CUNI and LMU Submission to the MRL 2024 Shared Task on Multi-lingual Multi-task Information Retrieval

Katharina Hämmerl[λ,μ,*] **Andrei Manea**[κ] **Gianluca Vico**[κ]
Jindřich Helcl[κ] **Jindřich Libovický**[κ]

[κ]Faculty of Mathematics and Physics, Charles University, Czech Republic
[λ]Center for Information and Language Processing, LMU Munich Germany
[μ]Munich Center for Machine Learning (MCML), Germany
haemmerl@cis.lmu.de {manea,vico,helcl,libovicky}@ufal.mff.cuni.cz

Abstract

We present the joint CUNI and LMU submission to the MRL 2024 Shared Task on Multi-lingual Multi-task Information Retrieval. The shared task objective was to explore how we can deploy modern methods in NLP in multi-lingual low-resource settings, tested on two sub-tasks: Named-entity recognition and question answering. Our solutions to the subtasks are based on data acquisition and model adaptation. We compare the performance of our submitted systems with the translate-test approach which proved to be the most useful in the previous edition of the shared task. Our results show that using more data as well as fine-tuning recent multilingual pre-trained models leads to considerable improvements over the translate-test baseline. Our code is available at https://github.com/ufal/mrl2024-multilingual-ir-shared-task.

1 Introduction

Over the past few years, large language models (LLMs) have attracted a fair share of attention from the research community. This is mainly caused by the remarkable in-context learning properties these models exhibit, especially in languages where there is plenty of data available (Wei et al., 2022).

Very recently, research advances have shown promising results in low-resource language processing by leveraging LLMs trained primarily on English data (Cahyawijaya et al., 2024; Nguyen et al., 2024, inter alia). Meanwhile, massively multilingual approaches also started to deliver good results (Zaratiana et al., 2024; Üstün et al., 2024). The MRL 2024 Shared Task on Multi-lingual Multi-task Information Retrieval aims to build upon this trend in tasks of named-entity recognition (NER; § 2) and question answering (QA; § 3) for Alsatian, Azerbaijani, Igbo, Indonesian, Turkish, Uzbek, and Yoruba.

In both subtasks, our submissions include fine-tuned mulitlingual models, compared to a translate-test baseline (Helcl and Libovický, 2023).

2 Named Entity Recognition

The goal of the NER subtask was to detect and classify words and phrases into one of three categories: person (PER), organization (ORG), and location (LOC). Unlike the previous year's edition, date and time entities were omitted from the task.

For development, the organizers released validation data in Alsatian, Azerbaijani, Turkish, and Yoruba, each of around 120 sentences.

We experiment with the translate-test approach and compare it with the most recent massively multilingual models (§ 2.1). We collect additional training data for the shared languages (§ 2.2) and fine-tune the best-scoring multilingual model (§ 2.3).

2.1 Baseline Models

Translate-test. Using the label-projection method from Helcl and Libovický (2023), we translate the validation data to English, then test two pre-trained models on them: An English SpaCy pipeline[1] and (English-only) GliNER[2] (Zaratiana et al., 2024). See Table 1, "Translate + Spacy" and "Translate + GliNER" for the respective validation set results.

Multilingual Models. We further test a multilingual baseline model from the UniversalNER project (Mayhew et al., 2024), as well as multilingual GliNER[3] (Zaratiana et al., 2024) on the original validation data. The model from UniversalNER is a version of XLM-R$_{large}$, fine-tuned on all of the project's annotated training data. Multilingual GliNER is an open-type NER model intitialised from mDeBERTa-v3$_{base}$ (He et al., 2023)

[*] Part of KH's work on this paper was done during a research visit to CUNI.

[1] en_core_web_lg
[2] urchade/gliner_large-v2.1; 459M parameters
[3] urchade/gliner_multi-v2.1; 209M parameters

Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), pages 357–364

Method	als	aze	tur	yor	Avg.
Translate + Spacy	43.7	51.7	42.6	52.0	47.5
Translate + GliNER	30.7	48.2	46.9	44.1	42.5
Universal NER	56.9	67.8	62.9	55.0	62.5
Multiling. GliNER	61.5	67.8	63.5	63.0	64.3
↳ + tuning	71.5	69.2	74.2	74.0	72.2

Table 1: Results of the explored methods on the shared task validation data.

