IMPROVING LLMS FOR LINGUISTIC DIVERSITY: IGBO TRANSLATION

Ebelechukwu Nwafor, Ph.D.
Minh Phuc Nguyen
Department of Computing Sciences

ABSTRACT

State-of-the-art large language models (LLMs) like GPT-4o excel in English but often struggle with low-resource languages like Igbo. This research enhances AI linguistic diversity by improving performance in Igbo-English translation using self-created and open-source benchmark datasets. Three finetuned models—NLLB (Meta), DeltaLM (Microsoft), and ByT5 (Google)—showed significant improvements over their original versions and GPT-4o in translation accuracy.

BACKGROUND

Large Language Models (LLMs) are advanced AI trained on extensive text data to generate human-like language. While models like GPT-4o excel in high-resource languages, they often falter in low-resource languages like Igbo, indicating a lack of linguistic diversity.
This research addresses this gap by finetuning three multilingual models—NLLB (Meta), DeltaLM (Microsoft), and ByT5 (Google)—on our Igbo benchmark dataset. Although less popular than GPT models, these multilingual models excel in Igbo-English translation and showed significant improvements, even surpassing GPT-4o, demonstrating the effectiveness of targeted finetuning.

METHODOLOGY

1. **Data Preparation:** To enhance LLM performance in Igbo-English translation, we created a dataset from reputable sources (BBC Igbo, TED Talks, Igbo National Broadcast) and annotated it by native Nigerian speakers for accuracy. We also included benchmark datasets from published research, resulting in approximately 500,000 sentences.
2. **Models Finetuning**: The prepared datasets were used to finetune three multilingual models: NLLB (Meta), DeltaLM (Microsoft), and ByT5 (Google). Each model was trained on the combined datasets to improve their Igbo-English translation performance.
3. **Benchmark Dataset Generation and Evaluation:** To evaluate performance, we use both the original and finetuned models to generate test datasets. These translations were compared to benchmark datasets with available parallel Igbo-English data, allowing us to measure improvements in translation accuracy and demonstrate the effectiveness of finetuning these multilingual models.

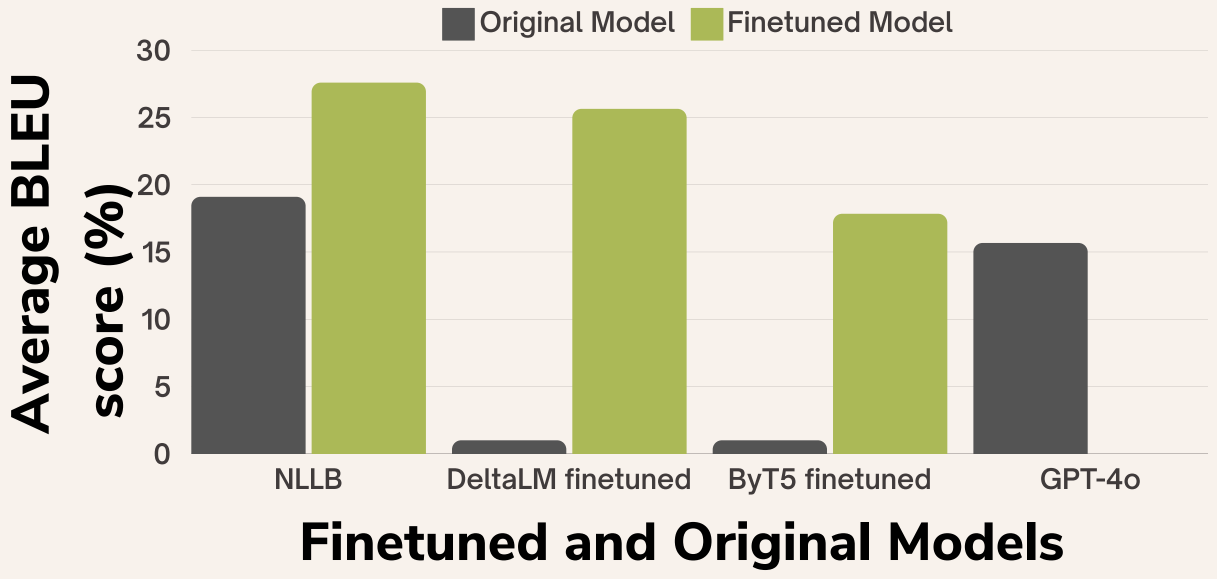
RESULT

For each original and finetuned model, English sentences from the benchmark datasets—Flores200, JW300, MAFAND-MT, and our self-collected data—were used for models to generate Igbo sentences, which are compared with corresponding Igbo sentences from the datasets, resulting in BLEU scores

