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@mail.mcgill.ca

Abstract

We present our systems for the three tasks and five languages included in the MRL 2024 Shared Task on Multilingual Multi-task Information Retrieval: (1) Named Entity Recognition, (2) Free-form Question Answering, and (3) Multiple-choice Question Answering. For each task, we explored the impact of selecting different multilingual language models for fine-tuning across various target languages, and implemented an ensemble system that generates final outputs based on predictions from multiple fine-tuned models. All models are large language models fine-tuned on task-specific data. Our experimental results show that a more balanced dataset would yield better results. However, when training data for certain languages are scarce, fine-tuning on a large amount of English data supplemented by a small amount of "triggering data" in the target language can produce decent results.[1]

1 Introduction

In this paper, we present our submission for the MRL 2024 shared task[2]. The shared task includes the following three tasks: Named Entity Recognition (NER), Free-form Question Answering (FFQA), and Multiple-choice Question Answering (MCQA). Each task involves a final test set for five languages: Igbo, Swiss German, Turkish, Azerbaijani, and Yorùbá. Our systems are designed to support all of these languages simultaneously.

Our systems leveraged the remarkable success of transformer-based (Vaswani et al., 2017), pre-trained Language Models (LMs) such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019), which have demonstrated outstanding performance in various Natural Language Processing (NLP) tasks in recent years. These models, with their large number of parameters and pre-training on vast datasets, have proven to be highly effective in extracting and representing information possessed by input sequences (Brown et al., 2020). Their strong generalization capabilities make them well-suited for fine-tuning on specific tasks, such as NER and translation. Multilingual pre-trained LLMs, like XLM-RoBERTa (Conneau et al., 2019), mT5 (Xue et al., 2021), and their variants, which were trained on extensive multilingual datasets, are particularly effective for multilingual tasks. These models capture semantic structures/knowledge shared across languages, enhancing their ability to transfer knowledge between languages. Fine-tuning these models for specific tasks allowed us to fully utilize their rich token-level and sentence-level semantic representations, which are essential for tasks requiring detailed language understanding. For instance, NER benefits from the token-level granularity learned during pretraining (Yan et al., 2019), while FFQA and MCQA require robust sentence-level comprehension, which these models provide (Robinson et al., 2023; Myrzakhan et al., 2024). The combination of pre-training on extensive multilingual corpora and task-specific fine-tuning enabled our system to achieve decent performance across all five target languages.

During the fine-tuning phase, in addition to hyper-parameter selection, our systems employed other strategies to promote a smoother and faster-converging learning process, such as using data from languages closely related to the target languages, applying curriculum learning (Bengio et al., 2009), and interleaving data from various languages to enhance model performance and smooth the learning process.

The experiment results show that different base models with a similar number of parameters exhibit varying advantages for different languages after fine-tuning. Ensembling the outputs from each model results in better and more robust overall performance. Additionally, given the limited availabil-

[*] Corresponding author
[1] Our codes will be made available at this link.
[2] The website of the shared task is available at this link.

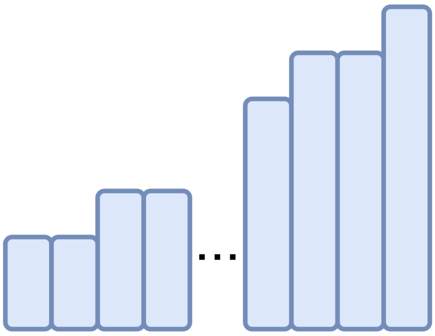

Figure 1: This figure illustrates the process of Curriculum Learning. Shorter data pieces appear earlier in the epoch, while longer data pieces are introduced later.

ity of data for certain languages, leveraging large amounts of task-specific data in English— which is the easiest to obtain— along with smaller amounts of data in the target language, allows knowledge transfer from English to the target language. This approach outperforms fine-tuning exclusively on the limited data available in the target language.

