Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models

Pitso Walter Khoboko [a,*], Vukosi Marivate [a,b], Joseph Sefara [c]

[a] Data Science for Social Impact, University of Pretoria, South Africa
[b] Lelapa AI , South Africa
[c] Council for Scientific and Industrial Research (CSIR), Meiring Naudé Road , Brummeria, Pretoria, South Africa

ARTICLE INFO

Keywords:
Large language model (LLM)
PEFT
Mistral 7B
Prompt Engineering
In-Context-Learning (ICL)
En-Zul (English-to-Zulu)
En-Xh (English-to-Xhosa)
BLUE-score
F1score
G-Eva (mean cosine-similarity score)

ABSTRACT

Training large language models (LLMs) can be prohibitively expensive. However, the emergence of new Parameter-Efficient Fine-Tuning (PEFT) strategies provides a cost-effective approach to unlocking the potential of LLMs across a variety of natural language processing (NLP) tasks. In this study, we selected the Mistral 7B language model as our primary LLM due to its superior performance, which surpasses that of LLAMA 2 13B across multiple benchmarks. By leveraging PEFT methods, we aimed to significantly reduce the cost of fine-tuning while maintaining high levels of performance.

Despite their advancements, LLMs often struggle with translation tasks for low-resource languages, particularly morphologically rich African languages. To address this, we employed customized prompt engineering techniques to enhance LLM translation capabilities for these languages.

Our experimentation focused on fine-tuning the Mistral 7B model to identify the best-performing ensemble using a custom prompt strategy. The results obtained from the fine-tuned Mistral 7B model were compared against several models: Serengeti, Gemma, Google Translate, and No Language Left Behind (NLLB). Specifically, Serengeti and Gemma were fine-tuned using the same custom prompt strategy as the Mistral model, while Google Translate and NLLB Gemma, which are pre-trained to handle English-to-Zulu and English-to-Xhosa translations, were evaluated directly on the test data set. This comparative analysis allowed us to assess the efficacy of the fine-tuned Mistral 7B model against both custom-tuned and pre-trained translation models.

LLMs have traditionally struggled to produce high-quality translations, especially for low-resource languages. Our experiments revealed that the key to improving translation performance lies in using the correct prompt during fine-tuning. We used the Mistral 7B model to develop a custom prompt that significantly enhanced translation quality for English-to-Zulu and English-to-Xhosa language pairs. After fine-tuning the Mistral 7B model for 30 GPU days, we compared its performance to the No Language Left Behind (NLLB) model and Google Translator API on the same test dataset. While NLLB achieved the highest scores across BLEU, G-Eval (cosine similarity), and Chrf++ (F1-score), our results demonstrated that Mistral 7B, with the custom prompt, still performed competitively.

Additionally, we showed that our prompt template can improve the translation accuracy of other models, such as Gemma and Serengeti, when applied to high-quality bilingual datasets. This demonstrates that our custom prompt strategy is adaptable across different model architectures, bilingual settings, and is highly effective in accelerating learning for low-resource language translation.

1. Introduction

Large language models (LLMs) have become central to advancements in Natural Language Processing (NLP), recognized for their impressive performance across a wide range of tasks. Notably, ChatGPT has gained widespread attention for its ability to generate human-like conversational responses (Brown et al., 2020). However, despite these achievements, a significant challenge remains due to the lack of transparency around the training data and methodologies used in these models. This opacity limits further research into the inner workings of LLMs and constrains efforts to improve them. Previous studies have also noted that LLMs struggle with translation tasks, particularly for

[*] Corresponding author.
E-mail addresses: u21824772@tuks.co.za (P.W. Khoboko), vukosi.marivate@cs.up.ac.za (V. Marivate), tsefara@csir.co.za (J. Sefara).

https://doi.org/10.1016/j.mlwa.2025.100649
Received 21 January 2025; Received in revised form 6 March 2025; Accepted 30 March 2025
Available online 9 April 2025

low-resource, morphologically complex African languages (Ojo et al., 2023). A key factor in enhancing LLM performance in translation appears to be crafting the right prompt to train the model effectively (Ghazvininejad et al., 2023; Pourkamali & Sharifi, 2024; Zhu et al., 2023). In this study, we conducted experiments to design a custom prompt that achieved state-of-the-art results in English-to-Zulu and Xhosa translation tasks, emphasizing the importance of selecting the right LLM for our experiments.

In response to this challenge, efforts have been made to democratize access to LLMs models by releasing models like Mistral to the public, by making the code publicly available (Jiang et al., 2023). By making these models available, researchers and practitioners alike can delve deeper into the intricacies of fine-tuning large models, thus advancing our collective understanding of state-of-the-art NLP techniques. Through the provision of accessible instructions, Large language model architecture (LLAMA) models aim to empower a broader community of researchers to explore novel applications and unlock new capabilities.

Fine-tuning the full parameters of LLMs for various NLP tasks is often prohibitively expensive due to their massive size and the computational resources required. This approach demands significant GPU/TPU memory, longer training times, and higher energy consumption, making it impractical for many applications, especially for organizations with limited resources. Until recently, the expense of fine-tuning posed a significant barrier to conducting research in this area. However, the introduction of Parameter Efficient Fine-tuning (PEFT) methods,this involves fine-tuning only a subset of parameters of LLMs while keeping the others frozen (Liu, Gao, & Belinkov, 2023). Such as the widely adopted low-Rank adaptation (LORA) method, has offered a solution to this challenge (Hayou et al., 2024; Hu et al., 2021). Thus, reducing the computational resources required for fine-tuning, thus it is more feasible to explore the capabilities of LLAMA models in the context of low-resource languages (Balne et al., 2024).

In this paper, we propose a methodology for fine-tuning the Mistral 7B model (Jiang et al., 2023), and not the latest LLAMA 3 (Dubey et al., 2024). We chose the Mistral 7B model as our main model, as it outperformed the LLAMA 13B, Zhu et al. (2023) proved the Mistral 7B model our 13B model thus we leverage training a small parameterized model and getting good performance of a bigger parameterized model. Addressing these underrepresented languages is crucial for bridging communication gaps and promoting inclusivity. While early NLP translation efforts yielded mixed results (Ojo et al., 2023), recent advancements, particularly in prompt engineering techniques, have significantly enhanced the performance of LLM models. Our research not only advances technical aspects of NLP but also has the potential to impact communities relying on Xhosa and Zulu, supporting education and cultural preservation. By demonstrating high-quality translations with limited data resources, we aim to amplify the digital presence and utility of these languages.

Studies have shown that leveraging In-Context Learning (ICL) with simple instruction prompts can significantly improve translation performance compared to baseline models (Bertsch et al., 2024). Additionally, incorporating prompts with dictionary definitions of words has shown promise (Bawden & Yvon, 2023), particularly in the context of translating between low-resource languages. Also there have been other research that has shed light on prompt strategies that assist LLMs in the translation task for low resource settings (Ghazvininejad et al., 2023; Pourkamali & Sharifi, 2024; Zhu et al., 2023). Building upon these insights, we developed a custom prompt tailored for translation between English and two low-resource native South African languages, Xhosa and Zulu.

Research has revealed that LLAMA models exhibit superior performance in multi language fine-tuning compared to bilingual fine-tuning (Lakew et al., 2021), especially for low-resource languages. By leveraging a custom prompt, previous simple prompt was used to train models on various NLP task and translation task for low resourced African language was had the worst results which made authors conclude that for better performance more data is needed (Ojo et al., 2023), that incorporates various prompt strategies we achieved a high BLEU score of 20 for English to Zulu translation—a noteworthy accomplishment in the realm of translation between English and two native South African languages on LLAMA.

LLAMA models represent a powerful tool in the field of NLP, continually pushing the boundaries of what is possible. Through ongoing research and innovation, these models can be harnessed to address critical challenges in translation and other NLP tasks, ultimately advancing cross-lingual communication and understanding.

Our scientific contribution highlights that prompt engineering is the key to improving translation performance for morphologically rich, low-resource African languages on LLMs. This breakthrough not only brings African languages into the rapidly growing landscape of LLMs but also enables these languages to contribute meaningfully to the ongoing improvement of large language models.

2. Literature study

2.1. PEFT transformer based LLMs

In the study by Zhang, Zhang, andXu (2023), the authors investigated the use of large language models (LLMs) for Visual Question Answering (VQA) on the ScienceQA dataset. The dataset used includes a comprehensive set of questions across various domains such as natural science (NAT), social science (SOC), language (LAN), text (TXT), and images (IMG). The methods employed include a comparison of multiple models: LLaMA-Adapter, GPT-3, ChatGPT, and GPT-4, among others. The LLaMA-Adapter model is an efficient fine-tuning method that transforms the LLaMA language model into a high-performing instruction-following model by inserting learnable adaptation prompts into its higher layers. These prompts are designed to inject new instructions into the model while preserving pre-trained knowledge through a zero-initialized attention mechanism with a learnable gating factor. The evaluation metric used was question answering accuracy. The results showed that the LLaMA-Adapter model achieved 78.31% accuracy with only 1.2 million parameters, demonstrating its efficiency. In comparison, GPT-4 with chain-of-thought prompting (GPT-4CoT) achieved the highest accuracy at 83.99%. This indicates a significant improvement over the baseline methods, such as the VisualBERT model, which had an accuracy of 61.87%. The study highlighted that the LLaMA-Adapter model is highly effective and efficient, particularly with its minimal parameter usage, and the languages involved in the VQA tasks were diverse, covering both high-resource and low-resource contexts.

Work by Lv et al. (2023) proposed an innovative approach to fine-tuning large language models (LLMs) with limited GPU resources, addressing a significant barrier in NLP research. The authors introduce a new optimizer called LOw-Memory Optimization (LOMO), which integrates gradient computation and parameter updates to drastically reduce memory usage. Their method allows for the full parameter fine-tuning of a 65 billion parameter model on a single machine with 8 RTX 3090 GPUs. The effectiveness of LOMO is demonstrated through experiments on the SuperGLUE benchmark, showcasing its ability to optimize LLMs efficiently with constrained resources. The results indicate that LOMO significantly lowers memory consumption, enabling broader participation in LLM research without compromising performance.

Research by Zhang et al. (2023) introduced Dynamic Sparse No Training (DS0T), a training-free fine-tuning approach that enhances the performance of sparse Large Language Models (LLMs) without requiring backpropagation or weight updates. This method was applied to sparse versions of the LLaMA model family, including LLaMA-V1 and LLaMA-V2, across various sparsity levels. They evaluated the models using datasets such as WikiText-2 for language modeling and a set of zero-shot tasks: PIQA, StoryCloze, ARC (Easy and Challenge), HellaSwag, and OpenBookQA. DS0T involves iterative pruning and growing of weights to minimize the reconstruction error between dense and sparse model outputs. The evaluation metrics included perplexity for language modeling and accuracy for zero-shot tasks. The results showed that DS0T consistently improved performance across different sparsity levels. For example, it reduced the perplexity of LLaMA-V1-7B models pruned with the Wanda metric from 7.26 to 7.12 at 50% sparsity and enhanced the average accuracy of SparseGPT-pruned models by 1.6%, achieving a mean accuracy of 52.7%. The best performance was seen with the LLaMA-V1-65B model pruned at 60% sparsity, where DS0T achieved a zero-shot accuracy of 63.3%. This demonstrates DS0T's effectiveness in maintaining high accuracy while managing computational efficiency, making it a significant advancement for fine-tuning sparse LLMs.

In the paper by Li et al. (2023), the authors introduce the Parameter-Efficient and Quantization-aware Adaptation (PEQA) method, which aims to bridge the gap between parameter-efficient fine-tuning (PEFT) and quantization, which is when model's weights and activation are represented with low-precision, in large language models (LLMs). The evaluation involves datasets like Wikitext2 and PennTreeBank, and the Alpaca dataset for instruction tuning. The evaluation metrics primarily include perplexity (PPL) for assessing the model's performance. PEQA is compared against traditional methods such as LoRA and OPTQ. The results demonstrate that PEQA achieves competitive performance, particularly when using low-bit quantization (3-bit and 4-bit), showing less performance degradation compared to other methods. The best results were observed with the PEQA method, achieving lower perplexity scores across various LLM sizes, including GPT-Neo, GPT-J, and LLaMA models, highlighting its efficacy in maintaining model performance while reducing memory usage and improving inference speed. In the study by Sun et al. (2023), the authors conducted a comparative analysis between full-parameter and LoRA-based fine-tuning on Chinese instruction data using the LLaMA model. The dataset used included various scales of instruction data, specifically 0.6 million, 2 million, and 4 million samples, evaluated across nine real-world use cases. The evaluation metric employed was the average score assigned by ChatGPT (0 to 1 scale), assessing tasks like translation, open QA, closed QA, generation, and more. The methods compared were full-parameter fine-tuning and LoRA-based tuning, with LoRA demonstrating significant training efficiency. Among the methods, full-parameter fine-tuning of LLaMA-7B + FT(2M) + FT(math_0.25M) achieved the highest average score of 0.738. This study highlights LoRA's efficiency, showing competitive performance with fewer additional parameters and reduced training time compared to full-parameter fine-tuning, and underscores its effectiveness in low-resource settings.

Work by Lialin et al. (2023) present a novel method, ReLoRA, for training large transformer language models using low-rank updates to enhance computational efficiency. The method is tested on the C4 dataset with transformer models of varying sizes (60M, 130M, 250M, 350M parameters). Evaluation is conducted using the perplexity metric, where lower values indicate better performance. ReLoRA achieves significant improvements over baseline methods, particularly in larger models, demonstrating a perplexity of 19.32 in the 350M parameter model, compared to 18.66 for full-rank training. The proposed method leverages a combination of restarts, partial optimizer resets, and a jagged learning rate scheduler to improve training efficiency and performance, making it a viable alternative to traditional high-rank training methods for large-scale neural networks.