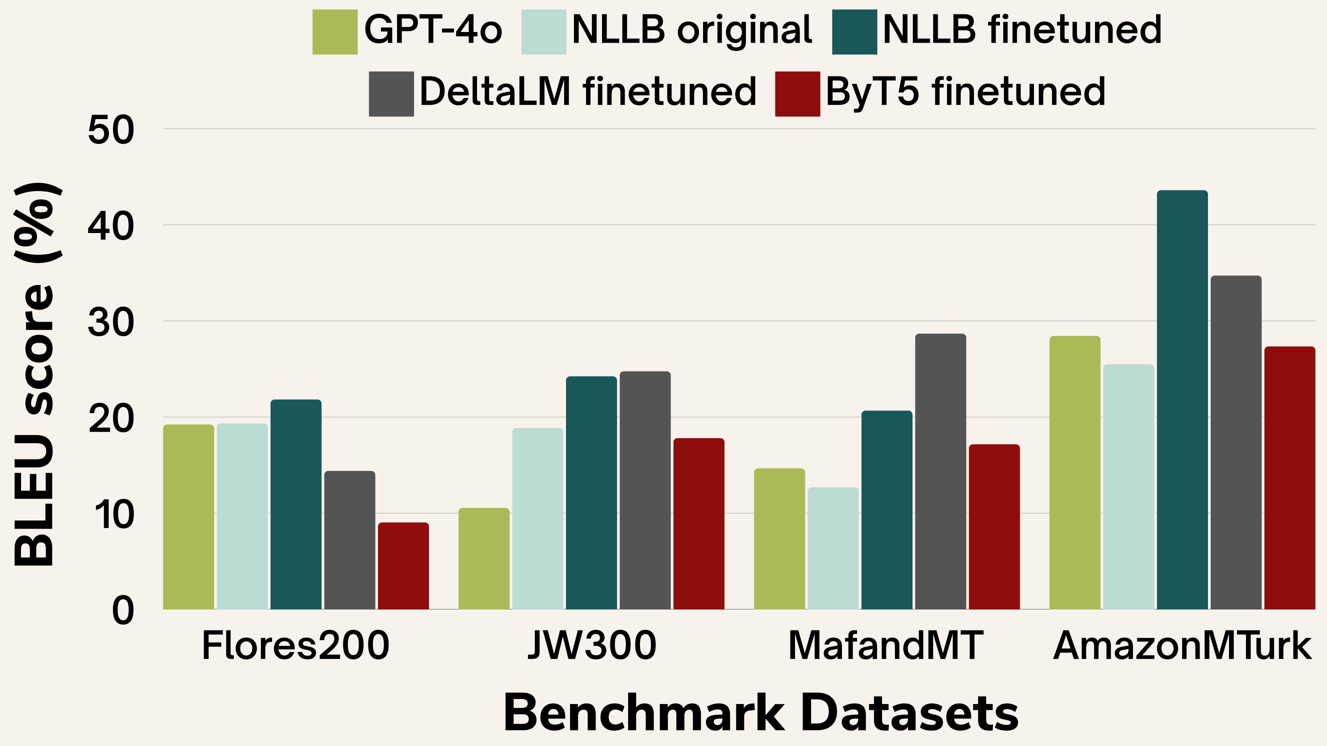
	GPT-4o	NLLB original	NLLB finetuned	DeltaLM finetuned	ByT5 finetuned
Flores200	19.23	19.34	21.83	14.4	9.04
JW300	10.55	18.87	24.24	24.76	17.81
MAFAND-MT	14.67	12.7	20.67	28.68	17.17
Amazon MTurk	18.23	25.5	43.6	34.72	27.35

ANALYSIS

Results demonstrate a significant improvement in BLEU scores for finetuned models on Igbo-to-English benchmark datasets. On average, NLLB's score increases by 44%, ByT5 by 274%, and DeltaLM by 250%.



Furthermore, all finetuned models achieve a higher average BLEU score than GPT-4o, which was not the case for DeltaLM and ByT5 before finetuning.



The chart shows that finetuned NLLB performs best on Flores200 and AmazonMTurk, while finetuned DeltaLM leads on MafandMT and JW300. Overall, finetuned NLLB has the highest average BLEU score at 27.59%, while GPT-4o, without finetuning, scores the lowest at 15.67%). The original DeltaLM and ByT5 models generate poor multilingual sentences, with BLEU scores close to 1%.

CONCLUSION

The research demonstrates that finetuning multilingual LLMs significantly enhance performance in Igbo-to-English translation, especially for low-resource languages. By refining models like NLLB, DeltaLM, and ByT5, we address critical gaps in linguistic diversity, achieving improvements that even surpass advanced models like GPT-4o. These findings highlight how targeted finetuning can improve translation accuracy and better support underrepresented languages.