

Improving Language Models for Low-Resource Languages

Introduction

Large Language models have transformed Natural Language Processing, doing tasks such as question answering, summarisation, and translation well. However, these benefits are limited to high-resource languages such as English, as they underperform for low-resource languages, such as isiZulu. LRL-LLM aims to reduce the gap in LLM usability for languages considered low-resource. Specifically, we close this gap for isiZulu.

Main Contributions

- We generated 3 sets of synthetic isiZulu data. We used machine translation and prompting to create the data
- We achieved better in-language performance for Gemma3-4B by fine-tuning it on our synthetic data
- We investigated how fine-tuning Gemma-3-4B-IT on diverse instruction-tuning datasets (AfridocMT, Aya Collection, and Inkuba-Instruct) affects accuracy across different AfriMMLU subject areas and analysed the impact of zero-shot and few-shot prompting strategies on model performance.
- We achieved substantial gains in model performance on isiZulu math word problems using fine-tuning on Chain-Of-Thought data.
- Group-Relative Policy Optimisation (GRPO) further improved reasoning ability.

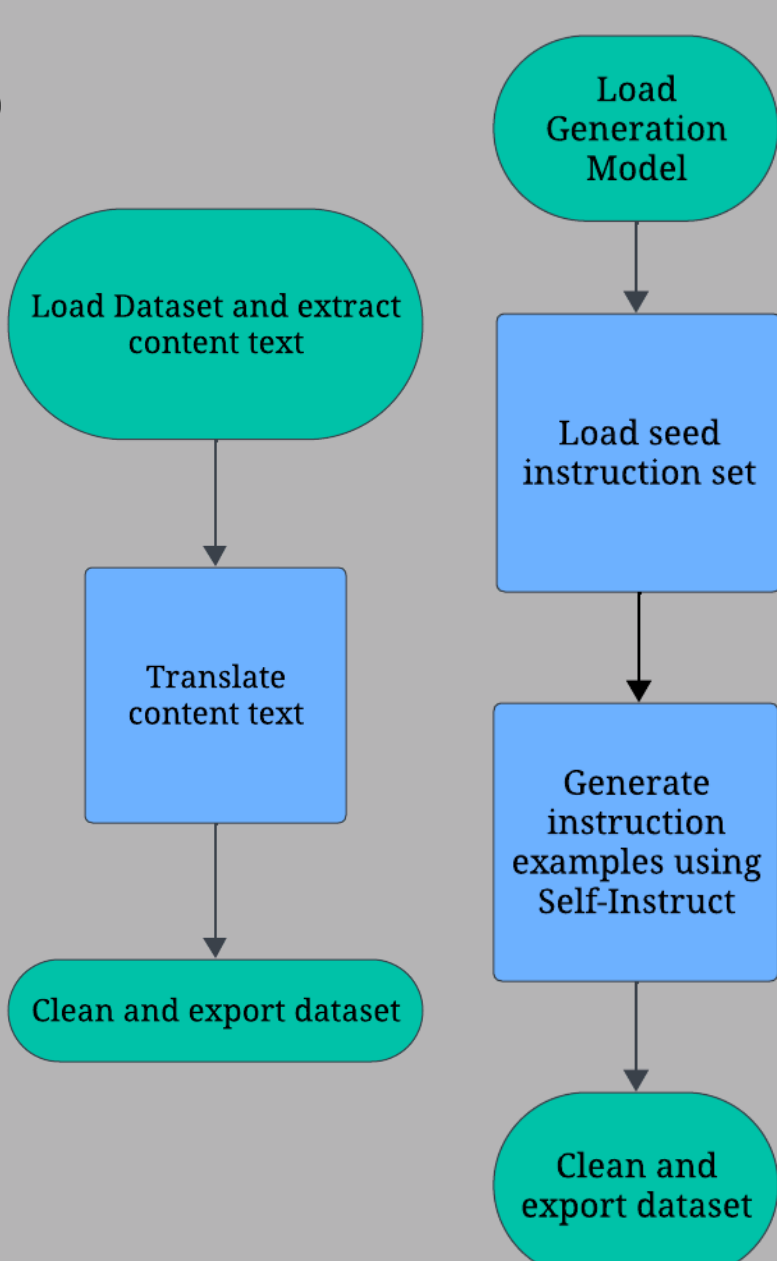
Methodology

Synthetic Data

(1) machine translation (MTL) of an existing English dataset into isiZulu

(2) synthetic isiZulu examples generated by prompting a LLM; and

(3) a combined dataset of the previous two methods.