In their study Dubey el. at. (Dettmers et al., 2024) introduce QLORA, an efficient fine-tuning technique for quantized large language models (LLMs). The study utilizes multiple datasets, including Alpaca and FLAN v2, and evaluates models using 5-shot accuracy tests on the MMLU benchmark. The methods explored include 4-bit NormalFloat (NF4) quantization, Double Quantization, and Paged Optimizers, designed to minimize memory usage while preserving performance. Their results demonstrate that QLORA can achieve performance comparable to full 16-bit fine-tuning methods, with significantly reduced resource demands. The best-performing configuration, leveraging NF4 with Double Quantization, achieved a mean MMLU accuracy of 63.9% on the FLAN v2 dataset for the 65B LLaMA model, closely matching the performance of BFloat16 methods. Given these advantages, we adopt PEFT strategy QLORA to fine-tune all of the models discussed in this paper.

In a separate study, Jiang et al. (2023) introduced Mistral 7B, a 7-billion-parameter model optimized for both efficiency and performance. By employing innovations such as grouped-query attention (GQA) and sliding window attention (SWA), Mistral 7B enhances inference speed and manages long sequences more effectively. It outperforms the 13B LLaMA 2 model across a variety of benchmarks and even surpasses the 34B LLaMA 1 model in tasks like reasoning, mathematics, and code generation. The datasets used in their study span a broad range of tasks, including commonsense reasoning, world knowledge, reading comprehension, mathematics, and code, with results reported on benchmarks such as Hellaswag, Winogrande, and HumanEval. With its balance of computational efficiency and high-level performance, Mistral 7B is an ideal model for fine-tuning specific applications.

Given these strengths, we have selected Mistral 7B as the primary model for our experiments and leverage QLORA as our fine-tuning approach. This combination enables us to maximize both performance and resource efficiency, effectively applying the innovations from Dettmers et al. (2024) while benefiting from the superior architecture and task-specific capabilities of Mistral 7B.

2.2. LLMs for translation

Work by Reynolds and McDonell (2021) explored the efficacy of different prompting strategies for GPT-3 in translation tasks, specifically using the WMT'14 French-English dataset. The study focuses on low-resource settings by analyzing zero-shot (0-shot) and few-shot (1-shot and 10-shot) prompts. The evaluation metric employed is the BLEU score, measured via SacreBLEU, which demonstrated that a simple colon prompt (i.e., specifying the language names followed by a colon) significantly outperformed the original GPT-3 few-shot results, achieving a BLEU score of 33.3 compared to 22.4 in the 0-shot scenario for the Curie model (13B parameters). This study highlights "prompt programming" as a novel method for improving translation quality without relying on extensive training examples, illustrating that properly engineered prompts can leverage the model's pre-existing knowledge to enhance performance significantly.

In their 2023, study Zhang, Zhang, andXu (2023) utilized the INT4-quantized version of the GLM-130B model to experiment with various high-resource languages and prompt templates. They focused on translations between English, German, and Chinese. The highest-performing prompts were those templated in English using the Latin alphabet, aligning with the model's original fine-tuning dataset. A simple prompt template that included the language names for input and output in a few-shot setting significantly improved their results. On Wikipedia data, this approach achieved a high score of 72.8 on the COMET evaluation metric for German to English translation. In their research, Ghazvininejad et al. (2023) introduced DIMPT, a method using phrase-level prompts to define rare words for improving translations in low-resource and out-of-domain settings. They evaluated their approach with OTP-175B and BLOOM models on the Flores-101 dataset, focusing on ten low-resource languages: Catalan, Croatian, Danish, Dutch, Filipino, Indonesian, Italian, Malay, Norwegian, and Slovak. For out-of-domain scenarios, they used a German to English dataset covering IT, Law, Medical, and Quranic texts. Their method significantly enhanced translation quality across all languages except Catalan, which scored 1 BLEU point less than the baseline of using a simple prompt. These results demonstrate DIMPT's potential to improve large language models' translation performance in challenging contexts.

In the study by Pourkamali and Sharifi (2024), the authors explore the application of LLMs for machine translation (MT) across

Persian, English, and Russian. They utilize different prompting methods including n-shot feeding and tailored prompting frameworks to enhance translation accuracy. The models tested include GPT-3.5, GPT-4, PaLM2, LLaMA-2-70B, and Claude 2. Evaluation metrics used in the study were BLEU, chrF++, and COMET scores, providing a comprehensive assessment of translation quality. Among the models, PaLM2 demonstrated competitive performance for both high-resource languages like English and Russian, and low-resource languages like Persian. This indicates a robust capability in multilingual contexts, highlighting the model's efficiency in handling varied language pairs. Conversely, while GPT-4 performed exceptionally well with high-resource languages, it showed limitations with less resourced and more linguistically distant languages. The study emphasizes the potential and challenges of using LLMs for MT, noting that while they achieve remarkable results, they are not free from generating linguistic errors and hallucination.

Work by Song et al. (2023) addressed the challenge of improving machine translation for low-resource languages (LRLs) through knowledge distillation and language-approaching techniques. They combined large parallel high-resource language (HRL) data with monolingual LRL data, using a multilingual teacher model (NLLB) to train smaller, efficient student models. Their method focused on translating Luxembourgish to English, adapting German resources to enhance translation quality. Evaluated using BLEU scores on datasets like Tatoeba and Flores-200, their approach significantly improved translation performance, demonstrating that knowledge distillation and related-language techniques are effective for creating high-quality, resource-efficient translation models for LRLs.

In their research, Bawden and Yvon (2023) evaluated the translation performance of the BLOOM model, focusing on both high-resource and low-resource language pairs using datasets like WMT, Flores-101, and DiaBLa. Their high performing prompt included having an equal sign in the input, to show the model that the source language sentence is equivalent to the target language sentence. They primarily used the BLEU score for evaluation. The study found that BLOOM's performance improved significantly in few-shot settings compared to zero-shot scenarios, although issues like overgeneration and producing text in the wrong language were noted. Among the low-resource languages tested, those with better performance included directions into English, while languages like Swahili and Yoruba performed less well. These findings underscore the model's variability and the importance of prompt design and model size for optimizing translation quality in multilingual contexts.

In their study, Zhu et al. (2023) assess the translation capabilities of various large language models (LLMs) using the FLORES-101 dataset, which includes 102 languages and 606 translation directions. The study focuses on low-resource languages, revealing that while models like XGLM, OPT, BLOOMZ, Falcon, LLaMA2, ChatGPT, and GPT-4 show improvements, significant challenges remain. They use evaluation metrics such as SentencePiece BLEU (spBLEU) and COMET, highlighting that cross-lingual examples enhance translation performance for low-resource languages. The research identifies the template.

"$\langle X \rangle = \langle Y \rangle$" as yielding the best results, achieving the highest average BLEU score among various formats. This template outperformed others, like "[SRC]: $\langle X \rangle$ n [TGT]: $\langle Y \rangle$", demonstrating the critical role of template design in in-context learning. The study underscores that despite some advancements, current LLMs, including GPT-4, still struggle with low-resource languages compared to systems like Google Translate.

In their research, Ojo et al. (2023) presented a study evaluating the performance of three popular large language models (LLMs) – mT0, LLaMa 2, and GPT-4 – on five natural language processing tasks: news topic classification, sentiment classification, machine translation, question answering, and named entity recognition. The evaluation spans 30 African languages, highlighting the models' capabilities in low-resource settings. The authors used human-translated data and designed English prompts for the tasks. They discovered that GPT-4 performed well on classification tasks but poorly on generative tasks like machine translation. Surprisingly, mT0 outperformed other models, including the state-of-the-art fine-tuned mT5, in cross-lingual question answering. LLaMa 2 recorded the worst performance due to its English-centric pre-training corpus. The study emphasizes the need for better representation of African languages in LLMs to bridge the performance gap observed when compared to high-resource languages.

Work by Mujadia et al. (2023) explore the multilingual capabilities of large language models, specifically focusing on machine translation between English and 22 Indian languages. They evaluated raw models (LLaMa-2-7b and LLaMa-2-13b) and fine-tuned them using parameter-efficient methods like LoRA and full fine-tuning. Their results show significant progress, with the fine-tuned LLaMa-13b models achieving notable BLEU and CHRF scores across multiple test sets including IN22, flores200-dev, flores200-devtest, and newstest2019. For English to Indian languages, the best-performing model achieved an average BLEU score of 15.93 and a CHRF score of 46.99. For Indian languages to English, the scores were slightly lower. The study underscores the potential of LLMs in translating underrepresented languages, demonstrating that their performance can be significantly enhanced through fine-tuning.

In their research, Andersland (2024) focused on improving translation and language understanding for low-resource languages, specifically Amharic. They utilized LLaMA-2 and LLaVA models, leveraging various methods to enhance performance. Key techniques included data augmentation through translation, expanding the dataset from millions to billions of tokens. They used Seamless M4T for text-to-text translations and aligned image encoders for multimodal capabilities. The fine-tuning datasets included translations of Alpaca, Dolly, and OpenAssistant datasets. Their approach significantly improved translation quality, especially when incorporating mixed English-Amharic datasets, although challenges remained in tokenization and handling specific topics like STEM .

In their study Bertsch et al. (2024) investigated the effectiveness of in-context learning (ICL) with long-context models for translation tasks, focusing on both low-resourced and out-of-domain languages. They used five classification datasets: TREC, TREC-fine, NLU, Banking-77, and Clinic-150. Utilizing Llama-2-7b and Mistral-7b models with extended context lengths (up to 80,000 tokens), they observed significant performance improvements with increased context lengths. The performance gap between random and retrieved examples narrowed with longer contexts, and ICL generally outperformed traditional finetuning with fewer examples. These results demonstrate the potential of long-context ICL to enhance translation tasks, particularly for low-resourced and out-of-domain language pairs.

Work by Merx et al. (2024) investigate the use of large language models (LLMs) for translating English into Mambai, a low-resource Austronesian language spoken in Timor-Leste. Utilizing a novel corpus derived from a Mambai language and additional sentences translated by a native speaker, the research evaluates few-shot LLM prompting for machine translation (MT). The methods employed include the strategic selection of parallel sentences and dictionary entries using TF-IDF and semantic embeddings. Evaluation metrics used were BLEU and ChrF++, with the best scores being 21.2 BLEU and 41.8 ChrF++ on sentences from the language manual, compared to a baseline score of 4.4 BLEU on a test set by a native speaker, highlighting the significant variance based on corpus origin.

Inspired by various works on low-resourced languages, we designed custom prompts leveraging different techniques to improve the fine-tuning of LLMs for low-resource NLP downstream tasks, specifically translation tasks. Our ultimate goal is to identify the most effective prompts that enhance LLMs' performance in translating low-resourced languages. We focused our experiments on translating from English to Zulu and Xhosa. Employing a multilingual setting, which has demonstrated superior performance (Ojo et al., 2023), we aim to optimize translation quality for these languages.

Table 1
Dataset sizes used for fine-tuning the Mistral model for translation in low-resource settings. Eng stands for English, Zul for isiZulu, and Xh for isiXhosa. The datasets were sourced from the OPUS project and consist of general topics such as religion and politics.

Language pair	Description	Dataset size
Eng-Zul	English to isiZulu	3,899,128
Eng-Xh	English to isiXhosa	8,877,078
Total		12,776,206

Table 2
Shows the test with only general topics from Flores-101 evaluation test dataset. Filtered to have general topics.

FLORES-101 Dataset	
Language pair	Language pair
Zul	109
Xh	109
Eng	109

Large Language Models (LLMs) often struggle with translation tasks for African languages due to limited pre-training on these rich, morphologically complex languages (Ojo et al., 2023). However, recent advances have enabled novel prompt design strategies to improve translation accuracy for low-resource languages (Ghazvininejad et al., 2023). Building on methods like DIMPT, which uses dictionary definitions of rare words, we redefine English terms in target languages such as Zulu and Xhosa to increase lexical exposure (Ghazvininejad et al., 2023). Further, by marking source–target pairs with an equal sign and leveraging balanced in-context examples, our custom prompt design aims to optimize translation quality (Bawden & Yvon, 2023; Bertsch et al., 2024; Ghazvininejad et al., 2023; Song et al., 2023; Zhang, Haddow, & Birch, 2023; Zhu et al., 2023). Using the Mistral model – a refined version of LLAMA – we test these strategies to address the longstanding gap in LLM support for African languages, contributing to a more inclusive approach in AI language processing (Ojo et al., 2023).

3. Material and methods

3.1. Dataset collection

Below is the representation of the datasets utilized in our experiments and evaluations. Notably, we highlight the multilingual dataset used in our primary experiments, which serves as a benchmark for assessing the effectiveness of our approach across diverse languages. Additionally, we discuss the bilingual dataset employed to demonstrate that our custom prompt engineering strategy can enhance performance even in bilingual settings across various LLM architectures. These datasets form the foundation of our investigation into improving translation quality and adaptability. We also detail the process of cleaning and preparing the datasets for fine-tuning, utilizing a single standardized script to ensure consistency and compatibility across different LLMs.

3.1.1. Multilingual dataset

Here we showcase the multilingual dataset compiled from various sources. We detail the size and composition of the training and test datasets used for fine-tuning and evaluating LLMs across different experiments. This comprehensive dataset enables robust assessments of model performance in multilingual scenarios.

Table 3
The above depicts the amount of data that was collected from the high quality Umsuka dataset (Mabuya et al., 2021). As in the previous table "Eng" means English and "Zul" means Zulu. Both test and training dataset have two translations from professional translators, thus being highlighted in the training dataset. And the test dataset we only use one of the translations.