2 Background

In this section, we provide a brief overview of the background knowledge for the three tasks involved in the shared task, along with the three techniques we employed to facilitate the learning process.

Named Entity Recognition NER is an NLP task that focuses on identifying and categorizing specific tokens or phrases in a text as belonging to predefined entity types, such as persons (PER), organizations (ORG), locations (LOC), dates (DATE), and other relevant categories. A named entity refers to a real-world object or concept that can be recognized by its proper name within the text. For example, in the sentence "Barack Obama visited Paris in 2015," the named entities are "Barack Obama" (person), "Paris" (location), and "2015" (date). In this shared task, we only consider three entity tags: persons, organizations, and locations.

Free-form Question Answering FFQA involves providing answers to natural language questions based on the information given. This task assumes an information-seeking scenario, where users ask questions without knowing the answer in advance, and the system is responsible for finding a relevant answer based on information presented in the passage (if one exists). In this task, the system is given a question, a title, and a passage, and must either generate a text sequence for the correct answer or indicate that there is no answer for the question based on information available in the passage by generating the text sequence “no answer”. For example, consider the passage: "Tom went to the supermarket and bought two apples." If the question is "What did Tom buy in the supermarket?", the system should return the answer "Two apples." However, if the question is "Which supermarket did Tom visit?", the system should respond with "no answer," as the passage does not specify the name of the supermarket.

Multiple-choice Question Answering Similar to FFQA, MCQA assumes a scenario where users seek information by asking questions without knowing the answer and are given a question, title, and passage. However, unlike FFQA, the MCQA system is also provided with four potential options, and its task is to identify the correct one based on the information in the passage. For instance, consider the passage: "Tom went to the supermarket and bought two apples." If the question is "How many apples did Tom buy?" and the four options are "A. 1", "B. 2", "C. 3", and "D. 4", the system should return "B".

Curriculum Learning Curriculum learning (Bengio et al., 2009) is a machine learning strategy that gradually introduces a model to progressively more challenging data pieces over multiple training iterations. This method can often produce better results compared to using a randomly shuffled training set. This approach is effective in the sense that, the model begins by learning general concepts through simpler examples, and then incrementally incorporates more detailed and complex information as more difficult examples are introduced. For our systems, we define "difficulty" by the length of the input text, where longer text equates to greater complexity and comes later in the epoch, as shown in Figure 1. Since curriculum learning is a paradigm that focuses solely on the selection and ordering of training data, it can be integrated with various other machine learning techniques, like Interleaving Multilingual Data Pieces which we will introduce later in this section.

Knowledge Transfer Knowledge transfer in multilingual LLMs refers to the model’s ability to leverage information, patterns, or representations learned in one language to enhance its performance or understanding in another. This happens because

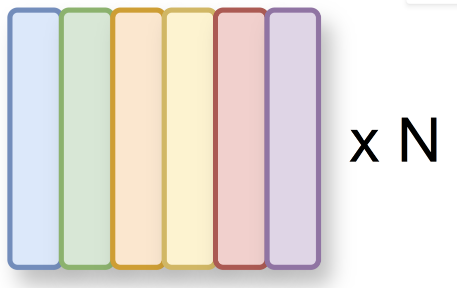

Figure 2: This figure illustrates the process of interleaving multilingual data. Each coloured tile represents a single data sample from a different language. This process is repeated for each data sample in every language, ensuring that each sample appears only once per epoch.

multilingual LLMs develop shared representations of concepts that can be applied across different languages. To facilitate the knowledge transfer for our base models, we fine-tuned the base models on diverse multilingual data. This includes a relatively small amount of data for the target languages, additional data for languages closely related to the target languages, and a large amount of data from high-resource languages like English.

