McGill NLP Group Submission to the MRL 2024 Shared Task: Ensembling Enhances Effectiveness of Multilingual Small LMs

Senyu Li[1,2] Hao Yu[1,2] Jessica Ojo[1,2] David Ifeoluwa Adelani[1,2,3]

[1]Mila - Quebec AI Institute, [2]McGill University, [3]Canada CIFAR AI Chair

{senyu.li,hao.yu2}@mail.mcgill.ca

🎯 Introduction & Background

Problem Statement

- Challenge of limited data availability in non-English languages
- Importance of knowledge transfer between languages
- Need for unified approaches across different NLP tasks

Task Definitions

- **Named Entity Recognition** (*NER*)

  Identification of entities (PER, ORG, LOC)

- **Free-form Question Answering** (*FFQA*)

  Generation of accurate answers from context

  Handling "no answer" scenarios

  e.g.: *"What did Tom buy?" → "Two apples"*

- **Multiple-choice Question Answering** (*MCQA*)

  Four-option selection format

  Context-based reasoning and Precise answer selection

📊 Datasets

NER

MasakhaNER 2.0 (20 African lan)

CoNLL03 (English/German)

Turkish Wiki NER, UZNER (Uzbek)

MCQA

Belebele, RACE (English)

Cosmos QA (English)

FFQA

XTREME-UP (88 languages)

NaijaRC (Nigerian languages)

MLQA (7 languages)

Belebele (Multilingual)

XQuAD (10 languages)

🧮 Methods & Technical Approach

Model Selection (5 Models)

1. **XLM-RoBERTa**
   - Multilingual, Extended BERT architecture
2. **Afro-XLMR(-76L)**
   - MLM adaptation of XLM-R-large
   - Coverage of **17/76** African languages
3. **mT5 (Multilingual T5) [FFQA]**
   - Text-to-text, 101 language
   - Common Crawl corpus training
4. **mT0 (Multilingual T0)**
   - Zero-shot and few-shot capabilities
   - Natural language instruction following
   - Multilingual task generalization
5. **AfriTeVa V2 [FFQA on IG and YO]**
   - T5 architecture derivative
   - Wura pretraining, 16 African languages

Training Techniques (3 Skills)

1. **Curriculum Learning Implementation**
   - Progressive complexity introduction
   - Length-based data organization
   - Improved model learning trajectory
2. **Knowledge Transfer Mechanism**
   - Cross-lingual representation sharing
   - High-to low-resource transfer
   - Shared conceptual understanding
   - Multilingual pattern recognition
3. **Multilingual Data Interleaving**
   - Systematic language mixing
   - Enhanced cross-lingual learning
   - Improved low-resource performance
   - Balanced language representation

FFQA
Task: free-form QA
Context: [Passage]
Question: [Question]
MCQA
Context: [Passage]
Question: [Question]
A. [Text of choice A]
B. [Text of choice B]
C. [Text of choice C]
D. [Text of choice D]

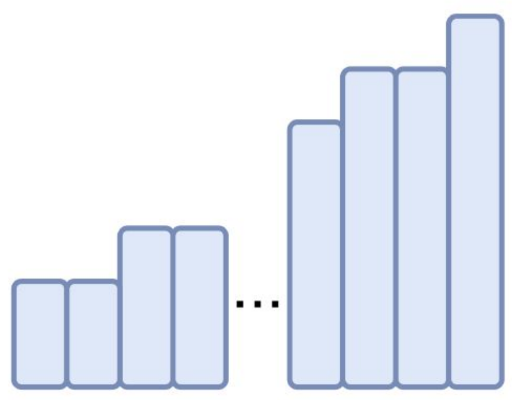

Figure 1: This figure illustrates the process of Curriculum Learning. Shorter data pieces appear earlier in the epoch, while longer data pieces are introduced later.

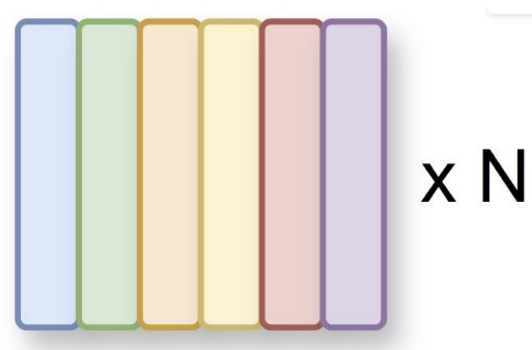

Figure 2: This figure illustrates the process of interleaving multilingual data. Each coloured tile represents a single data sample from a different language. This process is repeated for each data sample in every language, ensuring that each sample appears only once per epoch.