Umsuka English - Zulu Dataset		
Language pair	Training	Test
En-Zul	6444	1311

3.1.1.1. Training dataset. The bilingual English-to-Zulu and English-to-Xhosa datasets were collected to develop a multilingual model. Autshumato (Groenewald & Fourie, 2009), WTM22 (Pope et al., 2022), and OPUS (Tiedemann, 2012) is such dataset as which has various topic like religion, news, and more. The datasets were first cleaned to remove accents, numbers, and special characters, and to convert all text to lowercase. As shown in Table 1, the English-to-Xhosa corpus was initially larger. To create a balanced multilingual dataset, we matched the amount of English-to-Xhosa data to the English-to-Zulu data. A variety of prompt templates, which have been shown to improve translation for low-resource languages on various LLM models (Bawden & Yvon, 2023; Bertsch et al., 2024; Ghazvininejad et al., 2023; Song et al., 2023; Zhang, Haddow, & Birch, 2023; Zhu et al., 2023), were used to prepare the data for fine-tuning the Mistral 7-B model (Jiang et al., 2023). The combined multilingual dataset was then randomized. The model demonstrated high performance accuracy across different downstream NLP tasks, outperforming the LLAMA 13B model.

3.1.1.2. Test dataset. The FLORES-101 dataset (Goyal et al., 2022) is a highly accurate multi-directional translation corpus, validated by human annotators. From this dataset, we extracted English-to-Zulu and English-to-Xhosa translations, originally comprising 1,012 tokens from the dev-test set. To ensure alignment with the model's training domains, we reduced the dataset to 109 tokens. These tokens cover a variety of domains, including 'culture', 'disasters', 'crime', 'politics', 'conflicts', 'sports', 'accidents', 'tragedies', 'television', and 'entertainment.' This subset was designed to represent a general knowledge corpus, with the amount of data used summarized in Table 3.

3.1.2. Bilingual dataset

In this sub-subsection, we showcase the high-quality bilingual dataset used to demonstrate that, even in bilingual settings, our custom prompt engineering strategy can enhance translation for low-resource languages. Additionally, we highlight how phrase-level alignments between source and target sentences are incorporated into the custom prompt. This integration enables a deeper exploration of the prompt's full potential, maximizing its effectiveness in improving translation quality.

3.1.2.1. Umsuka dataset. Umsuka is an English-to-Zulu bilingual dataset created by professional translators fluent in both languages (Mabuya et al., 2021). We selected this dataset to demonstrate that our custom prompt is effective not only for the primary model used in our experiments but also for other large language models. Additionally, we sought to illustrate that our approach extends beyond multilingual contexts and performs equally well in bilingual settings.

This experiment further validates that, when leveraging high-quality datasets, translation accuracy can be significantly enhanced across various models. The training dataset initially comprised two sets: English-to-Zulu and Zulu-to-English translations. To streamline the learning process, we merged both sets and retained only the English-to-Zulu direction, forming a parallel dataset. This approach allowed the model to learn from two distinct Zulu translations of the same English sentences. During training, we utilized only the first translation for evaluation purposes in the test set.

Table 4
Depicts the multilingual statistical terminology which is a dataset that has definitions of English and Zulu phrases and words (Marivate et al., 2024).

Multilingual Statistical Terminology	
Language pair	Number of definitions
En-Zul	1153

This strategy provided a way to evaluate the model's performance on the second translation, which remained unseen during training. The total number of tokens used is presented in Table 2.

3.1.2.2. Multilingual statistical terminology dataset. Table 4 also presents the Multilingual Statistical Terminology Clean dataset, sourced from the Multilingual Statistical Terminology Project by Statistics South Africa (Marivate et al., 2024). This dataset includes phrases and word definitions in both English and Zulu, which were used to train the models with our final custom prompt. The full version of this prompt incorporates a crucial feature that provides dictionary hints to the model. This approach significantly aids in demonstrating that our prompt strategy can enhance the performance of various models, improving their ability to translate between English and low-resource or even extremely low-resource languages.

3.2. Dataset preparation

To prepare our datasets for effective model training and evaluation, we applied a uniform data-cleaning process across all sources using a single script, ensuring consistency and reliability in data quality. Data cleanliness is crucial in machine learning, especially in natural language processing (NLP) tasks, where inconsistent data can degrade model performance. Using a uniform cleaning method simplifies the workflow and minimizes human error, allowing for scalable modifications if further adjustments are necessary (Zhu et al., 2023).

In addition to standard pre-processing techniques, we implemented custom data-cleaning steps specifically designed to improve dataset quality for African languages such as Xhosa and Zulu. Our custom steps included sentence normalization to remove diacritical accents, as accent inconsistencies can interfere with tokenization and other language model processes (Pradha et al., 2019). Furthermore, we eliminated URLs, email addresses, Twitter handles, and usernames—elements that add noise without contributing valuable linguistic information, as found in studies addressing social media text processing (Zupon et al., 2021). Filtering out special characters and trimming extra white spaces created cleaner, more predictable input for our model (Ghafoor et al., 2021).

We also addressed the translation ratio challenge inherent in African languages, where morphological complexity often leads to longer translations of English sentences. In these cases, a single English word may translate to a multi-word expression in Xhosa or Zulu. To mitigate potential biases from uneven sentence lengths, we applied a 1:3 sentence-length ratio filter, retaining only sentence pairs where the target language did not exceed three times the English source length. This approach has been shown to improve low-resource language translation by focusing model learning on consistent sentence structures and avoiding overfitting to paragraph-length texts (Imankulova et al., 2017). By keeping sentence pairs concise, we aimed to guide the model's attention towards sentence-level accuracy, rather than extensive phrases that might degrade translation quality (Adamou et al., 2016).

Despite our thorough pre-processing, we acknowledge that high-quality training data for African languages remains a significant challenge. A considerable portion of the vocabulary in languages like Xhosa and Zulu has been directly borrowed from English, as highlighted in linguistic studies on African language corpora (Adebara, 2024). This phenomenon limits the model's capacity to learn native morphological structures, as borrowed terms do not follow traditional inflectional patterns within these languages. Without extensive, high-quality native linguistic data, it is difficult to fully train the model on the unique morphological richness of Xhosa and Zulu (Ranathunga et al., 2023). These limitations underscore the importance of ongoing efforts in data refinement and highlight the need for more expansive, representative datasets to enhance model performance in low-resource language settings (Ojo et al., 2023).

We selected the Opus and WTM 22 datasets as our primary training set, which offered a substantial volume of data for experimenting with English to Zulu and Xhosa multilingual pairs. As shown in Table 1, these datasets together provide over 12 million parallel corpus pairs. The main challenge with this dataset, however, is that it is not entirely clean—some Zulu and Xhosa words are borrowed from English, which affects linguistic purity. Nevertheless, once cleaned, it yields a robust dataset suitable for model training.

To evaluate our fine-tuned model, we used the widely adopted FLORES-101 dataset, which includes English, Xhosa, and Zulu. Although FLORES-101 presents similar issues to Opus and WTM 22, it allows us to assess how well our model has learned to translate the target languages in an unseen test set.

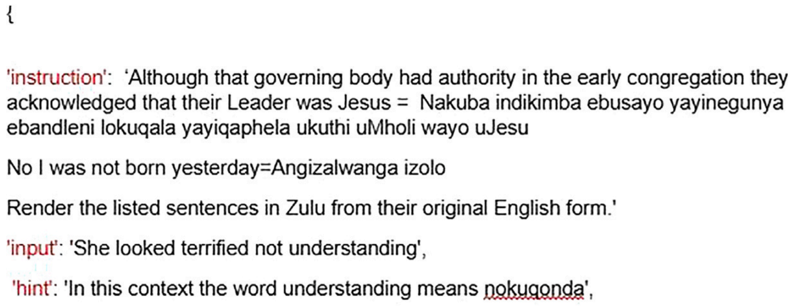
Additionally, we aimed to enhance translation quality in a bilingual setting by incorporating the high-quality Umsuka dataset, an English-to-Zulu dataset carefully curated by professional translators. This approach allowed us to test whether our custom prompts could improve translation not only in bilingual settings but also across various large language model architectures.

3.3. Prompt development

In this subsection, we present the various prompt templates used to prepare the dataset. Appendix A provides examples of the different prompts that were instrumental in creating the final JSON files for the datasets. These datasets were employed to fine-tune the Mistral 7B model, aiming to identify the most effective prompt template for enhancing LLM performance in translating low-resource South African languages from English.

Drawing inspiration from Bertsch et al. (2024), Brown et al. (2020), Groenewald and Fourie (2009) and Song et al. (2023), we adopted their prompt templates as foundational models. We chose these previous prompt strategies due to their demonstrated success in low-resource translation tasks within LLMs, experimenting with various combinations to derive the best-performing ensemble prompt for our fine-tuning. Custom templates were then developed to incorporate the most effective features from these foundational prompts, focusing on strengthening translation capabilities in low-resource settings—a well-known challenge for models like LLaMA 2, particularly in translating African languages (Ojo et al., 2023). A more detailed explanation of how various prompt engineering techniques affect the quality of translation can be found in the literature review and the previously discussed foundational research papers. These sources delve into the nuances of prompt design, highlighting how factors such as structure, phrasing, context length, and specificity significantly influence translation performance. By tailoring prompts to better align with the model's pre-training data and optimizing their clarity and relevance to the task, researchers have demonstrated marked improvements in translation accuracy, fluency, and semantic preservation. This study builds on those insights, drawing from prior research to explore how specific prompt engineering strategies can be applied to achieve more robust and contextually accurate translations in our experiments.

Through extensive experimentation with different prompts, our objective was to discover the optimal approach for improving translation accuracy in both the LLaMA 2 and Mistral models. Here, we focus on our custom-designed prompts rather than the initial universal templates, an example of which can be found in Appendix B.



{

'instruction': 'Although that governing body had authority in the early congregation they acknowledged that their Leader was Jesus = Nakuba indikimba ebusayo yayinegunya ebandleni lokuqala yayiqaphela ukuthi uMholi wayo uJesu

No I was not born yesterday=Angizalwanga izolo

Render the listed sentences in Zulu from their original English form.'

'input': 'She looked terrified not understanding',

'hint': 'In this context the word understanding means nokuqonda',

Fig. 1. Depicts how sentences are aligned in the prompt when fine tuning using 'Custom Prompt: Equal Sign ICL Dictionary Hint'.

3.3.1. *Custom prompt: Equal sign ICL dictionary hint*

The Equal Sign ICL Dictionary Hint prompt incorporates in-context learning (ICL) within the instruction, using one-shot input and output sentences. This approach, inspired by Bawden and Yvon (2023), employs a simple prompt format: "$\langle X \rangle = \langle Y \rangle$", where $\langle X \rangle$ represents the source sentence and $\langle Y \rangle$ represents the target language sentence. This method has been shown to improve translation for low-resource languages in LLMs. Building on this, we designed an in-context prompt where the sample input and output are identical, leveraging the benefits of ICL as demonstrated by Bawden and Yvon (2023).

Additionally, incorporating a dictionary hint at the phrase level to define rare words, as suggested by Ghazvininejad et al. (2023), further enhances translation quality. Our design includes sample input and output sentences in ICL that are identical, rather than just providing an instruction to translate. Moreover, instead of a dictionary of rare words, our dictionary hint provides definitions of rare English words translated into low-resource languages. This dual approach – using identical input and output in ICL and providing dictionary hints – aims to improve translation quality.

ICL has been shown to be essential in LLM translation tasks, particularly for low-resource languages (Brown et al., 2020), so we combined it with the equal-sign prompt structure to create a custom ensemble prompt. This combination aims to strengthen LLM translation capabilities for African languages. For an illustration of this approach, please refer to Fig. 1. A template for this custom prompt is also provided in Appendix A.

3.3.2. *Knowledge-infused few-shot prompt with dictionary hint*

This prompt is similar to the one discussed above, but it differs by incorporating a design that provides more input and output sentences in the in-context learning phase, significantly improving LLM translation as demonstrated by Bertsch et al. (2024). While we use the equal sign to represent source and target languages equally, this approach integrates a greater number of sentences, enhancing the context and thereby the translation quality. This method builds on our previous approach by extending the amount of data fed into the model during the learning phase, following the successful strategy outlined by Bertsch et al. (2024).

Building on our previous approach, this method extends the amount of contextual data fed into the model during the learning phase, aligning with the strategy outlined by Brown et al. (2020). By expanding our prompt to include more examples, we aim to assess the effects of increased contextual input, particularly for translations from English into morphologically rich languages like Xhosa. This approach allows us to explore how greater example exposure can optimize translation performance for languages with complex morphological structures. Fig. 2 depicts an example of the this prompt, template in Table B.12.

3.3.3. *Knowledge-infused few-shot prompt*

When employing the dictionary prompt, we encountered a reduction in dataset size. To identify rare words within the English source sentences that could be translated to Xhosa and Zulu, we used Google Translate (Tsai, 2019) alongside a curated list of high-frequency, non-domain-specific words. This list enabled us to detect rare terms in the source sentences that would benefit from translation into their target language counterparts. However, given the general nature of our dataset, many entries lacked a dictionary hint portion in the prompt, which led us to remove these entries.

To mitigate the impact of dataset reduction due to the missing dictionary hints, we trained an alternative model on the full multilingual dataset. This approach deviated from the original method by excluding the dictionary hint from the prompt. Through this adaptation, we aimed to maintain data volume without compromising the integrity of our translation task across low-resource African languages. Fig. 3 depicts an example of the this prompt, template in Appendix B.

3.3.4. *Knowledge-infused long-shot prompt*

With the loss of the full capabilities of our custom ensemble prompt, we were unable to experiment with the complete range of strategies initially planned. To compensate, we explored the use of long-shot in-context learning (ICL), following findings by Bertsch et al. (2024), which demonstrated improved performance compared to few-shot learning. However, their work also noted that providing too many examples could lead to overfitting, resulting in inaccurate translations. They found that an optimal number of examples for long ICL was around 10, which we adopted for our experiments.

By testing this long-shot ICL approach, we aimed to determine if it could offer similar benefits to those originally anticipated from the dictionary hint section of our prompt. This adjustment allowed us to explore whether long-shot ICL could serve as a viable substitute for the optimization lost due to the absence of the dictionary component. For a detailed example of this prompt, see Fig. 4, with the template provided in Table A.11.

