Findings of the 2nd Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2024

Francesco Tinner
University of Amsterdam
14497425@uva.nl

Raghav Mantri
New York University
raghav.mantri@nyu.edu

Mammad Hajili
Microsoft
mammadhajili@microsoft.com

Chiamaka Chukwuneke
Lancaster University, UK

Dylan Massey
University of Zurich
dylan.massey@uzh.ch

Benjamin A. Ajibade
University of Alabama
baajibade@crimson.ua.edu

Bilge Deniz Kocak
Villanova University
bkocak1@villanova.edu

Abolade Dawud
Masakhane
aboladedawud@gmail.com

Jonathan Atala
Anglia Ruskin University
Olaatala7@gmail.com

Hale Sirin
Johns Hopkins University
hsirin1@jhu.edu

Kayode Olaleye
University of Pretoria
kayode.olaleye@up.ac.za

Anar Rzayev
KAIST
rzayev.anar1@kaist.ac.kr

David Adelani
McGill University
david.adelani@mcgill.ca

Duygu Ataman
New York University
ataman@nyu.edu

Abstract

Large language models (LLMs) demonstrate exceptional proficiency in both the comprehension and generation of textual data, particularly in English, a language for which extensive public benchmarks have been established across a wide range of natural language processing (NLP) tasks. Nonetheless, their performance in multilingual contexts and specialized domains remains less rigorously validated, raising questions about their reliability and generalizability across linguistically diverse and domain-specific settings. The second edition of the Shared Task on Multilingual Multitask Information Retrieval aims to provide a comprehensive and inclusive multilingual evaluation benchmark which aids assessing the ability of multilingual LLMs to capture logical, factual, or causal relationships within lengthy text contexts and generate language under sparse settings, particularly in scenarios with under-resourced languages. The shared task consists of two subtasks crucial to information retrieval: Named entity recognition (NER) and reading comprehension (RC), in 7 data-scarce languages: Azerbaijani, Swiss German, Turkish and Yorùbá, which previously lacked annotated resources in information retrieval tasks. This year specifally focus on the multiple-choice question answering evaluation setting which provides a more objective setting for comparing different methods across languages.

1 Introduction

Recent advancements in organizing online knowledge facilitated by Large Language Models (LLMs) have fundamentally reshaped the way we approach information retrieval. This functionality creates exciting potential for new applications for education and media supporting seamless access to information on diverse subjects. However, this functionality is largely to limited in high-resourced languages, preventing equal access to potential applications in

Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), pages 365–376
November 16, 2024 ©2024 Association for Computational Linguistics

many under-resourced or studied languages across the world (Yong et al., 2023). Recently, initiatives for creating standardized benchmarks for evaluating natural language processing (NLP) systems in a more linguistically inclusive setting had been proposed by corpora like XTREME (Hu et al., 2020) and XTREME-UP (Ruder et al., 2023). Although these data sets bring together large multilingual corpora they lack in generative human prepared data related to information access.

The 2nd Shared Task on Multi-lingual Multi-task Information Retrieval (MMIR), provides a benchmark for evaluating multi-lingual large language models (LLMs) in terms of their applicability for information retrieval in various under-resourced and typologically diverse languages. Purely constructed using human annotated data consisting of examples of reading comprehension questions and named entity recognition in various context and languages, MMIR benchmark presents a challenging new task for testing and improving LLMs. As evaluation resource we use Wikipedia which we find representative of the inclusion of languages online. We pick five languages with varying degrees of resources and linguistic typology from three different language families: Azerbaijani and Turkish (Turkic), Igbo and Yoruba, (Niger-Congo) and Swiss German (Germanic), and produce annotations in two tasks crucial for IR: named entity recognition (NER) and reading comprehension (RC). We present our data curation and annotation process as well as the findings of the evaluation in the resulting benchmark including prominent LLMs trained on multi-lingual multi-task settings: LLAMA (Dubey et al., 2024), Aya (Üstün et al., 2024) and Gemini (Reid et al., 2024). Extending the data sets and competition from 2023, this year's edition allowed submissions both in open-ended and multiple-choice question answering to allow a more fine-grained and objective analysis. we received 3 submissions in the multiple-choice and 2 submissions in the open-ended RC tasks. The NER task also received 2 submissions. We provide more details on the data sets and a comparison of competing systems.