Parameter	Value
Learning rate	5×10^{-6}
Weight decay	0.01
Epochs	5
Batch size	16
Warmup ratio	0.1

Table 2: Hyperparameters used for the final tuned GliNER model.

and fine-tuned on Pile-NER (Zhou et al., 2024). The validation set results from these models are also listed in Table 1. Based on these initial results, we select Multilingual GliNER for further tuning.

2.2 Datasets

For fine-tuning Multilingual GliNER, we use a selection of NER datasets in different languages. We found relevant data for all target languages except Alsatian and decided to use Standard German data instead.

MasakhaNER2. Adelani et al. (2022) provide a high-quality NER dataset for 20 African languages. The data includes labels for *person*, *organisation*, *location* and *date* in the BIO format. Since the shared task does not include date labels, we discard those from the data before feeding it to our model. We use the Yoruba (6.8k) and Igbo (7.6k) training splits for the final tuned model. The validation splits (around 1k each) are also used for evaluation during model fine-tuning.

PolyglotNER. PolyglotNER (Al-Rfou et al., 2015) is a large, automatically created NER dataset for 40 languages. It includes labels for *person*, *organisation*, and *location*. We convert the labels to the BIO format before training. We use parts of the Turkish and German subsets in the final tuned model. In order to keep the training data to a similar size as Yoruba and Igbo, we only take around 10k examples for the training itself, and around 1k examples for validation during model fine-tuning.

LocalDoc NER. LocalDoc NER (LocalDoc, 2024) is an extensive collection of Azerbaijani NER data with 24 entity types. Since the shared task data includes only the target entities *person*, *organisation*, and *location*, we discard all other entity types, and transform the data to the BIO format, before feeding the data to our model. Since this leaves us with a somewhat large proportion of "empty" examples (with no labels other than "O"), we then discard such examples with a 50-50 chance. The original dataset includes almost 100k examples, but we only use around 10k examples for training in order to keep a similar proportion of training data as the other languages. We use an additional 1k examples for validation during model fine-tuning.

Additional Datasets. We further experimented with UZNER (Yusufu et al., 2023) and SwissNER (Vamvas et al., 2023) data for Uzbek and Swiss German, respectively, but found that including this data did not noticeably improve performance on the validation languages, so the final tuned model does not include them.

2.3 Model Tuning

We attempt tuning with different combinations of data, different learning rates, weight decay, and number of epochs. Table 2 shows the hyperparameters used in the selected model. Due to the comparatively small size of the base model (209M parameters), and limited training data especially for the smallest sets used, each training run is quite fast: Between one and two hours depending on epochs and data mix, on a single GPU.

2.4 Results

The validation results are presented in Table 1. The fine-tuned GliNER scores the best on all languages in the validation set, on average 8 F_1 points better than the pre-trained version. Multilingual models, even without fine-tuning, significantly outperformed translate-test baselines.

Based on these results, we submitted outputs of the fine-tuned GliNER to the shared task.

Table 3 shows test set results released by the organisers. Although our model is actually outpaced on most of the test languages by the system from McGill, we win on consistency, for an average performance lead of 4.2 F_1 points. Our result on Azerbaijani falls furthest behind, which may indicate that the distribution of the LocalDoc dataset

Team	als	az	ig	tr	yo	Avg.
CUNI	70.4	57.3	**73.9**	77.8	80.5	**71.9**
McGill	**78.9**	**82.1**	9.3	**82.6**	**85.7**	67.7
Ifeoma	0.8	2.0	2.0	4.0	0.8	1.9

Table 3: Results on the NER test set. The value for each language is the F1 metric.

was too different from the shared task set.

3 Question Answering

The goal of this task is to answer questions within a given context in two scenarios: First, select the correct answer from a set of four choices (multiple-choice questions). Second, generate a free-form answer in natural language (open questions).

The organizers provided 200 multiple-choice questions for all languages except Uzbek. All correct options in the development data were labeled as "A". To balance the dataset, we shuffle the ordering of the options in the data and report the performance on this shuffled dataset. Additionally, around 100 single-reference open questions for all languages were provided.

We experiment with LLMs in the zero-shot setup both in the task languages and when translating the test into English (§ 3.1). Then, we collect QA datasets that are available for the task languages (§ 3.2). We use the data to fine-tune the models (§ 3.3). Finally, we experiment with ensembling of the models outputs in the zero-shot setup (§ 3.4).