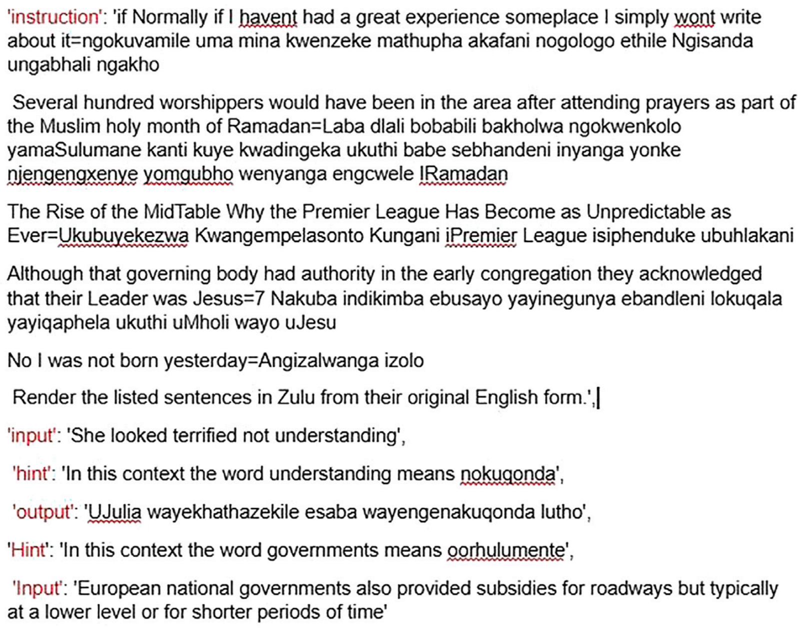
3.3.5. *Knowledge-infused long-shot prompt with Dic Hint for Gemma*

We designed a specialized template for the Gemma model, ensuring we adhered to the principles established in our custom model. The Gemma model requires a specific structure for user and model, including exact sentence boundaries and value placements. To meet these requirements, we used full supervised fine-tuning instead of zero-shot fine-tuning, facilitating the model's adaptation given its initial training on high-resource languages. Fig. 5 shows the format of the prompts used for fine-tuning and evaluation.

To enable a fair comparison between models, we kept experimental settings as consistent as possible between the Gemma model and our primary method, the Mistral 7B model.

3.3.6. *Knowledge-infused long-shot prompt with Dic Hint for Serengeti*

For the Serengeti model, a variant of the XML-RoBERTa model, we developed a slightly different template to ensure that the classification model could learn effectively while preserving the principles of our custom prompt. In this setup, all elements except the output were treated as inputs, with the output serving as the label. Since the



'instruction': 'if Normally if I havent had a great experience someplace I simply wont write about it=ngokuvamile uma mina kwenzeke mathupha akafani nogologo ethile Ngisanda ungabhali ngakho

Several hundred worshippers would have been in the area after attending prayers as part of the Muslim holy month of Ramadan=Laba dlali bobabili bakholwa ngokwenkolo yamaSulumane kanti kuye kwadingeka ukuthi babe sebhandeni inyanga yonke njengengxenye yomgubho wenyanga engcwele IRamadan

The Rise of the MidTable Why the Premier League Has Become as Unpredictable as Ever=Ukubuyekezwa Kwangempelasonto Kungani iPremier League isiphenduke ubuhlakani

Although that governing body had authority in the early congregation they acknowledged that their Leader was Jesus=7 Nakuba indikimba ebusayo yayinegunya ebandleni lokuqala yayiqaphela ukuthi uMholi wayo uJesu

No I was not born yesterday=Angizalwanga izolo

Render the listed sentences in Zulu from their original English form.',

'input': 'She looked terrified not understanding',

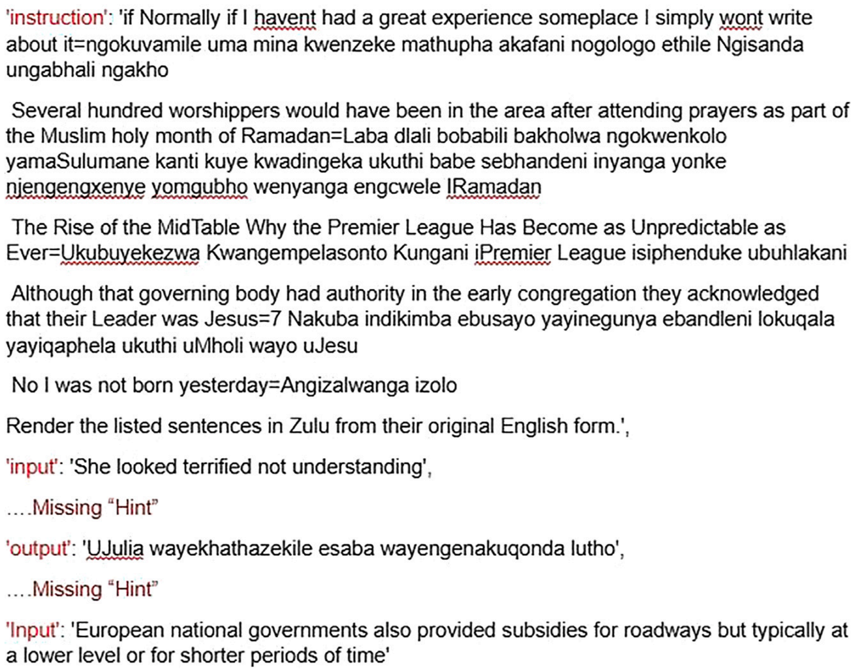
'hint': 'In this context the word understanding means nokuqonda',

'output': 'UJulia wayekhathazekile esaba wayengenakuqonda lutho',

'Hint': 'In this context the word governments means oorhulumente',

'Input': 'European national governments also provided subsidies for roadways but typically at a lower level or for shorter periods of time'

Fig. 2. Depicts how sentences are aligned in the prompt when fine tuning using 'Knowledge-Infused Few-Shot Prompt with Dictionary Hint'.



'instruction': 'if Normally if I havent had a great experience someplace I simply wont write about it=ngokuvamile uma mina kwenzeke mathupha akafani nogologo ethile Ngisanda ungabhali ngakho

Several hundred worshippers would have been in the area after attending prayers as part of the Muslim holy month of Ramadan=Laba dlali bobabili bakholwa ngokwenkolo yamaSulumane kanti kuye kwadingeka ukuthi babe sebhandeni inyanga yonke njengengxenye yomgubho wenyanga engcwele IRamadan

The Rise of the MidTable Why the Premier League Has Become as Unpredictable as Ever=Ukubuyekezwa Kwangempelasonto Kungani iPremier League isiphenduke ubuhlakani

Although that governing body had authority in the early congregation they acknowledged that their Leader was Jesus=7 Nakuba indikimba ebusayo yayinegunya ebandleni lokuqala yayiqaphela ukuthi uMholi wayo uJesu

No I was not born yesterday=Angizalwanga izolo

Render the listed sentences in Zulu from their original English form.',

'input': 'She looked terrified not understanding',

….Missing "Hint"

'output': 'UJulia wayekhathazekile esaba wayengenakuqonda lutho',

….Missing "Hint"

'Input': 'European national governments also provided subsidies for roadways but typically at a lower level or for shorter periods of time'

Fig. 3. Depicts how sentences are aligned in the prompt when fine tuning using 'Knowledge-Infused Few-Shot Prompt'.

model performs better in a supervised setting, we opted for supervised fine-tuning rather than zero-shot fine-tuning.

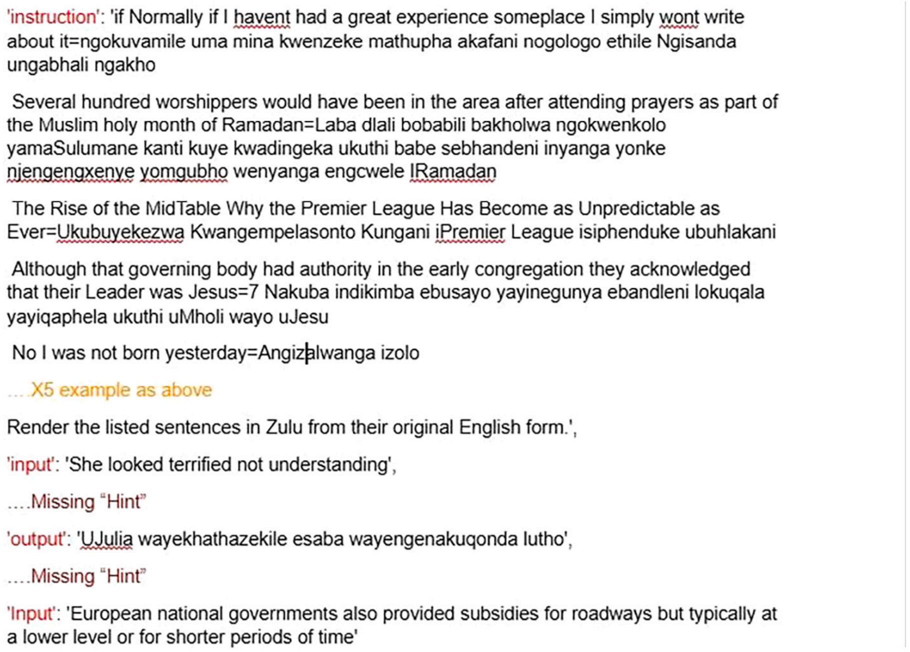
By adhering closely to the principles of our custom prompts, we were able to maintain consistency for a more reliable comparison between the Serengeti model and the Mistral model. The prompt used for this setup is shown in Fig. 6 above.

3.4. Experimental framework

In this section, we outline the models utilized in our experiments and detail the parameters used during fine-tuning. Additionally, we discuss our rationale for selecting specific evaluation metrics, highlighting their respective strengths and limitations. These metrics were chosen to comprehensively assess different aspects of translation quality, showcasing the performance of our fine-tuned LLMs across various tasks and settings.

3.4.1. Mistral model

The Mistral 7B model has proven to be a highly effective alternative to the LLAMA 13B model, demonstrating superior performance in our experimental evaluations. As the primary model for our research, Mistral 7B has been pivotal in validating the efficacy of our custom prompt strategies. These strategies involve a sophisticated ensemble of techniques designed to enhance translation accuracy, particularly



'instruction': 'if Normally if I havent had a great experience someplace I simply wont write about it=ngokuvamile uma mina kwenzeke mathupha akafani nogologo ethile Ngisanda ungabhali ngakho

Several hundred worshippers would have been in the area after attending prayers as part of the Muslim holy month of Ramadan=Laba dlali bobabili bakholwa ngokwenkolo yamaSulumane kanti kuye kwadingeka ukuthi babe sebhandeni inyanga yonke njengengxenye yomgubho wenyanga engcwele IRamadan

The Rise of the MidTable Why the Premier League Has Become as Unpredictable as Ever=Ukubuyekezwa Kwangempelasonto Kungani iPremier League isiphenduke ubuhlakani

Although that governing body had authority in the early congregation they acknowledged that their Leader was Jesus=7 Nakuba indikimba ebusayo yayinegunya ebandleni lokuqala yayiqaphela ukuthi uMholi wayo uJesu

No I was not born yesterday=Angizalwanga izolo

….X5 example as above

Render the listed sentences in Zulu from their original English form.',

'input': 'She looked terrified not understanding',

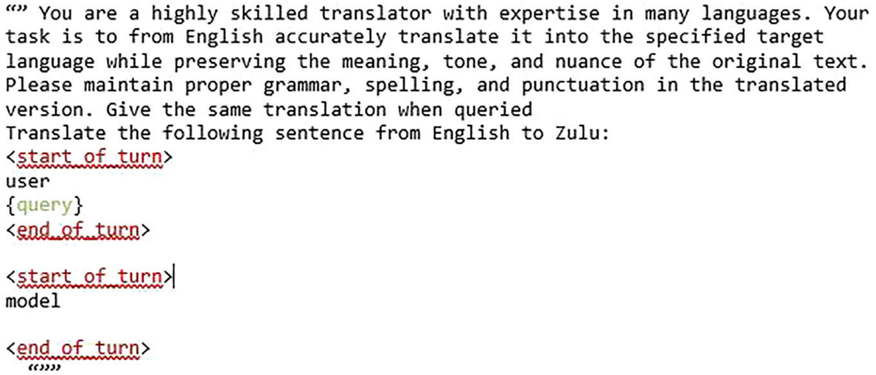
….Missing “Hint”

'output': 'UJulia wayekhathazekile esaba wayengenakuqonda lutho',

….Missing “Hint”

'Input': 'European national governments also provided subsidies for roadways but typically at a lower level or for shorter periods of time'

Fig. 4. Depicts how sentences are aligned in the prompt when fine tuning using ’Knowledge-Infused Long-Shot Prompt’.

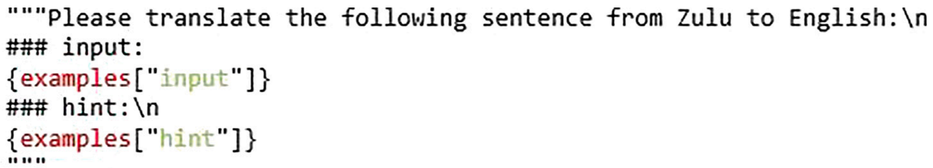


```
“” You are a highly skilled translator with expertise in many languages. Your
task is to from English accurately translate it into the specified target
language while preserving the meaning, tone, and nuance of the original text.
Please maintain proper grammar, spelling, and punctuation in the translated
version. Give the same translation when queried
Translate the following sentence from English to Zulu:
<start_of_turn>
user
{query}
<end_of_turn>

<start_of_turn>
model

<end_of_turn>
  “””
```

Fig. 5. Depicts how sentences are aligned in the prompt when fine tuning using ’Knowledge-Infused Long-Shot Prompt with Dic Hint for Gemma’.



```
"""Please translate the following sentence from Zulu to English:\n
### input:
{examples["input"]}
### hint:\n
{examples["hint"]}
"""
```

Fig. 6. Depicts how sentences are aligned in the prompt when fine tuning using ’Knowledge-Infused Long-Shot Prompt with Dic Hint for Serengeti’.

for under-resourced native South African languages. By implementing these advanced prompts, our goal is to achieve state-of-the-art translation performance, a critical objective given the resource constraints associated with these languages.