Interleaving Multilingual Data Pieces Interleaving Multilingual Data Pieces is a machine learning technique used to train multilingual models by interleaving data from various languages during training. This approach promotes cross-lingual knowledge transfer by encouraging the model to develop shared linguistic representations and structures, which improves its ability to generalize across languages. It is especially effective in cross-lingual information retrieval scenarios, allowing the model to utilize common features across languages and enhance performance in low-resource language settings. An illustration of this approach can be found in Figure 2

3 Methods

In this section, we provide a detailed illustration of each system we implemented for the three tasks involved in MRL 2024.

3.1 Models Included

XLM-RoBERTa XLM-RoBERTa (Conneau et al., 2019) is a transformer-based masked language model, which is a multilingual version of the RoBERTa model, designed to handle text in multiple languages by extending the BERT architecture.

Afro-XLMR AfroXLMR (Alabi et al., 2022) is a variant of the XLM-RoBERTa model specifically tailored for African languages. While XLM-RoBERTa is designed to work across 100 languages, it may underperform for African languages due to limited data in these languages during training. AfroXLMR addresses this by focusing on improving the model's performance in African linguistic contexts, by using MLM adaptation of XLM-R-large model on 17 African languages, covering the major African language families and three high-resource languages. Previous work(Adelani et al., 2022) has empirically demonstrated that this model performs strongly in NER for African languages.

mT5 The mT5 (Multilingual Text-to-Text Transfer Transformer) (Xue et al., 2021) is a variant of the T5 (Text-to-Text Transfer Transformer) architecture. mT5 is pre-trained on a massive multilingual dataset covering 101 languages from the Common Crawl corpus, which enables it to perform a wide range of natural language processing tasks. It operates using a text-to-text framework, where all tasks are framed as feeding text inputs and generating text outputs.

mT0 The mT0 (Multilingual T0) (Muennighoff et al., 2023) is a variant of the T0 model, designed to extend its zero-shot and few-shot learning capabilities to a multilingual context. It is based on the T5 architecture but trained to follow natural language instructions using multilingual data, allowing it to generalize across a wide range of languages and tasks without requiring task-specific training.

AfriTeVa V2 AfriTeVa V2 (Oladipo et al., 2023) is a multilingual sequence-to-sequence model derived from the T5 architecture, designed to support African languages. AfriTeVa V2 was pretrained on Wura which contains 20 languages, including 16 African languages, including Yorùbá and Igbo, alongside globally spoken languages like English and French.

3.2 Named Entity Recognition

We fine-tuned three models for the NER task: `xlm-roberta-large`, `afro-xlmr-large`, and `afro-xlmr-large-76L`. A linear layer was added to the final hidden states of the Transformer encoder for each model, followed by a softmax activation to predict the probability distribution for each token.

During training, each input sequence was first tokenized, meaning that each token was either tokenized as a whole or split into multiple tokens. For tokens that form parts of a word, only the first token is used for prediction. For example, if the word "eating" is tokenized into "eat" and "ing," only the prediction for "eat" will be considered as the final prediction for the word "eating" and the loss will be calculated only for the token “eat”

For the tokens in the test set, we gathered predictions from each model and applied majority voting. The prediction that occurred most frequently was selected as the final output for that token. In the case of a tie, where all three models produced different predictions, we chose the prediction from `afro-xlmr-large`, which is the best-performing model in the development phase.

3.3 Free-form Question Answering

We fine-tuned two models for the Free-form Question Answering task: mT5 Large and AfriTeVa V2 Large. The models were trained using a sequence-to-sequence (seq2seq) text generation approach. During training, the model was optimized to minimize the cross-entropy loss between the predicted tokens and the actual target tokens. The input data formatting template is shown in Table 1.

For the final submission, we chose to use the fine-tuned AfriTeVa V2 Large for the two African languages and mT5 Large for the other three non-African languages. This decision was based on the fact that AfriTeVa V2 Large is specifically adapted for African languages, while mT5, being designed for more general language tasks, performs better with non-African languages.