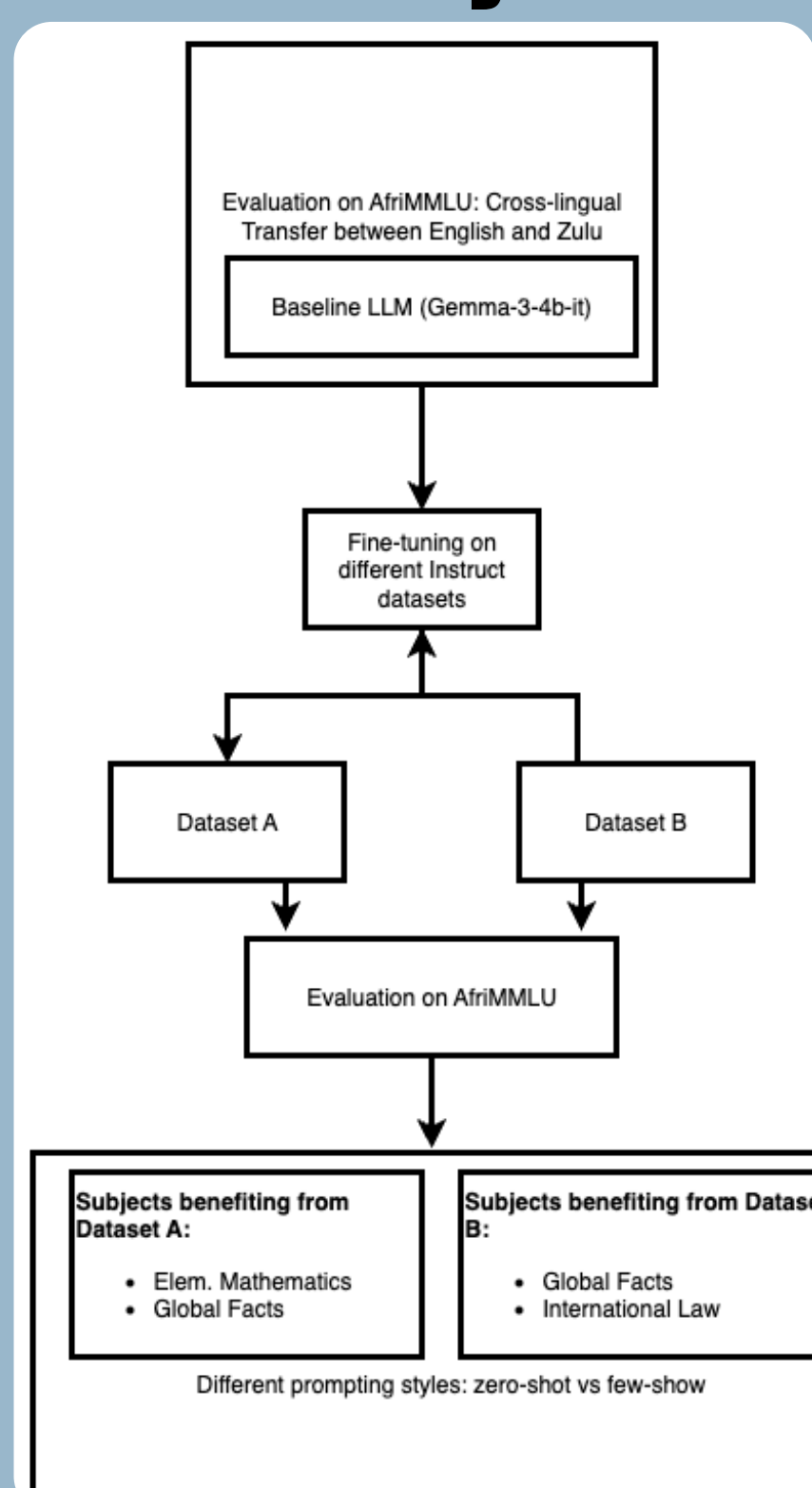


Instruction fine-tuning

1. Cross-lingual transfer evaluation

2. Create 3 fine-tuned models on each dataset

3. Compare baseline with