To rigorously test the effectiveness of our prompt strategies, we initially conducted a proof of concept phase, running experiments for a maximum of 2,000 steps followed by training the model for two epochs. Fine-tuning large language models (LLMs) for low-resource language translation requires efficient optimization strategies to balance computational costs and performance. Recent research has explored techniques such as low-rank adaptation, quantization, and memory-efficient optimizers like PagedAdamW 8-bit to enhance model training. For instance, studies have demonstrated that fine-tuning LLMs with 8-bit optimization reduces memory overhead while maintaining translation quality (Bawden & Sagot, 2023). Additionally, research on multilingual pre-trained models highlights the importance of batch size and learning rate selection, with configurations like a batch size of 32 and a learning rate of 0.0002 proving effective in adaptation tasks (Johnson et al., 2023). Moreover, prompt-based approaches and parameter-efficient tuning have been explored to adapt LLMs to unseen languages, reinforcing the significance of structured fine-tuning parameters in translation tasks (Wang et al., 2024). These findings align with the growing trend of optimizing LLMs through gradient accumulation steps and controlled warm-up strategies, ensuring reliable adaptation for low-resource languages. Future studies should continue refining these techniques to further enhance translation accuracy and computational efficiency.

In addition to these parameters, we configured the model to employ 4-bit quantization, significantly reducing its memory footprint while maintaining efficiency. The computational data was processed in float16 format to strike a balance between precision and performance. We also enabled double-quotation support to meet specific formatting requirements in our data. These configurations were critical in optimizing both the model's performance and its resource utilization.

Training was conducted on 2 GPUs with 24 GB of memory, ensuring sufficient computational power for handling the extensive requirements of the Mistral 7B model. The choice of SFTTrainer for fine-tuning was driven by its demonstrated effectiveness in supervised learning contexts. As highlighted by Ström Boman (2024), the SFTTrainer framework enhances the fine-tuning process for large language models by optimizing hyperparameters, efficiently managing extensive datasets, and tailoring the training process to specific tasks. This capability makes SFTTrainer particularly well-suited to the complex demands of training models like Mistral. By leveraging the SFTTrainer framework, we are able to achieve superior results with the Mistral model, meeting the rigorous standards required for advanced, state-of-the-art applications. Notably, each fine-tuning experiment with the Mistral model lasted up to 4 GPU days, reflecting the extensive computational resources necessary for achieving optimal performance.

3.4.2. Gemma model

In addition to our work with the Mistral 7B model, we extended our investigation to the Gemma model to further evaluate the effectiveness of our custom prompt strategies across different models. For this purpose, we fine-tuned the Gemma model over 2,000 steps, aiming to demonstrate the adaptability of our prompts in improving translation accuracy for low-resourced languages. In this experiment, we ensured that our dataset was well-aligned with the prompt requirements and allowed the model to tokenize the data during fine-tuning. We utilized the high-quality Umsuka English - isiZulu Parallel Corpus (Mabuya et al., 2021), which, despite being primarily an English-to-Zulu dataset, provided a robust foundation for our evaluation.

The use of this dataset did not pose a significant obstacle. Instead, it underscored the versatility of our prompt engineering strategies, confirming their effectiveness even when applied to a monolithic dataset. The experimental setup for fine-tuning the Gemma model was designed to mirror that of the Mistral model to ensure consistency. We employed a learning rate of 0.0002, set the maximum number of training steps to 2,000, and configured 100 warm-up steps. The optimizer used was PagedAdamW 8 bit, consistent with our previous experiments. The fine-tuning process for the Gemma model was conducted over a period of 3 GPU days, utilizing GPUs with 24 GB of memory.

This experiment with the Gemma model reinforces our confidence in the robustness of our custom prompt strategies and their applicability across various language models and datasets. Our findings align with the work of Team et al. (2024), who introduced the Gemma model and demonstrated its efficacy in multilingual translation scenarios. This reference highlights the Gemma model's capabilities and provides a contextual basis for our results.

3.4.3. Serengeti model

SERENGETI, a groundbreaking multilingual language model, encompasses over 17% of African languages and dialects, demonstrating exceptional performance across various language understanding tasks for African languages (Adebara, 2024). Our research investigates the efficacy of our novel prompt engineering techniques when applied to SERENGETI, particularly within the context of limited datasets. This endeavor contributes to the broader goal of enhancing African language representation in pre-trained multilingual models. By delving into SERENGETI's strengths and exploring avenues for further optimization, we aim to contribute to the development of more inclusive and effective language technologies for African languages. The innovative nature of our prompt design, as exemplified in Table A.11, distinguishes our approach and underscores its potential to improve model performance.

3.5. Chosen evaluation metrics

The introduction of BLEU (Bilingual Evaluation Understudy) by Papineni et al. (2002) marked a significant milestone in the field of machine translation evaluation. BLEU measures the precision of n-grams in the candidate translation with respect to one or more reference translations. Its simplicity and correlation with human judgment made it the de facto standard for many years. However, BLEU's reliance on exact n-gram matches often penalizes translations that are semantically correct but use different words or structures. This limitation is particularly pronounced in languages with rich morphology and idiomatic expressions, such as Zulu and Xhosa. Recognizing the need for improved evaluation metrics, Banerjee and Lavie (2005) introduced METEOR (Metric for Evaluation of Translation with Explicit ORdering). METEOR addresses some of BLEU's shortcomings by incorporating synonymy, stemming, and paraphrase matching. It calculates precision, recall, and an F-score to provide a more balanced evaluation. Despite these enhancements, METEOR still operates primarily on word-level matches and may not fully capture deeper semantic similarities, which are crucial for evaluating translations involving diverse linguistic structures and expressions. In parallel, Lin (2004) developed ROUGE (Recall-Oriented Understudy for Gisting Evaluation), a set of metrics designed for automatic summarization and machine translation. ROUGE measures the overlap of n-grams, longest common subsequences, and skip-bigrams between the candidate and reference translations. While useful for summarizing tasks, ROUGE shares similar limitations with BLEU, focusing on surface-level matches and often missing the semantic equivalence in translations.

The advent of large pre-trained language models has opened new avenues for translation evaluation. G-EVAL (G-Evaluation), introduced by Liu, Iter, et al. (2023), leverages the capabilities of models like GPT-4 to assess translations based on their semantic content. Unlike traditional metrics, G-EVAL uses embeddings to capture the contextual and semantic similarity between hypothesis and reference translations. This approach aligns more closely with human judgment, which tends to focus on meaning rather than exact word matches.

Using G-EVAL for evaluating translations between English, Zulu, and Xhosa provides significant advantages in capturing the semantic richness and contextual nuances of these languages. Traditional metrics like BLEU and METEOR offer valuable insights into surface-level accuracy but often miss deeper semantic equivalence. G-EVAL, by leveraging advanced language model embeddings, ensures a more comprehensive and accurate assessment of translation quality. This holistic approach is essential for developing reliable and culturally aware machine translation systems, especially for low-resource languages where data scarcity and linguistic diversity pose additional challenges.

ChFr++ (Yu et al., 2020), a more recent evaluation metric designed to address some of the limitations of traditional methods, particularly focuses on chunk- and frame-level translations. Unlike BLEU, which is

Table 5
Presents the results of all our experiments using the Mistral model, along with the different prompts tested in each experiment. A detailed analysis of these results, including the significance of the highlighted metrics, will be further discussed in Section 5.

Prompt	Language pair	BLUE	F1-score	G-Eva
Pre-Trained Mistral Model	Eng-Zul	0.005	10.52	46.56%
	Eng-Xh	0.4	14.03	55.17%
Simple Aligned Prompt	Eng-Zul	0.7	8.93	72.9%
	Eng-Xh	0.6	9.29	70.8%
Custom Prompt: ES-ICL D-Hint	Eng-Zul	4.35	27.01	81.69%
	Eng-Xh	4.93	25.45	78.91%
Knowledge-Infused FSP D-Hint	Eng-Zul	5.58	27.73	86.46%
	Eng-Xh	5.07	27.19	85.3%
Knowledge-Infused Few-Shot Prompt	Eng-Zul	5.65	19.9	87.46%
	Eng-Xh	5.04	19.3	88.28%
Knowledge-Infused Long-Shot Prompt	Eng-Zul	18.56	37.44	86.26%
	Eng-Xh	11.69	27.29	70.88%

heavily dependent on n-gram overlap, ChFr++ evaluates larger linguistic units such as grammatical chunks and syntactic frames, making it more suitable for languages with complex syntactic structures. For languages like Zulu and Xhosa, where morphology plays a critical role in sentence structure and meaning, ChFr++ provides a more linguistically grounded evaluation.

4. Results

In this section, we present the results of the various experiments conducted. We begin by highlighting the different prompt engineering strategies tested on the Mistral 7B model over 2000 steps, which ultimately led to the development of our custom prompt engineering approach, subsequently used for extended training. Additionally, we showcase the performance of other pre-existing models when evaluated on the same source and target languages. Furthermore, we present the results of different LLM architectures fine-tuned using the selected custom prompt engineering strategy derived from the 2000-step fine-tuning of the Mistral model.

4.1. Fine-tuning and prompt engineering experiments: Mistral 2000-step optimization results

Leveraging previous research on prompt engineering in the context of low-resource languages as highlighted in Section 3, we experimented with various prompt engineering strategies to find the optimal combination that could enhance the translation performance of large language models (LLMs) between English and two native South African languages: Zulu and Xhosa. We chose the Mistral 7B model, which outperformed the LLAMA 13B model on several downstream tasks as detailed in Section 3. With fewer parameters, the Mistral 7B model delivers performance comparable to larger models while requiring less GPU memory, making it more efficient. To identify the best-performing prompt, we conducted an initial experiment with 2,000 training steps. Table 3 below presents the results of these experiments, illustrating the effectiveness of different prompt engineering strategies on translation performance.

4.2. Final performance of the fine-tuned Mistral model: 31 GPU days of training

Extending the fine-tuning of the Mistral 7B model using our chosen prompt engineering ensemble strategy has yielded strong results, as shown in Table 5. These state-of-the-art outcomes will be further elaborated in Section 6, where we will discuss the extent of improvement in the Mistral 7B model's performance. Additionally, we will compare our results with other popular models available for translating between English and low-resource languages, with a focus on Zulu and Xhosa as our target languages.

4.3. Comparing various metrics on Mistral 7B model

In this section, we discuss the results of various metrics from our experiment fine-tuning the Mistral 7B model, aiming to identify the optimal prompt for improving translation quality in low-resource languages from English. Fig. 7 below presents these results in a bar graph format. The following sub-subsections provide a detailed analysis of the metrics used to evaluate the performance of the Mistral 7B model.

4.3.1. BLEU score evaluation of the Mistral 7B model

As observed, the original pre-trained model struggled significantly to translate between English-Zulu (En-Zul) and English-Xhosa (En-Xh), achieving BLUE scores of just 0.005 and 0.4, respectively. These low scores indicate that the model has difficulty identifying accurate target language words, resulting in many nonsensical translations. Moving forward, we will present only the scores, with the positioning indicating the language pair in question.

Using a simple, commonly adopted prompt in our second experiment yielded slight improvements, raising BLUE scores to 0.7 for En-Zul and 0.6 for En-Xh. Although modest, these gains suggest that the model is beginning to learn from the training data. In subsequent experiments, we saw exponential improvements in BLUE scores across both language pairs.

Notably, the “Knowledge-Infused Few-Shot Prompt” experiment reached BLUE scores of 5.65 (En-Zul) and 5.06 (En-Xh) after 2,000 training steps. However, the “Knowledge-Infused Long-Shot Prompt” experiment, which we prioritized, produced a substantial increase, achieving scores of 18.56 for En-Zul and 11.65 for En-Xh—a jump of over 12 points in En-Zul. This suggests the model has improved significantly in translating from English to Zulu, capturing more accurate words in the target language.

Interestingly, although En-Xh initially had a higher BLUE score than En-Zul with the pre-trained model, this trend reversed in our later experiments. En-Xh's lower BLUE score in the penultimate experiment (7 points below En-Zul) could be due to lower-quality translations in our training data for Xhosa, which may contain more English words mixed into Xhosa sentences compared to Zulu.

This prompted us to choose the “Knowledge-Infused Long-Shot Prompt” for fine-tuning the Mistral 7B model, aiming to improve translation performance further. However, we observed only a moderate improvement, reaching BLUE scores of 20 for En-Zul and 14 for En-Xh. This slight difference, although positive, suggests that overfitting might be occurring, possibly due to repetitive exposure to similar words.

While our custom prompt configuration significantly enhances translation, other human-aligned evaluation metrics, such as F1-score and G-Eva, show consistent improvement. Prior mentioned metric are more aligned with human translation than BLUE score metric as we have discussed in Section 3.

4.3.2. F1-score evaluation of the Mistral 7B model

The F1-score, which evaluates the overlap of n-grams between hypothesis and reference sentences, is a more human-aligned metric for translation evaluation. As shown in Fig. 7, the pre-trained model achieved F1-scores of 10.42 and 14.43 for En-Zul and En-Xh, respectively. However, with the simple prompt, the F1-scores decreased to 8.93 for En-Zul and 9.29 for En-Xh. This drop could be due to the adjustments in model weights and the limited training steps (2000), suggesting that while the model was adapting to new languages, the simple prompt was less effective in retaining performance.

In contrast, the “Custom Prompt: ES-ICL D-Hint” and “Knowledge-Infused FSP D-Hint” prompts, which include a dictionary-based hint for English to Zulu and Xhosa, showed steady improvements, achieving average F1-scores around 27. This suggests that the dictionary hints significantly enhance the model's translation accuracy in low-resource settings, as it correctly predicts more n-grams between the ground truth and the predicted sentences.