3.4 Multiple-choice Question Answering

We finetuned 3 models for the Multiple-choice Question Answering task: mT5 Large, mT0 Large and AfriTeVa V2 Large. The models were trained using the seq2seq text generation approach. Similar to the finetuning for FFQA, The input data formatting template is shown in Table 1

During the fine-tuning phase, we modified the target output that the model was optimized to predict. Instead of solely predicting the letter corresponding to the correct choice, we adjusted the model to predict both the letter and the text associated with the choice. For example, given the passage: "Tom went to the supermarket and bought two apples." and the question: "How many apples did Tom buy?" with the four options: "A. 1", "B. 2", "C. 3", and "D. 4", rather than training the model to predict only the letter "B," we trained it to predict "B. 2". During inference, we extracted the first token generated (the letter) as the final prediction. This adjustment led to improved performance and faster convergence during the development phase compared to using the original target text.

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]

Table 1: Input templates for MCQA and MMQA.

For each question in the test set, we collected predictions from each model and applied majority voting. The prediction that occurred most frequently was selected as the final answer for that question. In case of a tie, where all three models produced different predictions, we chose the prediction from mT5 Large, as it was the best-performing model during the development phase.

4 Experiment

In this section, we provide detailed information about our implementation, including the computational resources used to run the experiments, the specifics of the training process, and the datasets used to train the models for each of the three tasks. Additionally, we will present the results on the test set provided by the organizers of this shared task, along with an analysis of the experimental results.

4.1 Setup

We used one Nvidia A100 80G GPU for all experiments. We used the Trainer of huggingface transformers to fine-tune all the models.

4.2 Datasets

This section lists all the datasets used to train models for each of the three tasks. All datasets are publicly available. For datasets that were not associated with any papers, we listed them in the Ap-

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.

pendix B. For the final submission, we integrated the validation set provided by the organizers into our training set to reduce the gap between the training set and the test set.

4.2.1 Named Entity Recognition

We used data of 10 languages from 5 datasets to fine-tune models for the NER task. For each dataset, we masked out NER tags that were not included in this shared task.

MasakhaNER 2.0 MasakhaNER 2.0 (Adelani et al., 2022) is a human-annotated NER dataset for 20 African languages. For our study, we utilized the Yorùbá and Igbo data in this dataset. Additionally, we included data in Naija, Hausa, and chiShona to facilitate knowledge transfer.

We chose to include Naija, Hausa, and chiShona in our training data because Hausa and Naija are the top two transfer languages for Yorùbá, while chiShona is the best transfer language for Igbo (with Yorùbá as the second-best), as shown in the study by Adelani et al..

CoNLL03 CoNLL03 (Tjong Kim Sang and De Meulder, 2003) consists of annotations of NER tags across English and German languages. In our experiments, we used the data from both languages.

Turkish Wiki NER Dataset Turkish Wiki NER dataset (Altinok, 2023) is an NER dataset which contains 20,000 manually annotated sentences selected from TWNERTC dataset (Sahin et al., 2017).

UZNER UZNER(Yusufu et al., 2023) is a benchmark manually dataset specifically designed for NER tasks in the Uzbek language.

4.2.2 Free-form Question Answering

XTREME-UP XTREME-UP (Ruder et al., 2023) is a benchmark focus on the scarce data across 88 languages and 9 tasks. We used the Indonesian and English data of the "qa in lang" task in this dataset.

MLQA MLQA (Lewis et al., 2019) is an extractive QA evaluation benchmark contain across 7 languages. We used German data of this dataset.

XQuAD XQuaAD (Artetxe et al., 2019) is a cross-lingual question answering dataset composed of paragraphs and question-answer pairs selected from SQuAD v1.1 (Rajpurkar et al., 2016) translated into ten languages. We used German and Turkish data of this dataset.