2 Tasks

MMIR shared task provides a multi-task evaluation format to assess information retrieval capabilities of LLMs in terms of two tasks: named entity recognition (NER) and reading comprehension (RC).

Narendrabhai Damodardas Modi ni Mínśítà àgbà India kẹrìnlá àti mínísítà àgbà tí India lọ́wọ́ lọ́wọ́ lati ọdun 2014. O jẹ oloselu kan lati Bharatiya Janata Party, agbari-iṣẹ oluyọọda ara ilu Hindu kan. Oun ni Prime Minister akọkọ ni ita ti Ile-igbimọjọ ti Orilẹ-ede India lati ṣegun awọn ofin itẹlera meji pẹlu opoju to kun ati ekeji lati pari diẹ sii ju ọdun marun ni ọfiisi lẹhin Atal Bihari Vajpayee.

Table 1: Example of named entities in Yorùbá language. PER, LOC, and ORG are in colours red, green, and blue respectively. We make use of Label Studio for annotation (Tkachenko et al., 2020-2022).

2.1 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a classification task that identifies text phrases referring to specific entities or categories (e.g., dates, names of people, organizations, or locations). This is essential for systems handling entity look-ups for tasks like knowledge verification, spell-checking, or localization. Our training data in the shared task relies on the XTREME-UP dataset (Ruder et al., 2023) which is the most comprehensive data set that combines annotated data from MasakhaNER (Adelani et al., 2021b) and MasakhaNER 2.0 (Adelani et al., 2022) in a wide range of under-resourced languages including: Amharic, Ghomálá, Bambara, Ewe, Hausa, Igbo, (Lu)Ganda, (Dho)Luo, Mossi (Mooré), Nyanja (Chichewa), Nigerian Pidgin, Kinyarwanda, Shona, Swahili, Tswana (Setswana), Twi, Wolof, Xhosa, Yorùbá and Zulu.

The objective of the system is to tag the named entities in a given text, either as a person (PER), organization (ORG), or location (LOC). The NER data this year remains as same with 2023.

2.2 Reading Comprehension (RC)

RC is a challenging task often requiring different levels of natural language comprehension and reasoning for answering a given question based on a span of information distributed across a given context. Here we focus on the information-seeking scenario where questions can be asked without knowing the answer. It is the system's task to locate a suitable answer passage (if any). We provide 4 options for each question, where the systems are asked to pick one of the 4 answers as the correct one. Examples can be found in Table 2.

Information-seeking question-answer pairs typically display limited lexical and morphosyntactic overlap between the question and answer, as they