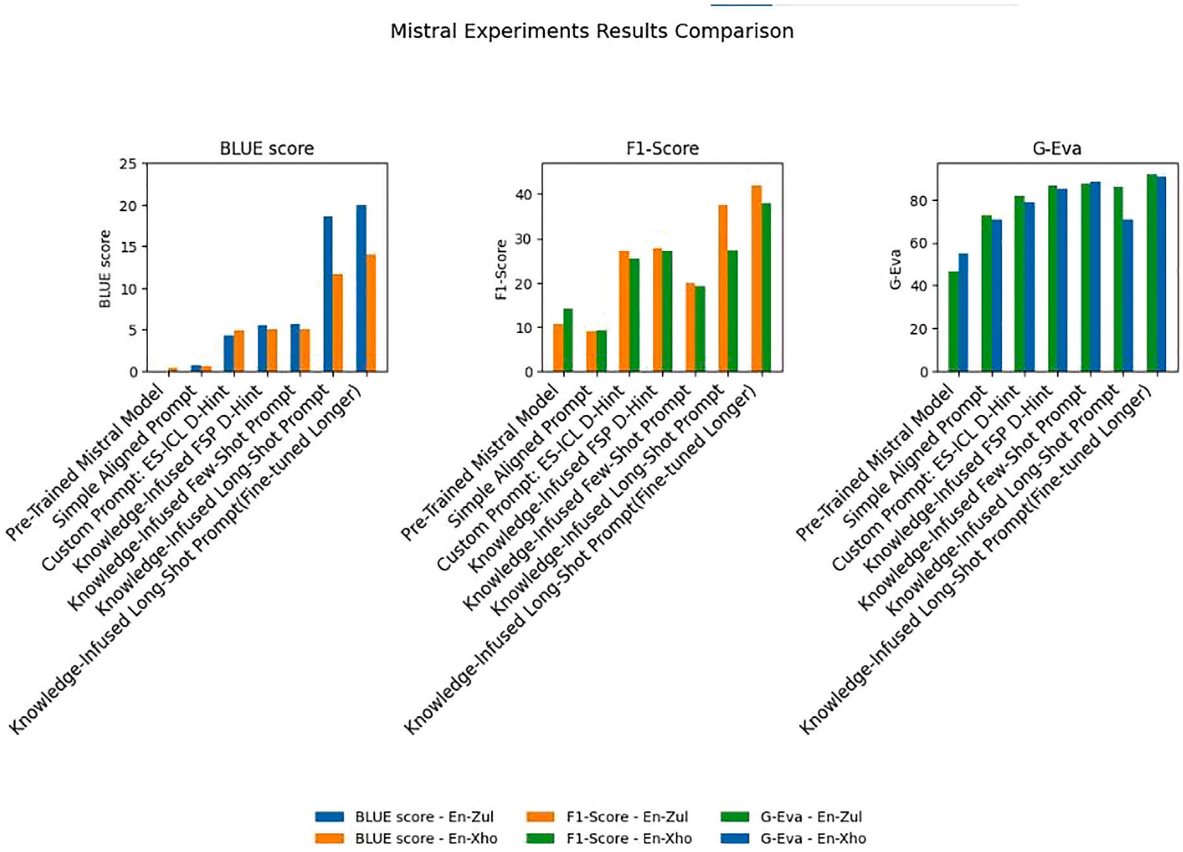


Fig. 7. Show experimental results using BLUE, F-1 and G-Eva score, experiments were conducted on Mistral 7B model.

When we removed the dictionary component in the “Knowledge-Infused Few-Shot Prompt” experiment, F1-scores dropped to 19.9 for En-Zul and 19.3 for En-Xh. This decrease emphasizes the importance of dictionary hints in improving translation quality. However, the “Knowledge-Infused Few-Shot Prompt” still outperformed the simple prompt, suggesting that exposure to more examples in the corpus can greatly aid in training the Mistral 7B model.

With additional training steps using the "Knowledge-Infused Long-Shot Prompt”, F1-scores rose significantly to 42 for En-Zul and 38 for En-Xh. These scores indicate that the model achieves a higher level of translation accuracy between these language pairs than suggested by the BLUE metric alone. Table A.11 includes examples of ground truth and predicted sentences that illustrate these F1-score improvements.

To further evaluate the Mistral 7B model’s capability to capture context and semantics accurately, we also incorporated the G-Eva score. This metric, calculated as the mean cosine similarity between sentence embeddings, provides insight into the alignment of contextual and semantic content in translations.

4.3.3. G-Eva score evaluation of the Mistral 7B model

The G-Eva score, calculated as the mean cosine similarity between sentence embeddings, provides a measure of how well the model captures context and semantics in the target sentence compared to the reference. The pre-trained model achieved G-Eva scores of 46.56% for En-Zul and 55.17% for En-Xh, suggesting that it somewhat captures sentence context and semantics, likely due to the reasoning abilities typical of large language models like Mistral 7B.

When using a simple prompt, the G-Eva scores increased significantly to 72.9% for En-Zul and 70.8% for En-Xh. This improvement indicates that while the model may not yet perform exact word-for-word translations, it now more accurately preserves the context and semantics in translations from the source to the target language.

With the “Custom Prompt: ES-ICL D-Hint” and “Knowledge-Infused FSP D-Hint”, both of which incorporate dictionary hints but utilize a reduced training dataset, G-Eva scores averaged around 85% and 82% across the language pairs. Notably, the ”Knowledge-Infused Few-Shot Prompt" outperformed these with scores of 87.46% for En-Zul and 88.86% for En-Xh. Interestingly, En-Xh performed better than En-Zul in this setup, contrary to previous experiments. This variation invites further investigation, as previous results generally favored En-Zul.

The “Knowledge-Infused Long-Shot Prompt” experiment, however, resulted in scores of 86.26% for En-Zul and 70.88% for En-Xh. This notable drop for En-Xh suggests some variability that warrants further analysis, particularly to understand why En-Xh, which performed well in earlier setups, has a comparatively low score here. Additionally, En-Zul showed a slight decrease.

In our final experiment, extending the training with the “Knowledge-Infused Long-Shot Prompt” led to significant gains, with G-Eva scores of 92% for En-Zul and 91% for En-Xh. These results show that the model effectively captures context and semantics, producing translations that convey the intended meaning even if the exact words differ from the reference. This capability is illustrated in Table A.11 with examples of predicted sentences that, even without exact word matches, maintain the sentence’s intended meaning.

Observing the progression of BLUE scores from top to bottom in Table 5 and comparing F1-scores in Table 6, we see a steady growth in translation quality. The G-Eva results show that the model begins with relatively high scores, likely due to the inherent reasoning capacity of the Mistral 7B model, allowing it to produce contextually and

Table 6
Shows the results obtained using the best-performing prompt, the Knowledge-Infused Long-Shot Prompt. We fine-tuned the Mistral model for a total of 31 GPU days. A more detailed analysis of these results will be provided in Section 5.

Prompt	Language pair	BLUE	F1-score	G-Eva
Knowledge-Infused Long-Shot Prompt	Eng-Zul	20	42	92%
	Eng-Xh	14	38	91%

Table 7
Presents the results of the NLLB model across different metrics on the Flores-101 dataset. As shown, the model performs exceptionally well. It is important to note that we did not fine-tune the model; the results reflect the performance of the pre-trained version used as-is.

Language pair	BLUE	F1-score	G-Eva
Eng-Zul	31.45	52.32	95.40%
Eng-Xh	25.72	49.73	95.39%

semantically accurate translations. This trend suggests that the model's ability to capture meaning is stronger than what the BLUE and F1-score metrics might reflect.

4.4. Benchmarking Mistral against popular English-to-Zulu and English-to-Xhosa translation models

In this subsection, we present the results of popular pre-existing translation models and compare them to the performance of the Mistral 7B model fine-tuned as described in the previous subsections. This comparison highlights the effectiveness of our custom training approach and demonstrates how Mistral 7B performs relative to widely used models in similar translation tasks.

4.4.1. NLLB results

The NLLB (Costa-jussá et al., 2022) model demonstrated strong performance in our evaluation of low-resource languages. These findings will inform a comparative analysis with our final experiment, in which the Mistral model was fine-tuned over 30 GPU days. The NLLB model was specifically designed to address the under representation of low-resource languages in the rapidly advancing field of LLMs, which have predominantly excelled in high-resource languages. Its aim is to ensure that low-resource languages also benefit from these developments in NLP tasks. As presented in Table 7, the NLLB model achieved high BLEU and cosine similarity scores, but it produced suboptimal results in terms of the F1 score when evaluated on the Umsuka dataset (Mabuya et al., 2021) for the English-to-Zulu translation task. A more detailed analysis of this performance will be provided in the subsequent section.

4.4.2. Google translate API

The Google Translate API (Tsai, 2019), a widely used free service that facilitates translation between numerous languages for millions of users globally, was included as one of our benchmark models. Its inclusion enables us to assess its performance in relation to the results of our final experiments, as discussed in Section 4.1. Moreover, we underscore the effectiveness of the prompt engineering strategies introduced in that section, which have successfully enhanced translation quality for the low-resource languages under investigation. The API demonstrates strong performance in terms of cosine similarity, yields reasonable results for the F1 score, and achieves a respectable BLEU score.

5. The influence of custom prompts on English-African language translation

In this section, we emphasize the impact of prompt engineering on fine-tuning and its role in enhancing LLM performance for translation tasks, particularly for low-resource languages, with a focus on African languages. We further analyze the results, demonstrating how different prompts influenced the outcomes. Additionally, we highlight the effectiveness of our custom prompt in improving translation quality across various LLM architectures in bilingual settings, utilizing a high-quality dataset.

5.1. Power of prompt engineering to enhance translation for LLMs in low-resource settings

In this subsection, we delve deeper into our results, analyzing how different prompts influence the outcomes across the various evaluation metrics we have selected. We explore the specific ways in which prompt variations impact model performance, shedding light on their effectiveness in different evaluation contexts.

5.1.1. Fine-tuning and prompt engineering experiments: analysis

In this sub-subsection, we delve deeper into the impact of various prompts on our results, examining how specific prompt formulations contributed to fluctuations in high scores across different experiments.

5.1.1.1. Custom prompt: Equal sign ICL dictionary hint. Selecting the appropriate prompt was critical for successfully extending the fine-tuning process of the Mistral 7B model. Drawing on previous research (Balne et al., 2024; Bawden & Yvon, 2023; Bertsch et al., 2024; Song et al., 2023; Zhang, Haddow, & Birch, 2023; Zhu et al., 2023) that focuses on enhancing translation quality in large language models (LLMs) for low-resource languages, we experimented with several ensemble approaches to identify the optimal configuration. The model was evaluated using three complementary metrics: BLEU, ChrF++, and G-Eval, each offering unique insights into the relationship between reference and hypothesis sentences.

Initially, we fine-tuned Mistral 7B using a basic prompt (an example is provided in Table A.11). As illustrated in Table 4, for both the En-Zulu and Xh-English translation pairs, the model achieved a BLEU score of less than 1, indicating significant difficulty in reproducing the correct lexical content of the target sentences. The model also registered a low ChrF++ F1-score, reflecting its struggles with syntactic structure and sentence formation. Interestingly, the G-Eval score was relatively high, which can be attributed to the presence of untranslated English words in the test dataset. This resulted in inflated semantic similarity between the source and target embeddings, despite the lexical and grammatical discrepancies.

5.1.1.2. Simple aligned prompt. Our first custom ensemble prompt yielded significant improvements, as shown in Table 4. Compared to the previous prompt, the scores across all evaluation metrics increased substantially. The structure of the new prompt, detailed in Table A.11, played a key role in this enhanced performance.

Notably, the BLEU score increased by over 3 points, a considerable improvement given only 2,000 additional training steps. This indicates that the model is producing a greater number of words that match the ground truth, reflecting more accurate translations. In terms of the F1-score (ChrF++), we observed a rise of more than 20 points, suggesting that the sentence structures generated by the model are becoming increasingly aligned with those of the reference sentences. This indicates that larger chunks of the model's output are now syntactically similar to the target sentences. Finally, the G-Eval score showed an average increase of 8% for both Zulu and Xhosa outputs, with the model now achieving over 75% similarity on this metric. While the model may not be producing word-for-word translations, this improvement suggests that it is capturing the underlying semantics and context more accurately in many cases.

Table 8
Shows that while the Google Translate API struggles to produce an exact match with the ground truth, as reflected in its lower BLEU score, it performs well on the ChrF++ and G-Eva metrics. This indicates that, although it may not replicate the exact wording, it captures the semantics and context of the translated sentences accurately, with minimal hallucination.

Language pair	BLUE	F1-score	G-Eva
Eng-Zul	5.93	55.12	95.6%
Eng-Xh	3.48	52.83	95.48%

5.1.1.3. Knowledge-infused few-shot prompt with dictionary hint. This prompt outperforms the one previously discussed across all metrics, even if the gains are modest in some cases. These improvements, though small, are crucial in identifying the most effective fine-tuning prompt for the model. The overall performance of all metrics increased when compared to the earlier prompt. Specifically, the BLEU score improved by an average of 1.4% for both Zulu and Xhosa, indicating that more sentences are now producing word choices that align with the ground truth. Similarly, the F1-score and G-Eval metrics also show consistent gains, confirming that this prompt yields better results than the one discussed in the prior section. These enhancements suggest the model's outputs are both structurally and semantically closer to the reference translations.

