, and I. Hatzilygeroudis, "Examining attention mechanisms in deep learning models for sentiment analysis," *Applied Sciences*, vol. 11, p. 3883, 2021.

[35] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Multilingual machine translation with large language models: Empirical results and analysis," in *Findings of the Association for Computational Linguistics: NAACL 2024* (K. Duh, H. Gomez, and S. Bethard, eds.), (Mexico City, Mexico), pp. 2765–2781, Association for Computational Linguistics, June 2024.

[36] M. Bu, S. Gu, and Y. Feng, "Improving multilingual neural machine translation by utilizing semantic and linguistic features," in *Findings of the Association for Computational Linguistics ACL 2024* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand and virtual meeting), pp. 10410–10423, Association for Computational Linguistics, Aug. 2024.

[37] T. Rama, L. Beinborn, and S. Eger, "Probing multilingual BERT for genetic and typological signals," in *Proceedings of the 28th International Conference on Computational Linguistics* (D. Scott, N. Bel, and C. Zong, eds.), (Barcelona, Spain (Online)), pp. 1214–1228, International Committee on Computational Linguistics, Dec. 2020.

[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2016.

[39] R. Diandaru, L. Susanto, Z. Tang, A. Purwarianti, and D. Wijaya, "Could we have had better multilingual llms if english was not the central language?," in *TDLE*, 2024.

[40] Z. Li, M. Karimi, H. I. Daume, and M. Hasegawa-Johnson, "Interpreting neural machine translation through visualization," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3479–3489, 2021.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[42] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, jan 2020.

[44] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency, "Towards debiasing sentence representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 5502–5515, Association for Computational Linguistics, July 2020.

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, (USA), p. 311–318, Association for Computational Linguistics, 2002.

[46] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers* (O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, eds.), (Brussels, Belgium), pp. 186–191, Association for Computational Linguistics, Oct. 2018.

[47] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation* (T. Luong, A. Birch, G. Neubig, and A. Finch, eds.), (Vancouver), pp. 28–39, Association for Computational Linguistics, Aug. 2017.

[48] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of Bleu in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics* (D. McCarthy and S. Wintner, eds.), (Trento, Italy), pp. 249–256, Association for Computational Linguistics, Apr. 2006.

[49] E. Reiter, "A structured review of the validity of BLEU," *Computational Linguistics*, vol. 44, pp. 393–401, Sept. 2018.

[50] O. Bojar, Y. Graham, and A. Kamran, "Results of the WMT17 metrics shared task," in *Proceedings of the Second Conference on Machine Translation* (O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, and J. Kreutzer, eds.), (Copenhagen, Denmark), pp. 489–513, Association for Computational Linguistics, Sept. 2017.

[51] Y. Graham, T. Baldwin, and N. Mathur, "Accurate evaluation of segment-level machine translation metrics," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (R. Mihalcea, J. Chai, and A. Sarkar, eds.), (Denver, Colorado), pp. 1183–1191, Association for Computational Linguistics, May–June 2015.

[52] N. Mathur, J. Wei, M. Freitag, Q. Ma, and O. Bojar, "Results of the WMT20 metrics shared task," in *Proceedings of the Fifth Conference on Machine Translation* (L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, and M. Negri, eds.), (Online), pp. 688–725, Association for Computational Linguistics, Nov. 2020.

[53] J. Daems, O. de clercq, and L. Macken, "Translationese and post-editese: How comparable is comparable quality?," *Linguistica Antverpiensia*, vol. 16, 01 2017.

[54] Q. Ma, O. Bojar, and Y. Graham, "Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, eds.), (Belgium, Brussels), pp. 671–688, Association for Computational Linguistics, Oct. 2018.

[55] L. Barrault, O. Bojar, M. R. Costa-Jussa, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, *et al.*, "Findings of the 2019 conference on machine translation (wmt19)," in *ACL 2019 Fourth Conference on Machine Translation (WMT19),*, ACL, 2019.

[56] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey, "Experts, errors, and context: A large-scale study of human evaluation for machine translation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1460–1474, 2021.

[57] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, "Findings of the 2016 conference on machine translation (wmt16)," in *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 131–198, Association for Computational Linguistics, Aug. 2016. First Conference on Machine Translation, WMT16 ; Conference date: 11-08-2016 Through 12-08-2016.

[58] T. Daybelge and I. Cicekli, "A ranking method for example based machine translation results by learning from user feedback," *Applied Intelligence*, vol. 35, pp. 296–321, 2011.

[59] A. Lommel, A. Burchardt, and H. Uszkoreit, "Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics," *Tradumàtica: tecnologies de la traducció*, vol. 0, pp. 455–463, 12 2014.