Context	Question	Options
Zaqatala" qəzeti redaksiyası 1923-cü ilin mart ayından fəaliyyətə başlamışdır. İlk əvvəllər "Zaqatala kəndlisi" adlanan qəzet sonralar "Kolxozun səsi", "Bolşevik kolxozu uğrunda", "Qırmızı bayraq" və s. başlıqlarla fəaliyyət göstərmişdir. 1991-ci ilin oktyabr ayından isə "Zaqatala" adı ilə fəaliyyətini davam etdirir. Hal-hazırda "Zaqatala" qəzeti redaksiyasında 5 nəfər çalışır.	İndi qəzetdə neçə nəfər çalışır?	**(1) İndi "Zaqatala" qəzetində 5 nəfər işləyir.** (2) "Zaqatala" qəzetinin hal-hazırki işçi sayı 7-dir. (3) İndi "Zaqatala" qəzetində 20 nəfər işləyir. (4) "Zaqatala" qəzetinin işçilərinin sayı bilinmir.
Noch de jüngere Version isch de Eurytos vom Herakles töödt woore. Us Raach nämmli, well de em sini Töchter Iole nöd hett wöle gee, hett er d Stadt Oichalia eroberet, de Eurytos und all sini Söö töödt und d Iole graubt.	Was isch de Grund gsi für di tötig vom Eurytos?	(1) Will de Eurytos de Herakles ermordet het. (2) Will das eh jüngeri Version vo de Gschicht isch gsi. **(3) Will de Eurytos am Herakles nöd sis Töchterli - d Iole - het welle geh.** (4) Will de Eurytos vom Herakles töödt woore isch.
A bi Aisha Adamu Augie ni Zaria, Ipinle Kaduna, Nigeria, Augie-Kuta jẹ ọmọbinrin oloogbe Senator Adamu Baba Augie (oloselu / olugbohunsafefe), ati Onidajọ Amina Augie (JSC). Augie-Kuta bere si ni nifẹ si fọtoyiya nigbati baba rẹ fun u ni kamẹra ni ọdọ.	Ki ni ibaṣepọ to wa laarin Aisha Adamu Augie ati Senator Adamu Baba Augie?	(1) Aisha Adamu j ìyàwó Senator Adamu Baba Augie **(2) Aisha Adamu jẹ ọmọ fun Senator Adamu Baba Augie** (3) Aisha Adamu jẹ àbúrò Senator Adamu Baba Augie (4) Aisha Adamu j ọbàkan Senator Adamu Baba Augie

Table 2: Examples from the RC validation data in different languages. Correct answers indicated in **bold.**

Language	Family
Azerbaijani	Turkic
Igbo	Niger-Congo
Swiss German	Indo-European
Turkish	Turkic
Yorùbá	Niger-Congo

Table 3: List of languages and language families.

are composed independently. This makes them ideal for evaluating languages with diverse typological features. In this task, the system receives a question, title, and passage, and must either provide the correct answer or indicate that no answer is present in the passage. Currently, the XTREME-UP benchmark includes data in Indonesian, Bengali, Swahili, and Telugu (Ruder et al., 2023), requiring competing systems to infer information from different language annotations. Our benchmark also contains correct text answers from 2023 edition (Tinner et al., 2023) for open-ended RC evaluation. This year we extend the benchmark in four languages with multiple-choice RC annotations. We allow both types of output for submission to the shared task.

3 Languages

Table 3 provides an overview of the variety in our data set in terms of language families.

3.1 Azerbaijani (AZ)

Azerbaijani, part of the Turkic language family, is mainly spoken in Azerbaijan and Iran. It shares many linguistic traits with other Turkic languages, particularly those in the Western Oghuz group like Turkish, Gagauz, and Turkmen. Azerbaijani features agglutinative morphology, uses a Subject-Object-Verb (SOV) word order, and lacks gender in its grammar. In Azerbaijan, the Latin script has been used since 1991, while Iranian Azerbaijanis use the Arabic script. This study's data preparation focuses on texts in the Latin script.

3.2 Igbo (IG)

Igbo, part of the Benue-Congo group within the Niger-Congo language family, is spoken by over 27 million people, primarily in southeastern Nigeria, as well as parts of Equatorial Guinea and Cameroon. While there are several dialects, Central Igbo, standardized in 1962, is the most widely used. Standard Igbo includes 28 consonants and 8 vowels, with two tones: high (marked by an acute accent) and low (marked by a grave accent), though these tones are usually not represented in writing. Igbo has been featured in various language benchmarks, such as MasakhaNER (Adelani et al., 2021b, 2022), AfriQA (Ogundepo et al., 2023), MasakhaPOS (Dione et al., 2023), AfriSenti (Muhammad et al., 2023).