NaijaRC NaijaRC (Aremu et al., 2024) is a multiple-choice reading comprehension dataset consisting of questions from high school reading comprehension exams in three native Nigerian languages. We used the Igbo, Yorùbá, and Hausa data from this dataset.

Belebele Belebele (Bandarkar et al., 2024) is a multilingual multiple-choice machine reading comprehension dataset. We transformed it into an FFQA dataset by removing the multiple-choice options and setting the text associated with the correct option as the target answer. We used the Azerbaijan, Igbo, Indonesian, English, German, Turkish, Uzbek, Yorùbá, and Hausa data from this dataset. We filtered out some questions if the question is not a closed question.

4.2.3 Multiple-choice Question Answering

Belebele For MCQA, we used the data from the same set of languages as for the FFQA dataset.

Cosmos QA Cosmos QA (Huang et al., 2019) is a commonsense-based reading comprehension dataset in English, formulated as multiple-choice questions.

RACE RACE (Lai et al., 2017) is a large-scale reading comprehension dataset in English

4.3 Results

The Table 2 demonstrates the final results of our model and other LLMs applied to these tasks. Currently, there is a lack of final results from the official leaderboard. We will only include the FFQA and MCQA results.

4.3.1 Free-form Question Answering

Our model achieved an average F1 score of 0.377 across all five languages. The performance varied across languages, with the highest scores observed for Azerbaijani and Swiss German (both 0.421), followed by Turkish (0.399), Yorùbá (0.361), and Igbo (0.331).

Compared to the baseline models, our system's performance was generally lower. The best-performing baseline was the 4-shot Llama-3.1-instruct 70B model, with an average F1 score of 0.513. The 0-shot o1-preview model also performed well, achieving the highest score for Azerbaijani (0.535) and competitive scores for other languages.

4.3.2 Multiple-choice Question Answering

Our MCQA system demonstrated strong performance, achieving an average accuracy of 0.879 across all languages. The system performed exceptionally well on Azerbaijani (0.969) and Igbo (0.969), followed by Yorùbá (0.853), Turkish (0.816), and Swiss German (0.777).

Among the individual models we fine-tuned, mT5 large performed the best with an average accuracy of 0.876, closely followed by mT0 large at 0.869. The AfriTeVa V2 large model, despite being specifically adapted for African languages, showed lower overall performance (0.772) but performed well on Igbo (0.949).

Our ensemble system outperformed all of our individual fine-tuned models, demonstrating the effectiveness of the ensemble approach. However, some of the larger baseline models, particularly the 0-shot o1-preview and the 4-shot versions of Llama-3.1-instruct 70B and gemma-2 27b, achieved higher average accuracies (0.941, 0.932, and 0.932 respectively).

4.3.3 Named Entity Recognition

Our NER system demonstrated strong performance across most languages in the shared task and achieved the highest F1 scores for four out of the five languages (Azerbaijani, Yorùbá, Turkish, and Swiss German) among all participant teams.

4.4 Analysis

4.4.1 Named Entity Recognition

Investigate Igbo Anomaly A detailed analysis of our model's behaviour on Igbo data is crucial. This could include examining the training data and the model predictions.

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. Incorporating elements from our system and CUNI's system might result in a more robust and universally effective NER model for diverse languages.

4.4.2 Free-form Question Answering

Language-specific performance Our system's performance varied across languages, with better results for Azerbaijani and Swiss German compared to African languages like Yorùbá and Igbo. This disparity might be due to differences in the quality or quantity of training data available for each language.

Gap with larger models The significant performance gap between our system and larger models like Llama-3.1-instruct 70B highlights the advantage of massive pre-training and model size in tackling complex FFQA tasks.

Zero-shot vs. few-shot Interestingly, for some baseline models (e.g., Llama-3.1-instruct 7B), the zero-shot performance was better than the few-shot performance. This suggests that for some languages, providing examples might not always lead to improved performance and could potentially introduce biases.