[60] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance," *Frontiers in Computer Science*, vol. 5, p. 1096257, 2023.

[61] J. M. Schraagen, P. Elsasser, H. Fricke, M. Hof, and F. Ragalmuto, "Trusting the x in xai: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 64, pp. 339–343, SAGE Publications Sage CA: Los Angeles, CA, 2020.

[62] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations* (M. Bansal and H. Ji, eds.), (Vancouver, Canada), pp. 67–72, Association for Computational Linguistics, July 2017.

[63] M. Artetxe and H. Schwenk, "Margin-based parallel corpus mining with multilingual sentence embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3197–3203, Association for Computational Linguistics, July 2019.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[65] A. M. Rush, "The annotated transformer." url: http://nlp.seas.harvard.edu/2018/04/03/attention.html, Accessed: 2024-09-03.

[66] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (M. R. Costa-jussà and E. Alfonseca, eds.), (Florence, Italy), pp. 37–42, Association for Computational Linguistics, July 2019.

[67] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (T. Linzen, G. Chrupała, Y. Belinkov, and D. Hupkes, eds.), (Florence, Italy), pp. 276–286, Association for Computational Linguistics, Aug. 2019.

[68] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2007.

[69] A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu, "Optimal kernel choice for large-scale two-sample tests," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, (Red Hook, NY, USA), p. 1205–1213, Curran Associates Inc., 2012.

[70] L. P. Ansu Berg, Rigardt Pretorius, "Exploring the treatment of selected typologial characteristics of tswana in lfg," in *Proceedings of the LFG12 Conference*, 2012.

[71] L. Pretorius and S. Bosch, "Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology," in *Proceedings of the First Workshop on Language Technologies for African Languages* (L. Levin, J. Kiango, J. Klavans, G. De Pauw, G.-M. de Schryver, and P. W. Wagacha, eds.), (Athens, Greece), pp. 96–103, Association for Computational Linguistics, Mar. 2009.

[72] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (E. Blanco and W. Lu, eds.), (Brussels, Belgium), pp. 66–71, Association for Computational Linguistics, Nov. 2018.

[73] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.

[74] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 4768–4777, Curran Associates Inc., 2017.

Appendices

Chapter A: Data Statement for the datasets used in the research

A.1 WMT22[1]

Data set name: WMT22 African dataset[1]
Link to Dataset: WMT22 African Huggingface
Data set developer(s): Meta AI (multiple research contributors)
Data statement author(s): Meta AI (multiple research contributors)

Curation rationale

Please refer to the Huggingface dataset documentation or github repository for detailed curation rationale. The description below motivates the use of the data in this research.
In order to evaluate the effect that data has on translation tasks, we looked at the datasets that were used in the training and testing of the M2M100 model. One of these datasets was the WMT dataset.
We chose WMT22 data for training the ONMT and M2M100 models because of its high quality and extensive coverage, making it an excellent benchmark for machine translation. WMT22 includes diverse language pairs and rich parallel corpora, which are crucial for developing models for low-resource languages. We were able to select the low resource language pairs Xhosa-Zulu and Tswana-Zulu based on the corpora mappings. Using this standardized dataset allows for consistent and fair comparisons between different models and techniques. Using this reputable dataset also allows of expanding this research across various models in future.

Language varieties/text characteristics

This dataset was created based on metadata for mined bitext released by Meta AI. It contains bitext for 248 pairs for the African languages that are part of the 2022 **WMT Shared Task on Large Scale Machine Translation Evaluation for African Languages**[1].

Licensing information

Dataset is released under the terms of ODC-BY, bound by Terms of Use as specified by the Internet Archive

Other

Please refer to the original paper: **WMT Shared Task on Large Scale Machine Translation Evaluation for African Languages**[1]

Provenance appendix

The dataset was accessed from Hugging Face in April 2023. The data and/or URL may change or be updated as per research considerations from Facebook Research/Meta AI.

A.2 The Vuk'uzenzele South African Multilingual Corpus[2, 3]

Data set name: The Vuk'uzenzele South African Multilingual Corpus[2, 3]
Link to Dataset: Github repository
Data statement author(s): Vukosi Marivate, Andani Madodonga, Daniel Njini, Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab

Curation rationale

Please refer to the github repository dataset documentation for detailed curation rationale. The description below motivates the use of the data in this research. We looked for a dataset that had aligned sentence pairs for Xhosa-Zulu and Tswana to Zulu that we could use for human evaluation. We also needed Zulu target sentences in order to measure MMD scores against for the generated model translations. This dataset provided the aligned sentence pairs for the language pairs. We also ensured none of the sentences overlapped with data in WMT22.