3.1 Zero-shot LLMs

We select a few multilingual LLMs tested on both the original and translated validation sets: Aya-101 (Üstün et al., 2024) and 4 versions of the LLaMA model (Touvron et al., 2023). Aya-101 is an encoder-decoder model trained in multiple tasks and 101 languages, while LLaMA is a causal language model.

Multiple Choice Questions. For this task, we extract the probability score for each option: "A)," "B)," "C)," or "D)." To do so, we use a prompt consisting of the context, the question, and the answer options. This is followed by "The correct answer is:". This way, we increase the chance that the next generated token is one of the answer letters. We translate this prompt into each task language so that the prompt and the question are in the same language.

We know the next token might not necessarily be in our range, as Wang et al. (2024) state. To overcome this, we use the system prompt: "You are an assistant trained to read the following context and answer the question with one of the options A), B), C) or D).". Upon inspection of the generated text, we found a minimal number of cases (1-2) where the generated answer starts with a different token.

We extract the probabilities of the four tokens corresponding to the options, and re-normalize them with softmax. Then, we choose the option with the maximum score. Another strategy is to generate text using nucleus sampling and extract the first label. This results in slightly lower accuracy for all languages; therefore, we use the probability scores.

Open Questions. For the open-question scenario, we use a different system prompt: "You are an assistant trained to read the following context and provide a succinct, accurate, and clear response in the same language." The user prompt consists only of the context and the question.

We use temperature 0.6, nucleus sampling with top p of 0.9, and maximum new tokens 80.

Translate test. We translate the multiple choice validation set into English using NLLB-200-3.3B (Team et al., 2022) and then use the same mentioned models as a baseline. Long samples are split into sentences with an English SpaCy pipeline[4] to fit the NLLB context size. The translations are then appropriately concatenated to have the context and the questions together. The prompt is in English, while in the multilingual case, it is translated into each input language.

3.2 Datasets

We use additional datasets for multiple-choice questions to fine-tune the LLaMA models and Aya-101. Similarly to the NER task, we use Standard German data instead of Alsatian, but we do not have Azerbaijani data. We also use additional English data. The domain of some datasets is broader than that of this shared task because these datasets can test for knowledge or include multiple-choice sentence completion. We standardize the format of all the datasets, including this shared task dataset: we combine the short text with the question and add

[4] `en_core_web_sm`

the prefixes "A)," "B)," "C)," and "D)" to the four possible answers.

MMLU. MMLU (Hendrycks et al., 2021b) (Hendrycks et al., 2021a) contains English multiple-choice questions testing various branches of knowledge. In contrast to this shared task, it does not contain a separate text with the information to answer the question. Moreover, some samples are about sentence completion, or the context is not always sufficient to answer the question. This set contains 115700 English samples.

AFRIMMLU. AFRIMMLU (Adelani et al., 2024) is a translation of MMLU in several African languages. We use the Igbo and Yoruba splits, which contain 608 examples.

M_MMLU. M_MMLU (Dac Lai et al., 2023) is a machine-translated version of MMLU. We use the Indonesian portion, which contains 14798 samples.

MMLU_TR. M_MMLU (Alhajar, 2024) is a Turkish machine-translated version of MMLU which contains 15263 samples.

Belebele. Belebele (Bandarkar et al., 2024) is a multiple-choice question dataset about reading comprehension. Each sample contains a short text, a question, and four possible answers (from 1 to 4, converted to A, B, C, and D). We use 900 samples for the following languages: Tosk Albanian (language code ALS), German, English, Igbo, Indonesian, Turkish, Uzbek, and Yoruba.

EXAMS. EXAMS (Hardalov et al., 2020) is a dataset that contains high school-level multiple-choice questions. Each sample has a short test, a question, and four possible answers. However, the short text does not answer the question, as the dataset aims to test knowledge. We use 1964 Turkish samples.

QASC. QASC (Khot et al., 2020) is a multiple-choice question dataset about grade school science questions. We use 9060 unique English samples with a short fact, a question, and eight possible answers. To adapt it for this task, we randomly discard four wrong answers from each sample and relabel the remaining ones.