4.4.3 Multiple-choice Question Answering

Strong overall performance Our MCQA system demonstrated robust performance across all languages, with particularly high accuracies for Azerbaijani and Igbo. This suggests that our approach of fine-tuning multilingual models and using ensemble methods is effective for MCQA tasks.

Ensemble effectiveness The superior performance of our ensemble system compared to individual fine-tuned models validates our approach of combining predictions from multiple models to improve overall accuracy.

Language-specific variations The performance variations across languages (e.g., lower accuracy for Swiss German) indicate that language-specific challenges persist even in MCQA tasks. This could be due to factors such as linguistic complexity, dataset quality, or the model's pre-training data distribution.

Competitiveness with larger models While some larger baseline models outperformed our system, the performance gap is smaller compared to the FFQA task. This suggests that our approach is particularly effective for MCQA, where the task structure might allow for better utilization of fine-tuning on limited data.

AfriTeVa V2 performance The specialized AfriTeVa V2 model showed strong performance on Igbo but underperformed on non-African languages. This highlights the trade-off between language-specific models and more general multilingual models.

5 Conclusion

Our study on multilingual multi-task information retrieval revealed key insights across NER, FFQA, and MCQA tasks. In the MCQA task, our ensemble models demonstrated particular strength, outperforming individual fine-tuned models. This underscores the benefits of combining predictions from multiple models to boost accuracy and robustness. For NER, our system showed strong performance across most languages, achieving the highest F1 scores in four out of five languages compared to the other participating systems. However, we observed a significant performance drop for Igbo, highlighting the challenges of consistent performance across diverse languages. We observed variable performance across tasks, with challenges particularly evident in FFQA and significant differences across languages, especially in low-resource settings. This variability was also present in NER, where our model's performance on Igbo lagged significantly behind other languages.

Looking forward, our findings suggest several promising areas for improvement. Enhancing FFQA performance through better fine-tuning strategies and exploring cross-lingual transfer methods is crucial. Developing task-specific model architectures that can better capture the nuances of each task while maintaining multilingual capabilities could lead to significant advances. Improving data augmentation and efficient fine-tuning approaches, especially for low-resource languages, remains a key challenge. Increasing model interpretability will be vital to better understand and address performance discrepancies across languages and tasks. For NER, investigating the causes of the performance anomaly in Igbo and refining our ensemble method could create a more universally effective model across diverse languages. While our approach shows promise, particularly for MCQA and most languages in NER, there is substantial room for further research. The goal remains to develop robust, multilingual, multi-task information retrieval systems that can overcome language barriers, address performance inconsistencies, and improve access to global information across a wide range of languages and task types.

Acknowledgements

This research was supported by MILA compute. 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.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Duygu Altinok. 2023. A diverse set of freely available linguistic resources for Turkish. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13739–13750, Toronto, Canada. Association for Computational Linguistics.

Anuoluwapo Aremu, Jesujoba O. Alabi, Daud Abolade, Nkechinyere F. Aguobi, Shamsuddeen Hassan Muhammad, and David Ifeoluwa Adelani. 2024. Naijarc: A multi-choice reading comprehension dataset for nigerian languages.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Yoshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 1856–1884. Association for Computational Linguistics.

H. Bahadir Sahin, Caglar Tirkaz, Eray Yildiz, Mustafa Tolga Eren, and Ozan Sonmez. 2017. Automatically annotated turkish corpus for named entity recognition and text categorization using large-scale gazetteers.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. TENER: adapting transformer encoder for named entity recognition. *CoRR*, abs/1911.04474.

Aizihaierjiang Yusufu, Liu Jiang, Abidan Ainiwaer, Chong Teng, Aizierguli Yusufu, Fei Li, and Donghong Ji. 2023. Uzner: A benchmark for named entity recognition in uzbek. In *Natural Language Processing and Chinese Computing*, pages 171–183, Cham. Springer Nature Switzerland.