3.3 Swiss German (ALS)

Swiss German, part of the Alemannic dialects within the Germanic language family, poses a significant challenge for multilingual NLP due to its non-standardized nature. It varies greatly in lexicon, phonetics, morphology, and syntax, with no official orthography. Individuals often write words based on their interpretation of phonetics, resulting in inconsistent spellings. Unlike Standard German, Swiss German is not an official language of Switzerland and is primarily used in spoken or informal contexts, with formal writing done in Standard German. Due to this, textual resources are scarce. A notable exception is a text corpus for PoS tagging, compiled from sources like Alemannic Wikipedia, novels, reports, and articles (Hollenstein and Aepli, 2014). Further resources are only available in spoken format, including the SDS-200 corpus (Plüss et al., 2022), Swiss Parliaments Corpus (Plüss et al., 2020), SwissDial corpus (Dogan-Schönberger et al., 2021), Radio Rottu Oberwallis corpus (Garner et al., 2014), ArchiMob corpus (Samardžić et al., 2016), SST4SG-350 (Plüss et al., 2023).

3.4 Turkish (TR)

Turkish, the most widely-resourced language in the Turkic family, is known for its agglutinative morphology and Subject-Object-Verb (SOV) word order. It has no grammatical gender but includes a complex case system. Verbs are inflected to show tense, mood, and person, while personal pronouns are used for person reference. Key linguistic features include vowel harmony, palatalized consonants, and phonemic vowel length, which influences word meaning. Turkish lacks definite or indefinite articles, relying on context for clarity. Despite its uniqueness compared to Indo-European languages, its use of the Latin script allows for easier comparisons. Corpus studies in Turkish include plenty monolingual (Aksan et al., 2012) and parallel resources (Tyers and Alperen, 2010; Cettolo et al., 2012; Ataman, 2018). Turkish NLP resources include many inclusive tree banks, such as for Universal Dependencies (Sulubacak et al., 2016; Sulubacak and Eryiğit, 2018), semantic parsing (Şahin and Adalı, 2018) and a WordNET (Ehsani et al., 2018). It is also included in prominently used public multilingual benchmarks including the mc4 corpus (Raffel et al., 2019), and it is recognized in benchmarks, such as for machine translation (Cettolo et al., 2013; Bojar et al., 2017) and morphological analysis (Pimentel et al., 2021). There are also annotated resources for Turkish which were created through automatic annotation using label transfer from other languages or translating existing resources, in tasks including natural language inference (Conneau et al., 2018), NER (Sahin et al., 2017), and summarization (Scialom et al., 2020).

		# Sentences/ # Passages		# Tokens	
Lang	**Task**	Val	Test	Val	Test
AZ	NER	126	124	7,774	8,200
	RC-OE	202	291	13,268	25,487
	RC-MC	202	291	16,147	31,447
IG	NER	711	143	54,526	11,668
	RC-OE	202	748	15,620	58,963
	RC-MC	202	748	21,987	79,761
ALS	NER	130	166	8,761	11,610
	RC-OE	202	651	16,949	50,045
	RC-MC	202	651	21,113	58,182
TR	NER	113	151	7,375	11,736
	RC-OE	197	148	16,336	12,384
	RC-MC	197	148	22,059	16,169
YO	NER	100	303	4,166	11,490
	RC-OE	202	673	20,497	67,816
	RC-MC	202	673	22,891	79,529

Table 4: Dataset statistics for the validation and test splits. NER annotations are at the sentence level while RC questions include passages and questions related to the passage. RC-MC denote the multiple-choice setting where the question is accompanied with 4 potential answers for systems to pick the correct answer.

3.5 Yorùbá (YO)

Yorùbá part of the Volta-Niger subgroup of the Niger-Congo language family, is spoken by over 45 million people, primarily in southwestern Nigeria, as well as in Benin and Togo. It ranks among the top five most spoken African languages, after Nigerian Pidgin, Swahili, Hausa, and Amharic (Eberhard et al., 2021). Yorùbá makes use of the Latin script with modified alphabet: it omits the letters "c,q,v,x,z" and adds "ẹ, gb, ọ, ṣ". The language is tonal, the tones includes high, low, and neutral. The high (as in à) and low (as in á) tones are indicated when writing texts in the language. The tones are important for the correct understanding and pronunciation of the words in Yorùbá. Despite the importance of the tones, many texts written online do not support the writing of the tonal marks, and this may pose a challenge on some downstream NLP applications e.g. machine translation (Adelani et al., 2021a) and text-to-speech (Ogunremi et al., 2023).