Language varieties/text characteristics

The data set contains parallel statements from the Vukuzenzele magazine publication across the 11 official SA languages. The magazine is managed by the Government Communication and Information System (GCIS).

Licensing information

Dataset is licenced under Creative Commons.

Other

Please refer to the original paper: **Preparing the Vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora[2]**

Provenance appendix

The dataset was accessed from the DSFSI GitHub repository in April 2024. The data and/or URL may change or be updated as per research considerations from the DSFSI research group.

Chapter B: Human evaluation methods

B.1 Multidimensional Quality Metrics (MQM) questionnaire

This questionnaire aims to evaluate the correctness of translations generated by both the LLM and the small-scale model. The evaluation focuses on several dimensions, including accuracy, fluency, terminology, and overall quality. Please read each statement and rate the translation based on the criteria provided. Use the scale below for each question:

Rating scale:

- **5 - Excellent:** The translation is flawless concerning the specific criterion.
- **4 - Good:** The translation has minor issues but meets the criterion effectively.
- **3 - Neutral:** The translation has noticeable issues that moderately affect quality.
- **2 - Poor:** The translation has significant issues that considerably affect quality.
- **1 - Very Poor:** The translation does not meet the criterion and has critical flaws.

Section 1: Accuracy

Faithfulness

1.1 The translation accurately conveys the meaning of the source text.

1.2 The translation preserves the key information from the source text.

Completeness

1.3 The translation includes all important details from the source text.

1.4 No significant information is missing in the translation.

Consistency

1.5 The translation is consistent in terminology throughout the text.

1.6 Repeated phrases or terms are translated consistently.

Section 2: Fluency

Grammaticality

2.1 The translation is grammatically correct.

2.2 The translation follows the syntax rules of the target language.

Punctuation and spelling

2.3 The translation uses correct punctuation.

2.4 The translation is free from spelling errors.

Section 3: Terminology

Technical terms

3.1 The translation uses appropriate technical terms.

3.2 The technical terms are used correctly in context.

Consistency in terminology

3.3 The translation uses consistent terminology throughout the document.

3.4 The terminology used is appropriate for the subject matter.

Section 4: Locale convention

4.1 The translation represents dates, currency and names in the correct format

4.2 The translation meets the expected standards for the given task.

Section 5: Style

5.1 The translation maintains a natural flow and is easy to read.

5.2 The style of the translation is appropriate for the target audience.

5.3 The translation is a non-translation (Impossible to reliably characterize distinct errors)

Overall evaluation

6.1 The overall quality of the translation is satisfactory.

6.2 The translation meets the expected standards for the given task.

Demographics (optional)

- What is your level of expertise in machine translation? (Beginner, Intermediate, Advanced)
- What is your level of familiarity with the low-resource language pair? (Fluent)

B.2 Explanation Satisfaction Scale (ESS) questionnaire

This questionnaire aims to measure human evaluators' satisfaction with the post-model evaluation techniques for both the LLM and the small-scale model in the context of low-resource language translation. The post-model evaluation techniques include BLEU scores, attention heatmaps, attention pattern analysis, and MMD distance measurements.

Please rate the following statements on a scale of 1 to 5, where 1 means "Strongly Disagree", 2 means "Disagree", 3 means "Neutral", 4 means "Agree" and 5 means "Strongly Agree."

Section 1: BLEU scores

Clarity

1.1 The BLEU scores provide clear numerical indications of translation quality.

1.2 I find the interpretation of BLEU scores straightforward.

Detail

1.3 The BLEU scores offer detailed quantitative information about translation performance.

1.4 The explanations accompanying the BLEU scores are comprehensive and clear.

Helpfulness

1.5 The BLEU scores are helpful in evaluating the overall quality of translations.

1.6 These scores aid in understanding the relative performance of different models.

Trust

1.7 The BLEU scores increase my trust in the model's translation quality assessment.

1.8 I feel more confident in the model's translations after reviewing the BLEU scores.

Overall satisfaction

1.9 Overall, I am satisfied with the explanations provided by the BLEU scores.

Section 2: Attention heatmaps

Clarity

2.1 The attention heatmaps provide clear visualizations of the attention patterns between source and target translations.