5.1.1.4. Knowledge-infused few-shot prompt. As highlighted in Section 3, this prompt operates by requiring the model to generate output without being provided with a direct example. For further details on the structure and design of this prompt, please refer to Section 3, as it has influenced our results in a distinct way compared to the prompts discussed throughout Section 5.1.1. A detailed example of the prompt can be found in Table A.11. Compared to the prompt in Section 5.1.1.4 (hereafter referred to as the "previous section"), this prompt yields a higher BLEU score. However, a decline in the F1-score was observed for both the En-Zulu and En-Xhosa translation pairs. This decrease can be attributed to two factors: first, as noted earlier in this section, the prompt's design omits the dictionary hint component that was included in the previous prompt. Second, the model's fine-tuning in a semi-supervised manner may have affected the syntactic accuracy of the predicted sentences when compared to the ground truth. On the other hand, the G-Eval similarity score increased, indicating that the model produced outputs that are semantically and contextually closer to the reference translations, even if the sentence structure was less accurate.

5.1.1.5. Knowledge-infused long-shot prompt. Our final experiment yielded remarkable results. The development of the prompt has already been discussed in Section 3, with a sample available in Appendix A. As seen in Table 4, all prior prompts averaged around 5 BLEU scores, but this final experiment saw a significant increase, achieving over 18 BLEU for En-Zulu and 11 BLEU for En-Xh. This dramatic improvement highlights the importance of incorporating more examples into the instruction section of the prompt, coupled with in-context learning (ICL). In terms of the F1-score, the best previous scores (excluding this experiment) were 27.73 for En-Zulu and 27.19 for En-Xh, discussed in Section 5.1.1.3 and 5.1.1.4. In those sections, we noted a reduction in F1-scores to 19.9 and 19.3, respectively. However, the prompt from this experiment boosted the F1-score to 37.44 for En-Zulu and 27.44 for En-Xh. This significant increase, particularly for the En-Zulu pair, indicates that the predicted sentence structures align more closely with the ground truth. Although the G-Eva similarity score decreased slightly—by 1% for En-Zulu and 3% for En-Xh these differences are minimal. We therefore concluded that this prompt would be the best candidate for further fine-tuning of the Mistral 7B model to enhance its translation accuracy from English to Zulu and Xhosa. Interestingly, when comparing this experiment to the pre-trained model results (0.005 BLEU for En-Zul and 0.4 for En-Xh), it is clear that En-Xh initially performed better across the evaluation metrics. However, in this final experiment, En-Zulu outperformed En-Xh, likely due to the larger portion of En-Zulu data in our dataset. This demonstrates that our custom ensemble prompt strategy effectively improves translation accuracy for low-resource languages.

Table 9
Showcases the performance of three models with distinct architectures. Notably, the Gemma model shares architectural similarities with the Mistral 7B model, while Serengeti is derived from a BERT-based architecture that we fine-tuned for our specific tasks. The results clearly demonstrate that the Gemma model outperforms both Mistral 7B and Serengeti across all evaluated metrics..

Model	Language pair	BLEU	F1-Score	G-Eva
Serengeti	Eng-Zul	1.2	21.12	21.12%
	Eng-Xh	4.97	24.84	97.62%
Mistral	Eng-Zul	4.46	26.19	73.23%
	Eng-Xh	0.4	14.03	55.17%
Gemma	Eng-Zul	4.97	24.84	97.62%
	Eng-Xh	0.4	14.03	55.17%

5.2. Bench-marking Mistral against popular translation models

We present a comparative analysis of our fine-tuned Mistral 7B model, trained over 30 GPU days, with the NLLB and Google Translator API, both of which leveraged substantially larger English-Zulu (En-Zul) and English-Xhosa (En-Xh) datasets. Despite the disparity in training data, our model demonstrates competitive performance.

On the FLORES-101 general domain dataset, NLLB achieves an average BLEU score of 31.45, an F1 score of 52.32, and a mean cosine similarity of 95.4% for En-Zul. For En-Xh, NLLB scores 25.72 (BLEU), 49.73 (F1), and 95.39% (cosine similarity). In contrast, Google Translator API produces a BLEU score of 5.93, an F1 score of 55.12, and a cosine similarity of 95.66%. NLLB's stronger performance, particularly in BLEU and F1 metrics, can be attributed to its designation of Xhosa and Zulu as "high-resource" languages, allowing for the collection of more extensive datasets during training. This confers a significant advantage over our Mistral 7B model. However, the difference in cosine similarity is relatively modest, with only a 3% difference for En-Zul and a 4% difference for En-Xh, indicating that our model may perform better than the BLEU and F1 scores alone suggest—especially given its more limited training data.

Notably, our Mistral 7B model outperforms the Google Translator API in BLEU scores, with improvements exceeding 15 points for En-Zul and 9 points for En-Xh. However, Google Translator API surpasses our model in both F1 and cosine similarity scores. The F1 score differences (around 10 points) and cosine similarity differences (approximately 3%) imply that Google Translator API captures sentence-level semantics and context more effectively. Nonetheless, the superior BLEU scores of our model highlight the potential of our custom ensemble prompting technique to substantially improve translation quality for low-resource languages. In conclusion, while NLLB benefits from access to larger datasets, our Mistral 7B model performs robustly given its more constrained training environment. Furthermore, it surpasses Google Translator API in key translation metrics for low-resource languages, making a strong case for the efficacy of our custom approach in improving translation for such languages (see Table 9). .

5.3. Comparative analysis: Fine-tuned LLMs on a high-quality Zulu-English dataset

Details about the models can be found in Section 3.5. Due to the limited size of our dataset, as shown in Table 2, we trained the

Table 10
Comparison of Mistral's performance against other models analyzed in this paper.

Model	Language	BLUE Score	F1-Score	G-Eva
Mistral	En-Zul	20	42	92%
Google Translate Api	En-Zul	5.93	55.12	95.6%
Gemma	En-Zul	4.97	24.84	97.62%
Serengeti	En-Zul	1.2	21.12	21.12%
NLLB	En-Zul	31.45	52.32	95.40%

models for only two epochs. The datasets used include the Umsuka English-Zulu Dataset (Mabuya et al., 2021) and the Multilingual Statistical Terminology Project by Statistics South Africa (Marivate et al., 2024). These datasets allowed us to fully explore the potential of the "Knowledge-Infused Long-Shot Prompt" by incorporating dictionary data, as demonstrated in our "Knowledge-Infused Few-Shot Prompt with Dictionary Hint" custom ensemble prompts. As depicted in Table 9, the Serengeti model performs poorly, with a BLEU score of 1.2 and an F1 score of 21.12, indicating that it struggles to produce accurate word-for-word translations when compared to the ground truth. In contrast, the Mistral 7B model delivers more promising results with a BLEU score of 4.46 and an F1 score of 24.46. This suggests that, despite limited training, Mistral 7B learned to structure sentences more accurately and produce words closer to the ground truth. Finally, the Gemma model outperforms both Serengeti and Mistral 7B, achieving a BLEU score of 4.46 and an F1 score of 26.19. This demonstrates that, even with a limited dataset, if the data is of high quality, the model can learn more efficiently. The success of the Gemma model further validates the effectiveness of our custom ensemble prompts, which accelerate learning even in bilingual datasets. The model can predict more words about the ground truth and correctly capture sentence structures, proving that our custom prompts are not restricted to specific models, multilingual settings, or architectural designs. We remain cautious in interpreting the cosine similarity scores for this model, as they do not align well with other evaluation metrics discussed here. Both Gemma and Serengeti achieve nearly 100% cosine similarity, despite their low BLEU and F1 scores. This suggests that the embeddings may be capturing the wrong sentence meanings or that the sub-word tokenizers might be incorrectly separating tokens, leading to artificially inflated cosine similarity scores. Therefore, we consider these scores to be insignificant for analyzing the models' performance in this context.

5.4. Mistral vs other LLMs for English to Zulu translation

As shown in Table 10, our Mistral 7B model demonstrates a mixed performance across the evaluated metrics when compared to other models. While the NLLB model outperforms Mistral 7B on most evaluation metrics, our Mistral 7B model achieves a higher BLEU score compared to Google Translate API, Gemma, and Serengeti models. However, it is outperformed by the NLLB model on the BLEU score, as well as on the F1-score and G-Eval metrics, which are dominated by Google Translate API and NLLB.

The BLEU score results indicate that our Mistral 7B model aligns words more accurately when translating from English to Zulu compared to Google Translate API, Gemma, and Serengeti models. On average, Mistral 7B achieves a BLEU score that is 14 points higher than these models, though it falls short of the NLLB model, which outperforms Mistral 7B by 11 BLEU points. This suggests that Mistral 7B is effective in generating translations with better word alignment, despite being outperformed by NLLB.

In terms of the F1-score, our Mistral 7B model outperforms only the Gemma and Serengeti models. This is likely due to the fact that Gemma and Serengeti are classification models that were fine-tuned for translation tasks, as discussed in earlier sections. However, Mistral 7B struggles to match the performance of Google Translate API and NLLB, which may be attributed to their training on larger, cleaner, and more accurate datasets, as well as their more extensive parameter optimization. On average, Google Translate API and NLLB achieve an F1-score that is 10 points higher than Mistral 7B, indicating that our model may face challenges in accurately translating certain English words into Zulu.

Similarly, on the G-Eval metric, Mistral 7B outperforms only the Serengeti model. This suggests that the translation embeddings generated by Mistral 7B are less accurate when compared to the ground truth embeddings, potentially due to some English words remaining untranslated or being preserved as-is in the output. This highlights areas for improvement in the model's ability to fully capture the nuances of English-to-Zulu translation.

In summary, while our Mistral 7B model shows competitive performance in terms of BLEU score, it faces challenges in matching the F1-score and G-Eval performance of more robust models like Google Translate API and NLLB. These results underscore the potential of Mistral 7B for translation tasks but also highlight the need for further refinement, particularly in handling complex translations and improving embedding accuracy.

5.5. Future and practical applications

In future work, we aim to refine the dataset quality for fine-tuning the model, as the current training data contains some English words embedded within Xhosa and Zulu sentences. By securing a more accurate and cleaned dataset, we hope to improve translation precision. Additionally, we plan to extend training time for the Mistral 7B model and experiment with a broader range of prompt strategies. Investigating how different components of our custom prompts impact translation quality – by selectively omitting parts – may reveal which elements are essential for enhancing translation, particularly in low-resource language settings. Furthermore, exploring how high-quality datasets could benefit different architectures, such as the Serengeti model, may offer insights into achieving superior translation quality for these language pairs.

We also plan to fine-tune the tokenizer of the LLM to better accommodate African languages, expanding its vocabulary to enhance translation accuracy for low-resource languages.

LLMs, like ChatGPT-4, have demonstrated exceptional capabilities in reasoning and cross-domain knowledge, revolutionizing access to information. Similarly, models like Mistral 7B open doors to further research on LLM applications. If these models can understand and translate African languages accurately, they hold transformative potential for education, economic development, and knowledge accessibility.

For example, accurate African language models could improve educational outcomes by allowing students to engage with content in their native languages, aiding comprehension in subjects like science. Mathematics, for instance, could be taught through a language-focused application, supporting students from primary school to university. This would reduce the cognitive load of translating academic content from a secondary language, breaking down language barriers in the education system.

Moreover, small-scale farmers could benefit from receiving climate and agricultural information in their own languages, helping them adapt to changing conditions and reducing poverty. Ultimately, accurate translations into African languages could enable people to learn new skills and access valuable knowledge, fostering economic empowerment and social advancement across the continent.

6. Conclusion

We have demonstrated that the key to achieving high performance in LLMs for translation tasks lies in prompt engineering. Through a series of experiments, we identified an effective prompt template that significantly enhances translation performance for LLMs, specifically using the Mistral 7B model. Our results show that the Mistral 7B model, fine-tuned with our custom prompt template, delivers competitive performance compared to leading translation models.

Additionally, we demonstrated that this prompt template can be successfully applied to other models, such as Serengeti and Gemma, to enhance translation accuracy. Our results show that the effectiveness of the template is not restricted by model architecture or multilingual settings. It performs consistently across various architectures, including BERT-based models, and in bilingual contexts, underscoring its versatility in improving translation accuracy regardless of the underlying model or language configuration.

In the future, we aim to acquire higher-quality datasets, expand the embedding layer of the Mistral model, and develop a new tokenizer capable of handling subword units for languages like Zulu and Xhosa. Additionally, we plan to investigate hyperparameters more thoroughly to optimize translation performance for LLMs and determine the optimal parameters for fine-tuning. Further research is necessary to understand Gemma's exceptional performance and explore ways to enhance Mistral 7B and Serengeti for superior results. We also intend to experiment further with improving translation between Zulu and Xhosa bidirectionally, advancing towards a many-to-many translation model that incorporates all official written languages of South Africa.

Finally, we intend to fine-tune Mistral for a longer period using additional GPU resources, and gain deeper insights into which weights within an LLM influence translation tasks. This could help us optimize models more effectively to improve translation performance.

CRediT authorship contribution statement

Pitso Walter Khoboko: Writing paper, Conducting research, and Performing experiments. **Vukosi Marivate:** Primary supervisor, Providing guidance in identifying and refining the research problem, Reviewing and improving the article, offering feedback to enhance its clarity and coherence for readers. **Joseph Sefara:** Technical challenges during the research process. When the experimental results were not aligning, [Second Supervisor] suggested improvements to the code and experimental setup. Additionally, [Second Supervisor] ensured steady progress through fortnightly meetings, during which feedback and direction were provided.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pitso Walter Khoboko reports article publishing charges was provided by University of Pretoria. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Highlights various prompt engineering strategies that were explored. It showcases the custom prompts that were experimented with, leading to the final design. This final prompt demonstrated that prompt engineering plays a pivotal role in enhancing the transactional capabilities of LLMs, particularly in low-resource settings (see Table A.11).

Table A.11
The table presents the various prompt designs used in our experiments.