fine-tuned models on AfriMMLU zulu split

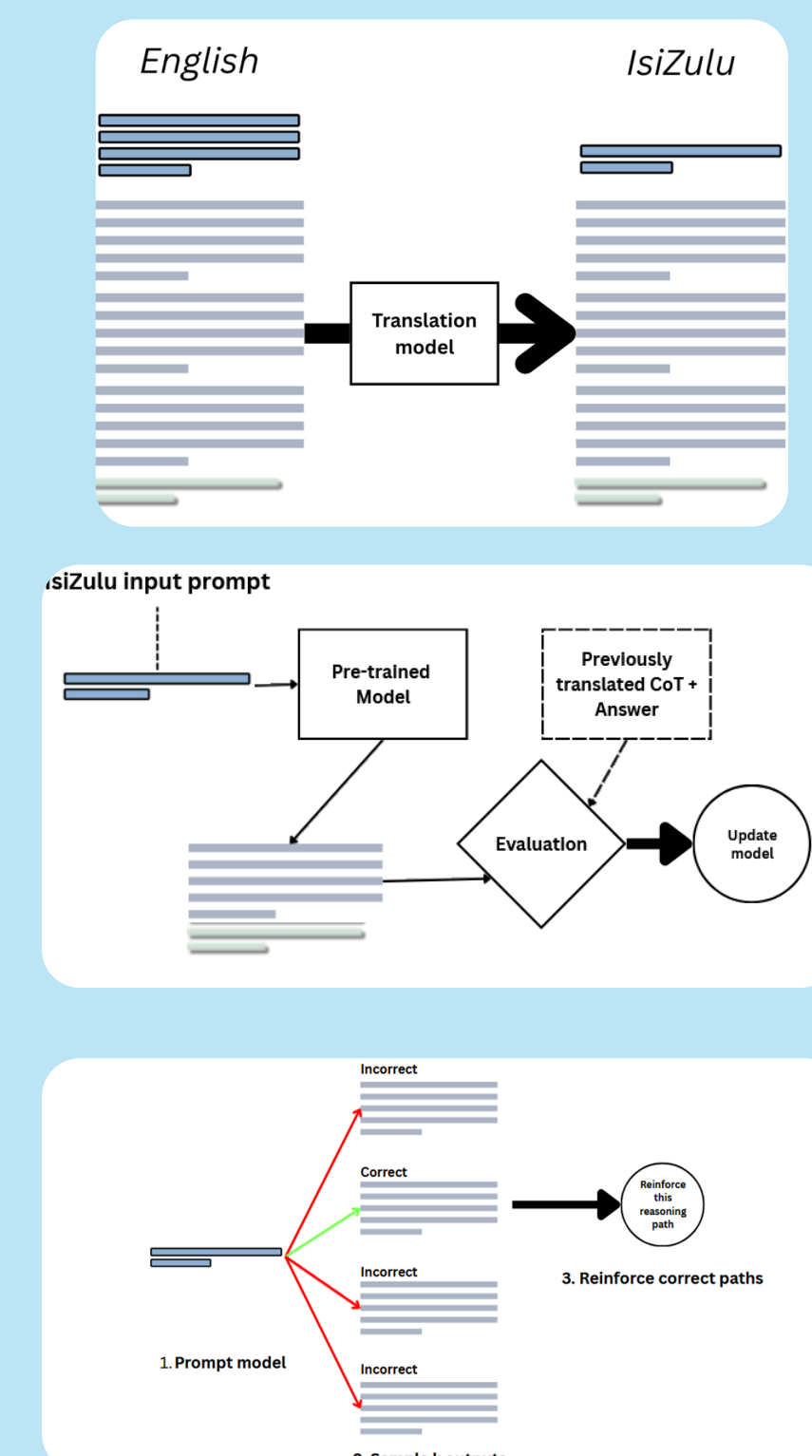


Eliciting reasoning

1. Translate dataset

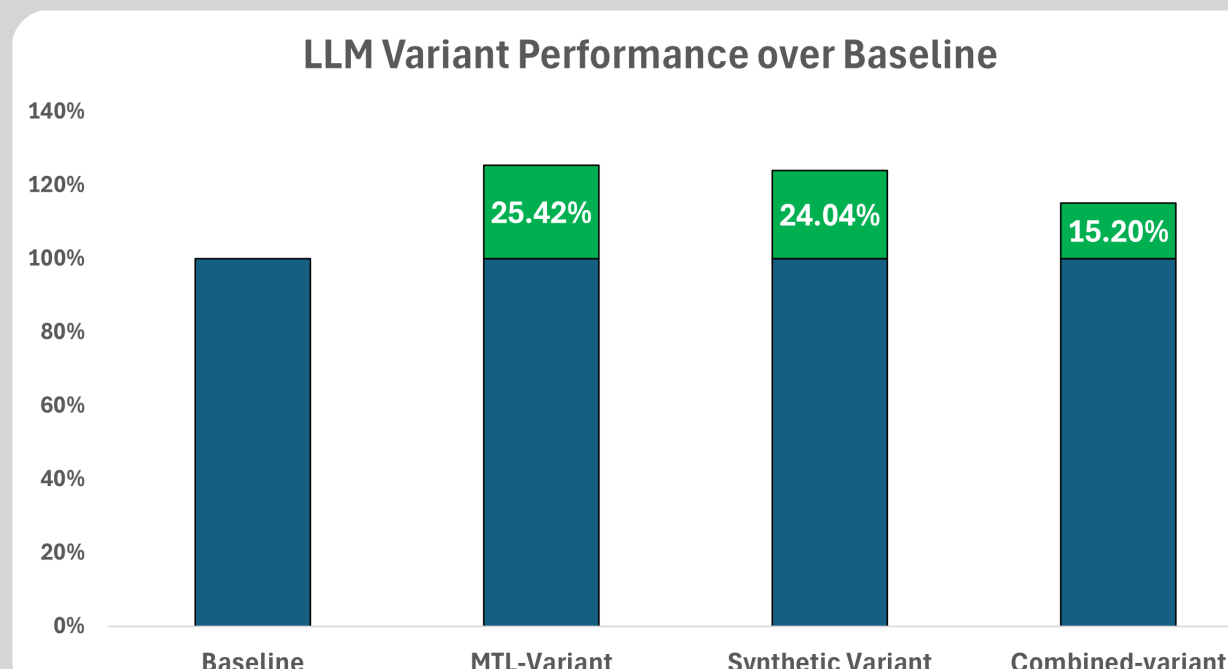
2. Fine-tune