🚀 Results

Models	AZ	YO	TR	IG	ALS	Avg	Mdn
Named Entity Recognition							
Ours	**0.821**	**0.857**	**0.826**	0.093	**0.789**	0.677	**0.821**
CUNI	0.573	0.805	0.778	**0.740**	0.704	**0.720**	0.740
Free Form Question Answering							
Ours	0.421	0.361	0.399	0.331	0.421	0.377	0.399
0-shot Llama-3.1-instruct 7B	0.536	0.468	0.472	0.536	0.425	0.485	0.472
4-shot Llama-3.1-instruct 7B	0.501	0.373	0.451	0.520	0.435	0.452	0.451
0-shot Llama-3.1-instruct 70B	**0.540**	0.508	0.491	0.491	0.478	0.498	0.491
4-shot Llama-3.1-instruct 70B	0.506	0.436	0.460	**0.616**	**0.488**	**0.513**	0.488
0-shot gemma-2 27b	0.448	0.490	0.423	0.347	0.474	0.434	0.448
4-shot gemma-2 27b	0.453	0.458	0.425	0.449	0.478	0.458	0.453
0-shot aya-101 13B	0.398	0.444	0.370	0.318	0.419	0.390	0.398
4-shot aya-101 13B	0.404	0.451	0.364	0.453	0.422	0.434	0.422
0-shot o1-preview	0.535	**0.525**	**0.520**	0.428	0.458	0.480	**0.520**
Multiple Choice Question Answering							
Ours	0.969	0.853	0.816	**0.969**	0.777	0.879	0.853
FT mT5 large	0.966	0.848	0.810	0.965	0.778	0.876	0.848
FT mT0 large	0.966	0.824	0.830	0.965	0.769	0.869	0.830
FT AfriTeVa V2 large	0.807	0.784	0.592	0.949	0.580	0.772	0.784
0-shot Llama-3.1-instruct 7B	0.969	0.731	0.884	0.954	0.788	0.849	0.884
4-shot Llama-3.1-instruct 7B	0.931	0.737	0.701	0.933	0.782	0.827	0.782
0-shot Llama-3.1-instruct 70B	0.979	0.896	0.939	0.959	0.917	0.932	0.939
4-shot Llama-3.1-instruct 70B	0.976	0.881	0.966	0.963	**0.923**	0.932	0.963
0-shot gemma-2 27b	0.979	0.891	0.946	0.963	0.886	0.925	0.946
4-shot gemma-2 27b	**0.983**	0.905	0.932	0.967	0.898	0.932	0.932
0-shot aya-101 13B	0.969	0.881	0.905	0.967	0.834	0.906	0.905
4-shot aya-101 13B	0.969	0.860	0.871	0.967	0.834	0.898	0.871
0-shot o1-preview	0.976	**0.911**	**0.973**	0.967	0.922	**0.941**	**0.967**

Table 2: The final results of each model on the test set for each task.

🔍 Results Analysis and Findings

NER

Results ⭐ Poor in Igbo to lower the average score. But get top performance in **4/5** languages.

Analysis 💡 Ensemble Method Refinement Given the strong performance of our system in most languages, further refinement of our base methods could potentially improve the final results, especially if we can address the models' performance issue on Igbo.

FFQA

Results ⭐ Larger models (e.g., Llama-3.1-70B) consistently outperformed smaller ones even with 0-shot setup. Our model only performance well in Azerbaijani a little bit.

Analysis 💡 *Gap with larger models* The significant performance gap between our system and larger models. *Zero-shot vs. few-shot* Fewshot or Not?

MCQA

Results ⭐ Our system gets an average accuracy of 0.879 across all languages and performs exceptionally well on Azerbaijani and Igbo, followed by Yorùbá, Turkish, and Swiss Germam. Besides, all models also do well in MCQ, hard to find the gap. But the 4-shot Gemma-2 27b (open source) and 0-shot o1-preview (closed source) **both** shows competitive results.

Analysis 💡 SLM has competitive performance compare with larger models. MCQ is a easy for language pre trained knowledge model, but still lack of generalizatioin ablilty across test dataset.

Takeaway Findings

1. **Model Size Impact**: Larger models like Llama-3.1-instruct 70B consistently outperformed smaller models. Performance gap was more pronounced in FFQA than MCQA.
2. **Language-Specific Variations**: Performance varied significantly across languages. Generally better results for Azerbaijani and Swiss German. African languages (Yorùbá and Igbo) often showed lower performance. Specialized African language model (AfriTeVa V2) performed well on Igbo but struggled with non-African languages.
3. **Ensemble Effectiveness**: Ensemble approach proved effective, particularly for MCQA. Combined predictions from multiple models improved overall accuracy.

📋 Acknowledgements

This research was supported by MILA compute. Prof. David Adelani is supported by the Canada CIFAR AI Chair program. We are grateful to OpenAI for providing API credits through their Researcher Access API programme to Masakhane for the evaluation of GPT LLMs.

Feel Free to Check More Deatils on Our Paper 🤗