Language pair	Number of definitions
Custom Prompt: Equal Sign ICL Dictionary Hint	*Instruction: If [SRC] to [TRG] translation for [random1_source_sentence] is [random1_target_sentence] following that [random_instruction_ sentence (e.g Please provide the [TGT] translation for the following sentences)]* *Input:[random2_source_sentence]* *Output:[random2_target_sentence]*
Custom Prompt: Equal Sign ICL Dictionary Hint	*Instruction: If [SRC] to [TRG] translation for [random1_source_sentence] = [random1_target_sentence] following that [random_instruction_sentence] Input [random2_source_sentence] Output:[random2_target_sentence]* *Input: [random2_source_sentence]* *Hint: In this context, [word_from_ random2_source_sentence] means [Translated word to target language]* *Output: [random2_target_sentence]* *Instruction: Instruction: If [SRC] to [TRG] is [random3_source_sentence]=[random3_target_sentence] following that [random_instruction_sentence (e.g Please provide the [TGT] translation for the following sentences)]* *Input:[random4_source_sentence]* *Hint: In this context, [word_from_ random4_source_sentence] means [Translated word to target language]* *Output:*
Knowledge-Infused Few-Shot Prompt with Dictionary Hint	*Instruction: If [SRC] to [TRG] is [random1_source_sentence]=[random1_target_sentence] [random2_source_sentence]=[random2_target_sentence](random up to 1 pairs) following that [random_instruction_ sentence (e.g Please provide the [TGT] translation for the following sentences)* *Input: [random2_source_sentence]* *Hint: In this context, [word_from_ random2_source_sentence] means [Translated word to target language]* *Output: [random2_target_sentence]* *Instruction: Instruction: If [SRC] to [TRG] is [random3_source_sentence]=[random3_target_sentence] following that [random_instruction_sentence (e.g Please provide the [TGT] translation for the following sentences)]* *Input:[random4_source_sentence]* *Hint: In this context, [word_from_ random4_source_sentence] means [Translated word to target language]* *Output:*

(*continued on next page*)

Table A.11 (*continued*).

Language pair	Number of definitions
Knowledge-Infused Few-Shot Prompt	*Instruction: If [SRC] to [TRG] is [random1_source_sentence]=[random1_target_sentence] [random2_source_sentence]=[random2_target_sentence](random up to 5 pairs) ...... following that [random_instruction_ sentence (e.g Please provide the [TGT] translation for the following sentences)]* *Input:[random3_source_sentence]* *Output:[random3_target_sentence]* *Output: [random2_target_sentence]* *Instruction: Instruction: If [SRC] to [TRG] is [random4_source_sentence]=[random4_target_sentence]* *Instruction: Instruction: If [SRC] to [TRG] is [random4_source_sentence]=[random4_target_sentence] [random5_source_sentence]=[random_target_sentence] ... following that [random_instruction_ sentence (e.g Please provide the [TGT] translation for the following sentences)]* *Input:[random_source_sentence]* *Output:*
Knowledge-Infused Long-Shot Prompt	*Instruction: If [SRC] to [TRG] is [random1_source_sentence]=[random1_target_sentence] [random2_source_sentence]=[random2_target_sentence](random up to 10 pairs) ...... following that [random_instruction_ sentence (e.g Please provide the [TGT] translation for the following sentences)]* *Input:[random3_source_sentence]* *Output:[random3_target_sentence]* *Output: [random2_target_sentence]* *Instruction: Instruction: If [SRC] to [TRG] is [random4_source_sentence]=[random4_target_sentence]* *Instruction: Instruction: If [SRC] to [TRG] is [random4_source_sentence]=[random4_target_sentence] [random5_source_sentence]=[random5_target_sentence] ... following that [random_instruction_ sentence (e.g Please provide the [TGT] translation for the following sentences)]* *Input:[random_source_sentence]* *Output:*

Appendix B

"The tables below present randomly selected sentences" G-Eva and BLEU scores using the FLORES101 test dataset. While the BLEU scores are relatively low, the mean cosine similarity (G-Eva) between the reference and hypothesis sentences is consistently high. Upon reviewing the sentences, it is evident that G-Eva, being a more human-aligned evaluation metric, clearly demonstrates that our model performs better than the BLEU score suggests in all instances (see Table B.12) .

Data availability

Data will be made available on request.

Table B.12
The table shows the randomly selected ground truth and model-predicted sentences used for evaluating both BLEU and G-Eva metrics.

Sentences	BLUE score	G-Eva (cosine similarity score)
BluerefEnZulu: Ijas iaeroplane yaphahlazeka yawa endaweni yokupaka izindiza ezindizayo cishe ngezithuba zabo eziyi UTC futhi yashiswa yavala isikhumulo sezindiza ukuze kuqhutshwe izindiza zezentengiselwano *BluehypZul: IJAS 39C Gripen yashayeka emgwaqweni okugijima kuwo indiza cishe ngawo 9 30 ekuseni ngekhathi sendawo 0230 UTC yaqhuma okwavala izindiza zabantu bonke esikhumulweni sezindiza*	*0.17*	*0.70*
BluerefEnXhosa: ia jas ia gripen yaphahlazeka kwisitalato nge ngezithuba zabo ezilishumi elinesithathu ebusuku i UTC kwaye yatshisa ivala isikhululo seenqwelomoya *BluehypXh: IJAS 39C Gripen yawa endleleni yeenqwelo moya pha ngentsimbi ye 930 kusasa kwixesha lasekuhlalenie 0230UTC yaze yadubula isikhululo seenqwelo ntaka savalwa ekubeni kufikele kuso iinqwelo ntaka njengesiqhelo*	*0.04*	*0.65*
Eng: The JAS 39C Gripen crashed onto a runway at around 930 am local time 0230 UTC and exploded closing the airport to commercial flights		
BluerefEnZulu: Umshayeli wendiza wabonwa njengomkhokheli wesquadron uDilokrit Pattavee *BluehypZul: Umshayeli wabhanoyi wahlonzwa njengoMholi weSquadron uDilokrit Pattavee*	*0.43*	*0.93*
BluerefEnXhosa: "Umphathi weqela uDilokrit Pattavee", *BluehypXh: Umqhubi ufumaneke engu Squandron Leader Dilokrit Pattavee*	*0.14*	*0.66*
Eng: The pilot was identified as Squadron Leader Dilokrit Pattavee		
BluerefEnZulu: Imithombo yezokuxhumana ithi imoto yezokuphepha yaseMpangeni yashayisana nabanye abezimo eziphuthumayo ngesikhathi bephendula esigamekweni *BluehypZul: Abezindaba bendawo babika ukuthi imoto yomlilo yasesikhumulweni sezindiza yaginqikangenkathi iphuthuma*	*0.074*	*0.76*
BluerefEnXhosa: "Umphathi weqela uDilokrit Pattavee", *BluehypXh: Umqhubi ufumaneke engu Squandron Leader Dilokrit Pattavee*	*0*	*0.73*
Eng: The pilot was identified as Squadron Leader Dilokrit Pattavee		

References

Adamou, E., Breu, W., Scholze, L., & Shen, R. (2016). Borrowing and contact intensity: A corpus-driven approach from four slavic minority languages. *Journal of Language Contact*, *9*(3), 513–542.

Adebara, I. (2024). *Towards afrocentric natural language processing* (Ph.D. thesis), University of British Columbia.

Andersland, M. (2024). Amharic LLaMA and LLaVA: Multimodal LLMs for low-resource languages. arXiv preprint arXiv:2403.06354.

Balne, C., Bhaduri, S., Roy, T., Jain, V., & Chadha, A. (2024). Parameter efficient fine tuning: A comprehensive analysis across applications. arXiv preprint arXiv:2404.13506.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Bawden, R., & Sagot, B. (2023). RoCS-MT: Robustness challenge set for machine translation. *10*, In *WMT23-Eighth conference on machine translation* (pp. 198–216).

Bawden, R., & Yvon, F. (2023). Investigating the translation performance of a large multilingual language model: the case of bloom. arXiv preprint arXiv:2303.01911.

Bertsch, A., Ivgi, M., Alon, U., Berant, J., Gormley, M., & Neubig, G. (2024). In-context learning with long-context models: An in-depth exploration. arXiv preprint arXiv:2405.00200.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Costa-jussá, M., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., et al. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, *36*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., et al. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Ghafoor, A., Imran, A., Daudpota, S., Kastrati, Z., Batra, R., & Wani, M. (2021). The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, *9*, 124478–124490.

Ghazvininejad, M., Gonen, H., & Zettlemoyer, L. (2023). Dictionary-based phrase-level prompting of large language models for machine translation. arXiv preprint arXiv:2302.07856.

Goyal, N., Gao, C., Chaudhary, V., et al. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, *10*, 522–538.

Groenewald, H., & Fourie, W. (2009). Introducing the autshumato integrated translation environment. In *Proceedings of the 13th annual conference of the European association for machine translation*.

Hayou, S., Ghosh, N., & Yu, B. (2024). LORA+: Efficient low-rank adaptation of large models. arXiv preprint arXiv:2402.12354.

Hu, E., Shen, Y., Wallis, P., et al. (2021). LORA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Imankulova, A., Sato, T., & Komachi, M. (2017). Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th workshop on Asian translation* WAT2017, (pp. 70–78).

Jiang, A., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Johnson, O. V., XinYing, C., Johnson, O. E., Khaw, K. W., & Lee, M. H. (2023). Learning rate schedules and optimizers, a game changer for deep neural networks. In *International conference of reliable information and communication*. Cham: Springer Nature Switzerland.

Lakew, M., Federico, M., Negri, M., & Turchi, M. (2021). Multilingual neural machine translation for low-resource languages. *IJCoL*, *4*(1), 11–25.

Li, Y., Yu, Y., Liang, P., Karampatziakis, N., Chen, W., & Zhao, T. (2023). LoftQ: LoRA-fine-tuning-aware quantization for large language models. arXiv Preprint arXiv:2310.08659.

Lialin, V., Muckatira, S., Shivagunde, N., & Rumshisky, A. (2023). Relora: High-rank training through low-rank updates. In *The twelfth international conference on learning representations*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Liu, P., Gao, Y., & Belinkov, Y. (2023). G-EVAL: NLG evaluation using GPT-4 with a holistic approach to diversity and error analysis. arXiv Preprint arXiv:2306.03804.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-EVAL: NLG evaluation using GPT-4 with better human alignment. arXiv Preprint arXiv:2303.16634.

Lv, K., Yan, H., Guo, Q., Lv, H., & Qiu, X. (2023). ADaLoMo: Low-memory optimization with adaptive learning rate. arXiv preprint arXiv:2310.10195.

Mabuya, R., Abbott, J., & Marivate, V. (2021). Umsuka English-Isizulu parallel corpus.

Marivate, V., Banda, F., Lastrucci, R., Nakeng, M., Olalaye, K., & Sindane, T. (2024). MAVITO: South African terminology, lexicon, and glossary project. Available at: https://github.com/dsfsi/za-mavito.

Merx, R., Mahmudi, A., Langford, K., de Araujo, L., & Vylomova, E. (2024). Low-resource machine translation through retrieval-augmented LLM prompting: A study on the mambai language. arXiv preprint arXiv:2404.04809.

Mujadia, V., Urlana, A., Bhaskar, Y., Pavani, P., Shravya, K., Krishnamurthy, P., & Sharma, D. (2023). Assessing translation capabilities of large language models involving English and Indian languages. arXiv preprint arXiv:2311.09216.

Ojo, J., Ogueji, K., Stenetorp, P., & Adelani, D. (2023). How good are large language models on african languages? arXiv preprint arXiv:2311.07978.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*.

Pope, M., Libovický, J., & Helcl, J. (2022). CUNI systems for the WMT22 czech-ukrainian translation task. arXiv preprint arXiv:2212.00486.

Pourkamali, N., & Sharifi, S. (2024). Machine translation with large language models: Prompt engineering for Persian, English, and Russian directions. arXiv preprint arXiv:2401.08429.

Pradha, S., Halgamuge, M., & Vinh, N. (2019). Effective text data preprocessing technique for sentiment analysis in social media data. In *2019 11th international conference on knowledge and systems engineering* (pp. 1–8).

Ranathunga, S., Lee, E., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, *55*(11), 1–37.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1–7).

Song, Y., Ezzini, S., Klein, J., Bissyande, T., Lefebvre, C., & Goujon, A. (2023). Letz translate: Low-resource machine translation for Luxembourgish. In *2023 5th international conference on natural language processing* (pp. 165–170).

Ström Boman, A. (2024). Identifying sensitive data using named entity recognition with large language models: A comparison of transformer models fine-tuned for named entity recognition.

Sun, X., Ji, Y., Ma, B., & Li, X. (2023). A comparative study between full-parameter and LoRA-based fine-tuning on Chinese instruction data for instruction following large language models. arXiv preprint arXiv:2304.08109.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Riviére, M., Kale, M., Love, J., & Tafti, P. (2024). Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th international conference on language resources and evaluation*. [PDF] Available at: https://example.com.

Tsai, S. (2019). Using google translate in EFL drafts: A preliminary investigation. *Computer Assisted Language Learning*, *32*(5-6), 510–526.

Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., & Yang, F. (2024). Parameter-efficient fine-tuning in large models: A survey of methodologies. arXiv preprint arXiv:2410.19878.

Yu, Y., Zuo, S., Jiang, H., Ren, W., Zhao, T., & Zhang, C. (2020). Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. arXiv preprint arXiv:2010.07835.

Zhang, B., Haddow, B., & Birch, A. (2023). Prompting large language model for machine translation: A case study. In *International conference on machine learning* (pp. 41092–41110).

Zhang, X., Zhang, F., & Xu, C. (2023). Reducing vision-answer biases for multiple-choice VQA. *IEEE Transactions on Image Processing*.

Zhang, Y., Zhao, L., Lin, M., Sun, Y., Yao, Y., Han, X., Tanner, J., Liu, S., & Ji, R. (2023). Dynamic sparse no training: Training-free fine-tuning for sparse LLMs. arXiv preprint arXiv:2310.08915.

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023). Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:2304.04675.

Zupon, A., Crew, E., & Ritchie, S. (2021). Text normalization for low-resource languages of Africa. arXiv preprint arXiv:2103.15845.