NaijaRC. NaijaRC (Aremu et al., 2024) is a multiple-choice question dataset about reading comprehension. As the dataset for this shared task, NaijaRC contains a short text, a question, and four possible answers. We used 89 Igbo samples and 191 Yoruba samples.

Model	Method	als	aze	ibo	tur	yor	Avg.
LLaMA	score	**83.0**	**83.5**	**88.0**	88.2	**87.5**	**86.0**
	gen.	80.5	81.0	86.5	**89.2**	**87.5**	83.7
Aya 101	score	85.5	**96.0**	**95.0**	88.2	**90.5**	**91.0**
	gen.	**88.0**	95.0	92.5	**89.2**	87.5	90.5

Table 4: Comparison of the *score* versus *generate* (gen.) method in the zero-shot multilingual inference. LLaMA model refers to 3.1 8B version.

3.3 Model Tuning

In the multiple-choice task scenario, we fine-tune the models using Low-Rank Adaptation (LoRA; Hu et al., 2021). We format the data in the same way as in the zero-shot experiments. After the context, questions, and multiple choices, we repeat the correct answer with "The right answer is X):" prepended.

For training the model using LoRA, we set rank r to 64, scaling factor α to 16. We use a dropout of 0.1, with no bias, and we only adapt the attention layers. We tested the fine-tuned models on the open task with the prompt mentioned in Section 3.1.

Table 7 and 8 contains the preliminary results of the test set. These were the only submissions that were publicly listed on Codabench.

Multilingual Fine-tuning. We compile the training dataset from all datasets listed in the previous section, except for EXAMS, which we omit so Turkish is not overrepresented.

Monolingual Fine-tuning. We fine-tune the multilingual models with monolingual data to make a comparison. We train each model for 8 epochs with a learning rate of $2 \cdot 10^{-4}$ and tested on the same language. Since we do not have Azerbaijani data, this language is not included. For Alsatian, we use Standard German and Tosk Albanian (ALS), which was included by accident because of the same unofficial abbreviation as the ISO code for Alsatian.

3.4 Ensembling

For the multiple-choice scenario, we experiment with model ensembling to increase robustness.

Three Models. We combine the scores of our best models: LLaMA 3.0 70B, LLaMA 3.1 70B, and Aya 101. Each model outputs scores for each answer choice. We merge the scores with either

	Model	als	aze	ibo	tur	yor	Avg.
Translate-test	LLaMA 3.0 8B	55.0	80.0	86.0	87.7	74.0	76.5
	LLaMA 3.1 8B	52.0	81.5	87.5	85.1	75.5	76.3
	Aya 101	49.0	79.0	85.0	81.0	73.5	73.5
Multilingual Zero-Shot	LLaMA 3.0 8B	84.5	88.0	91.5	90.8	80.5	87.3
	LLaMA 3.0 70B	92.5	96.5	93.0	95.4	83.5	92.2
	LLaMA 3.1 8B	83.0	83.5	88.0	88.2	87.5	86.0
	LLaMA 3.1 70B	92.5	96.5	93.0	95.4	83.5	92.2
	Aya 101	85.5	96.0	95.0	88.2	90.5	91.0
Multilingual Tuned	LLaMA 3.1 8B	78.0	94.5	89.5	85.6	79.0	85.3
Monolingual Tuned	LLaMA 3.1 8B	85.0	—	95.0	80.5	82.5	—
	Aya 101	85.5	—	91.0	73.8	85.0	—
Three Models	hard	92.0	**99.5**	94.5	**96.4**	92.0	94.9
	soft	**93.0**	**99.5**	**95.5**	**96.4**	**92.5**	**95.4**
Three Prompts	Aya 101	86.5	97.0	94.5	89.7	90.5	91.7
	LLaMA 3.1 70B	92.5	97.5	91.5	**96.4**	89.5	93.5

Table 5: Results for the multiple choice model on the validation set, using the *score* method