3. Reinforcement learning

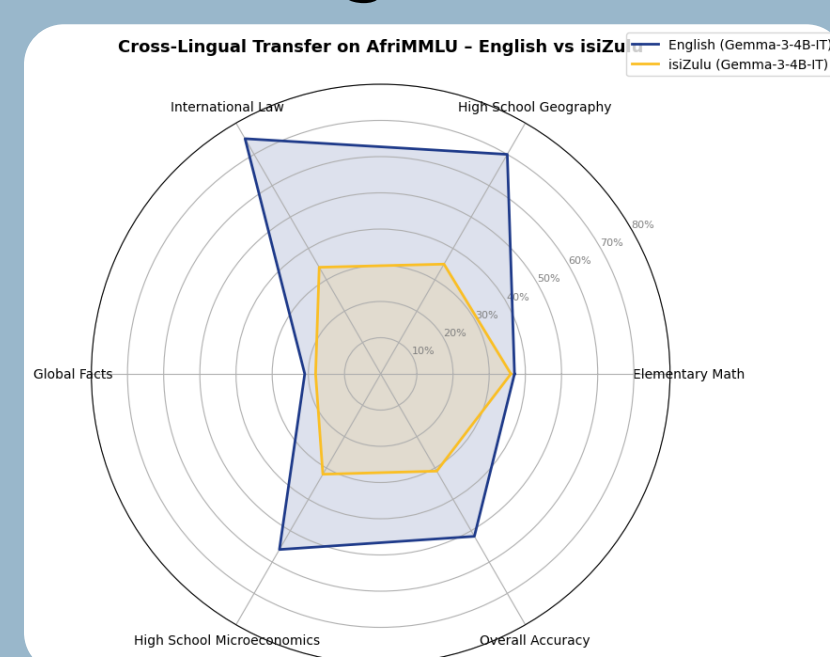


Results

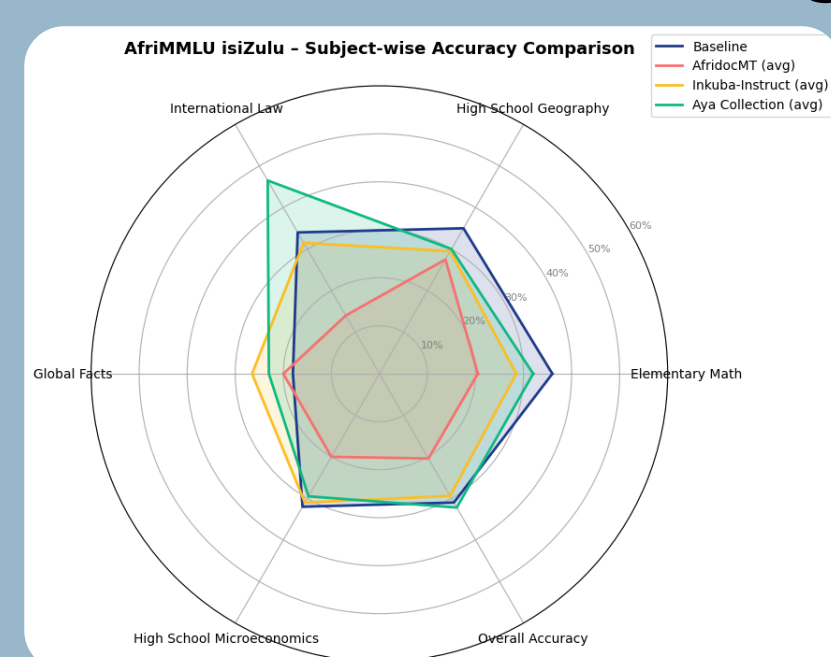
Synthetic Data



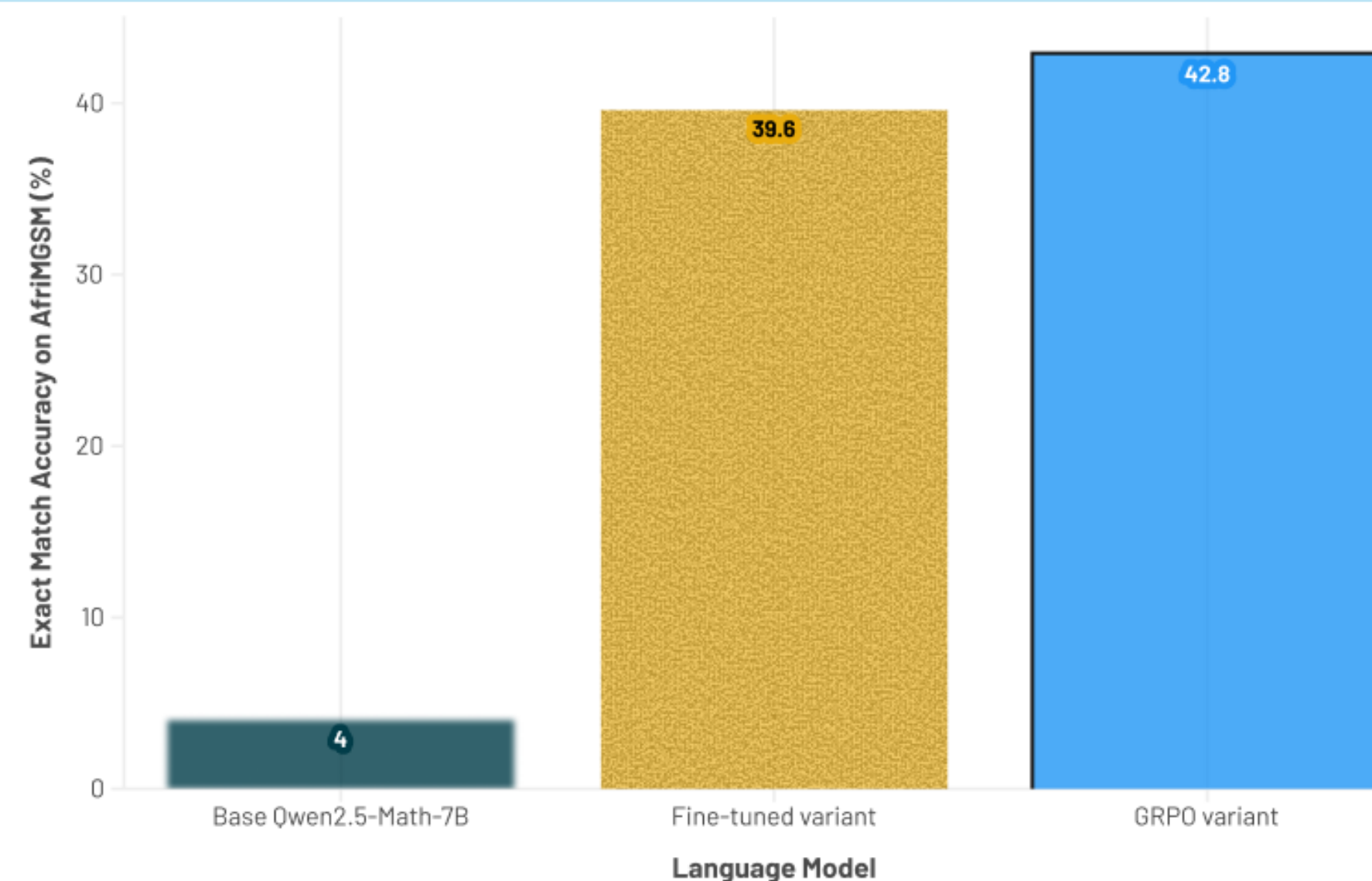
Cross-lingual transfer



Instruction fine-tuning



Eliciting reasoning with fine-tuning and reinforcement learning:



Authors

- Dineo Chiloane chldin002@myuct.ac.za
- Mitchell Johnson jhnmit002@myuct.ac.za
- Corey Webb wbbcor001@myuct.ac.za

Supervisor

Dr. Jan Buys
janbuys@uct.ac.za

Private Bag X3, Rondebosch, 7700
+27 21 650 26663
cs.uct.ac.za

IT | SCHOOL OF IT