A Prompt

A.1 Zero-Shot Prompt

A.1.1 FFQA Prompt

```
You are an AI assistant designed to answer questions based on given passages.
Your task is to provide accurate and concise answers to questions using only the information provided in the passage.
If the passage doesn't contain enough information to answer the question, respond with 'The passage does not provide sufficient information to answer this question.'
```

```
Do not use any external knowledge
    or make assumptions beyond
   what is explicitly stated in
   the passage. Response should
   be in one line without any
   additional information and
   response in source language.

Passage: {Passage}

Question: {Question}

Your answer:
```

A.1.2 MCQA Prompt

```
You are an AI assistant designed
   to answer multiple-choice
   questions.
Your task is to select the most
   appropriate answer from the
   given options (A, B, C, D)
   based on the question provided
   .
Analyze the question and options
   carefully before making your
   selection.
Your response should only contain
    the letter of the correct
   option (A, B, C, or D).
If none of the options seem
   correct or if there isn't
   enough information to make a
   selection, respond with '
   Unable to determine the
   correct answer based on the
   given options.'

Passage: {Passage}
Question: {Question}
Options:
A) {OptionA}
B) {OptionB}
C) {OptionC}
D) {OptionD}
Answer:
```

A.2 Few-Shot Prompt

A.2.1 FFQA Prompt

```
You are an AI assistant designed
   to answer questions based on
   given passages.
Your task is to provide accurate
   and concise answers to
   questions using only the
   information provided in the
   passage.
If the passage doesn't contain
   enough information to answer
   the question, respond with '
   The passage does not provide
   sufficient information to
   answer this question.'
Do not use any external knowledge
    or make assumptions beyond
   what is explicitly stated in
   the passage. Response should
   be in one line without any
   additional information and
   response in source language.

Passage: {Passage1}
Question: {Question1}
Answer: {Answer1}

Passage: {Passage2}
Question: {Question2}
Answer: {Answer2}

Passage: {Passage3}
Question: {Question3}
Answer: {Answer3}

Passage: {Passage4}
Question: {Question4}
Answer: {Answer4}

Passage: {Passage}
Question: {Question}
Answer: "
```

A.2.2 MCQA Prompt

```
You are an AI assistant designed
   to answer multiple-choice
   questions.
Your task is to select the most
   appropriate answer from the
   given options (A, B, C, D)
   based on the question provided
   .
Analyze the question and options
   carefully before making your
   selection.
Your response should only contain
    the letter of the correct
   option (A, B, C, or D).
If none of the options seem
   correct or if there isn't
   enough information to make a
   selection, respond with '
   Unable to determine the
   correct answer based on the
   given options.'

Passage: {Passage1}
Question: {Question2}
Options:
A) {OptionA1}
B) {OptionB1}
C) {OptionC1}
D) {OptionD1}
Answer: A)

Passage: {Passage2}
Question: {Question2}
Options:
A) {OptionB2}
B) {OptionA2}
C) {OptionC2}
D) {OptionD2}
Answer: B)

Passage: {Passage3}
Question: {Question3}
Options:
A) {OptionC3}
B) {OptionB3}
C) {OptionA3}
D) {OptionD3}
Answer: C)

Passage: {Passage4}
Question: {Question4}
Options:
A) {OptionD4}
B) {OptionB4}
C) {OptionC4}
D) {OptionA4}
Answer: D)

Passage: {Passage}
Question: {Question}
Options:
A) {OptionA}
B) {OptionB}
C) {OptionC}
D) {OptionD}
Answer:
```

B Additional Datasets

B.1 NER

LocalDoc/azerbaijani-ner-dataset[3]

B.2 FFQA

LocalDoc/databricks-dolly-azerbaijan (closed qa)[4]

[3] https://huggingface.co/datasets/LocalDoc/azerbaijani-ner-dataset

[4] https://huggingface.co/datasets/LocalDoc/databricks-dolly-azerbaijan