	Model	Metric	als	aze	ibo	tur	yor	ind	uzb	Avg.
Multilingual Zero-Shot	LLaMA 3.1 8B	ChrF	27.7	61.7	37.7	52.3	25.3	42.2	49.6	42.3
		RougeL	9.1	55.3	28.7	35.9	15.9	35.5	38.8	31.3
		BERTscore	64.5	83.6	70.3	67.0	66.2	69.7	72.4	70.5
	LLaMA 3.1 70B	ChrF	**32.1**	**69.5**	**57.0**	**53.4**	33.2	41.7	56.9	**49.1**
		RougeL	**22.5**	**70.7**	**43.2**	**46.3**	23.4	36.4	46.4	**41.3**
		BERTscore	**85.0**	**96.1**	**86.5**	**91.1**	82.8	83.5	87.1	**87.5**
	Aya 101	ChrF	22.4	53.2	24.5	42.0	**34.8**	**44.1**	**60.2**	40.2
		RougeL	16.7	52.4	25.0	39.2	**29.1**	**43.6**	**48.1**	36.3
		BERTscore	82.5	91.6	83.5	89.1	**84.8**	**86.5**	**87.8**	86.5
Multilingual Tuned	LLaMA 3.1 8B	ChrF	24.6	34.0	22.0	22.8	17.9	39.8	31.1	27.5
		RougeL	15.2	31.2	15.5	19.3	13.1	33.0	19.2	20.9
		BERTscore	67.9	72.1	65.1	50.3	63.9	72.5	64.7	65.22

Table 6: Results for the open question models on the validation set.

Team	als	az	ig	tr	yo	Avg.
isidoratourni	**0.92**	**0.98**	**0.98**	**0.97**	**0.92**	**0.95**
CUNI	**0.92**	**0.98**	**0.98**	0.96	0.86	0.93
McGill NLP Group	0.78	0.97	0.97	0.82	0.85	0.88

Table 7: Preliminary results of multiple-choice questions leaderboard, extracted from Codabench. The final column is the weighted accuracy.

Team	als	az	ig	tr	yo	Avg.
CUNI	**0.43**	**0.61**	**0.68**	**0.55**	**0.47**	**0.54**
McGill NLP Group	0.42	0.42	0.33	0.4	0.36	0.38

Table 8: Preliminary results of open questions leaderboard, extracted from Codabench. The final column is the weighted average of all the metrics.

hard or *soft* voting. In *hard* voting, we select the choice with the highest score for each model and then choose the final answer with a majority vote. In *soft* voting, we average the scores for each choice and then select the one with the maximum score.

Three Prompts. Since the models produce answers in different formats, we use two additional prompts. In addition to the original, "The correct answer is: ", we add "It is: " and the empty prompt. We translate the prompts into the shared task languages and use them with the model. We average the probabilities for the respective prompts and return the option with the maximum score.

3.5 Results

Table 4 shows the difference between the scoring method and the generated for the multiple-choice tasks. The results were much better with scoring, allowing us more flexibility, such as using *soft* vot-

ing in ensembling models.

After tuning the models, we observe a performance drop in every language. We believe this is due to a domain mismatch between the training and shared task test data. Therefore, we decided to proceed with the zero-shot setup. Table 5 shows the performance on the validation set in the multiple-choice task. For the submission, three models *soft* voting was selected as the best submission.

Table 6 contains the result from the open task, with the best submission being LLaMA 3.1 70B (zero-shot).

4 Conclusions

We presented our submissions to the MRL Shared Task on Multi-lingual Multi-task Information Retrieval. Our methods based on data acquisition and fine-tuning of multilingual pre-trained models achieve good results compared to the translate-test approach, which was the key idea of the winning system from 2023 (Helcl and Libovický, 2023). For NER, we achieved our best results by finetuning state-of-the-art models specifically for the shared task languages and entities. In the QA subtask, we achieved our best results using the LLMs in the zero-shot setup.

Acknowledgments

This work was supported by the Charles University project PRIMUS/23/SCI/023 and SVV project number 260 698.

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.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *Preprint*, arXiv:2406.03368.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30- May 2, 2015*.

Mohamad Alhajar. 2024. mmlu_tr-v0.2. https://huggingface.co/datasets/malhajar/mmlu_tr-v0.2.

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. *Preprint*, arXiv:2308.09768.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jindřich Helcl and Jindřich Libovický. 2023. CUNI submission to MRL 2023 shared task on multi-lingual multi-task information retrieval. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 302–309, Singapore. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*.

LocalDoc. 2024. Azerbaijani NER dataset (Revision 7bf7e0a).

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The multilingual language model for Switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69, Neuchatel, Switzerland. Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7407–7416, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Aizihaierjiang Yusufu, Jiang Liu, 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*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.