2.2 I find the attention heatmaps easy to interpret.

Detail

2.3 The attention heatmaps provide detailed information about the attention weights at each position.

2.4 The granularity of the attention heatmaps is sufficient for understanding the model's focus.

Helpfulness

2.5 The attention heatmaps are helpful in identifying translation errors.

2.6 The attention heatmaps aid in understanding how the model handles different linguistic features.

Trust

2.7 The attention heatmaps increase my trust in the model's translation process.

2.8 I feel more confident in the model's translations after examining the attention heatmaps.

Overall satisfaction

2.9 Overall, I am satisfied with the explanations provided by the attention heatmaps.

Section 3: Attention pattern analysis

Clarity

3.1 The attention pattern analysis provides clear insights into the model's focus areas.

3.2 The visualizations of attention patterns are easy to understand.

Detail

3.3 The attention pattern analysis offers detailed information about the attention weights at each position.

3.4 The explanations accompanying the attention patterns are thorough and clear.

Helpfulness

3.5 The attention pattern analysis helps identify which parts of the sentence the model prioritizes during translation.

3.6 This analysis aids in understanding how the model processes linguistic structures.

Trust

3.7 The attention pattern analysis increases my trust in the model's translation quality.

3.8 I feel more confident in the model's translations after reviewing the attention pattern analysis.

Overall satisfaction

3.9 Overall, I am satisfied with the explanations provided by the attention pattern analysis.

Section 4: MMD distance measurements

Clarity

4.1 The MMD distance measurements provide clear numerical insights into the similarity between target and reference translations.

4.2 I find the MMD distance measurements easy to interpret.

Detail

4.3 The MMD distance measurements offer detailed quantitative information about translation quality.

4.4 The explanations accompanying the MMD distances are thorough and clear.

Helpfulness

4.5 The MMD distance measurements are helpful in assessing the quality of translations.

4.6 These measurements aid in understanding the differences between model translations and reference translations.

Trust

4.7 The MMD distance measurements increase my trust in the model's translation evaluation process.

4.8 I feel more confident in the model's translation quality after examining the MMD distances.

Overall satisfaction

4.9 Overall, I am satisfied with the explanations provided by the MMD distance measurements.

Additional questions (optional)

Comparative evaluation

5.1 Which post-model evaluation technique do you find most useful overall? (BLEU scores, Attention Heatmaps, Attention Pattern Analysis, MMD Distance Measurements)

5.2 Which post-model evaluation technique do you find least useful overall? (BLEU scores, Attention Heatmaps, Attention Pattern Analysis, MMD Distance Measurements)

5.3 Please provide any additional comments or suggestions for improving the post-model evaluation techniques.

Demographics

6.1 What is your level of expertise in machine translation? (Beginner, Intermediate, Advanced)

6.2 What is your level of familiarity with low-resource languages? (Beginner, Intermediate, Advanced)

B.3 Qualitative responses to ESS survey

As per ESS survey in Appendix B.2, we posed a list of optional additional questions to the human evaluators. The three evaluators were in agreement in their analysis, providing the responses below for LLM and small-scale model.

B.3.1 Qualitative survey responses for LLM

Comparative evaluation

- **Which post-model evaluation technique do you find most useful overall?** (BLEU scores, Attention heatmaps, Attention pattern analysis, MMD)
  - MMD
  - Attention pattern analysis
  - BLEU scores
- **Which post-model evaluation technique do you find least useful overall?** (BLEU scores, Attention heatmaps, Attention pattern analysis, MMD)
  - Attention heatmaps
- **Please provide any additional comments or suggestions for improving the post-model evaluation techniques.**
  - Better visualization of attention pattern distribution across different layers, not just the last layer of the decoder, to obtain more fine-grained perspectives on model focus.

  - Consider integrating different evaluation metrics such as COMET for a more comprehensive assessment that aligns with human evaluation.

Demographics

- **What is your level of expertise in machine translation?** (Beginner, intermediate, advanced)
  - Advanced
- **What is your level of familiarity with low-resource languages?** (Beginner, intermediate, advanced)
  - Advanced

B.3.2 Qualitative survey responses for small-scale model

Comparative evaluation

- **Which post-model evaluation technique do you find most useful overall?** (BLEU scores, Attention heatmaps, Attention pattern analysis, MMD)
  - MMD
  - Attention pattern analysis
- **Which post-model evaluation technique do you find least useful overall?** (BLEU scores, Attention heatmaps, Attention pattern analysis, MMD)
  - BLEU scores
- **Please provide any additional comments or suggestions for improving the post-model evaluation techniques.**
  - Exploring other methods for evaluation of why ONMT has so many failures would be ideal.
  - Although suggested methods do serve as a good starting point into evaluation, more can be done.

Demographics

- **What is your level of expertise in machine translation?** (Beginner, intermediate, advanced)
  - Advanced
- **What is your level of familiarity with low-resource languages?** (Beginner, intermediate, advanced)
  - Advanced