4 Data Preparation

The textual data for the generative task are based on Wikimedia downloads[1]. RC annotations are prepared by sampling articles, splitting into paragraph-wise for question and answer annotations. In the extension of the benchmark this year, we annotate additional questions and wrong answer options for creating the multiple-choice QA setting (Tinner et al., 2023). For the NE annotation, we ensure we sample only biographical articles and also only include articles available in all six languages.
We use Label Studio for RC and NER annotation (Tkachenko et al., 2020-2022) with the tag set (Person (PER), Organization (ORG), Location (LOC)) and ensure an annotation overlap of 2% for NER. The question-answer pairs were always produced from two separate annotators. We recruited two annotators per language, for IG and TR respectively four annotators contributed, and five persons annotated YO. The resulting data statistics for the validation and test splits can be found in Table 4. The scripts used to obtain the data, as well as pre- and post-processing methods required to create and export Label Studio annotation projects is included in this GitHub repository [2].

5 Experimental Methodology

5.1 Baseline Systems

GPT-4 OpenAI (2023) is a large-scale, multimodal AI model capable of processing both text and image inputs to generate text outputs. GPT-4 achieves human-like performance on various professional and academic benchmarks. It is a

[1] https://dumps.wikimedia.org/
[2] https://github.com/Fenerator/wikiDataProcessingForQAandNER

Transformer-based model, pre-trained to predict the next word in a sequence. A post-training alignment phase enhances its factual accuracy and ensures it behaves according to specific guidelines. Key to its development was creating infrastructure and optimization methods that scale reliably. The instruction training is based on Reinforcement Learning from Human Feedback (RLHF), similar to InstructGPT (Ouyang et al., 2022).

Gemini-1.5 Pro (Reid et al., 2024) is a mid-size multimodal model optimized for scalability across various tasks, performing on par with the 1.0 Ultra, the largest model to date. It introduces a breakthrough feature in long-context understanding, with a standard 128,000 token context window. Built on cutting-edge research in Transformer and Mixture of Experts (MoE) architecture, Gemini 1.5 uses multiple smaller "expert" neural networks instead of a single large one, enhancing efficiency and performance.

LLAMA-3.2 (Touvron et al., 2023) is a set of large language models (LLMs) that have been pre-trained and fine-tuned, with 1B and 3B models handling multilingual text only, while the 11B and 90B models accept both text and image inputs and produce text outputs.

Claude 3.5 SonnetV2 is an AI language model developed by Anthropic, designed to handle complex tasks and conversations while prioritizing user safety and ethical AI use. It is named after Claude Shannon, a pioneer in information theory. The model is built with a focus on creating helpful, honest, and harmless interactions, with an emphasis on reducing biased or harmful outputs. Its architecture supports advanced reasoning, summarization, and in-depth conversations, making it ideal for a wide range of applications.

	Prompt Template
mT0	<CONTEXT> <QUESTION>
GPT-4	I will provide you with a passage and a question, please provide a precise answer Passage: <CONTEXT> Question: <QUESTION>

Table 5: Zero-shot prompt template used to obtain open-ended answers from the systems.

	Prompt Template
mT0	<CONTEXT> <QUESTION>
GPT-4	I will provide you with a passage and a question, please provide a precise answer Passage: <CONTEXT> Question: <QUESTION> Answers: <A> ... <B> ... <C> ... <D> ...

Table 6: Zero-shot prompt template used to obtain answers in the multiple-choice setting.

5.2 Evaluation

We evaluate and report results in the generative task using ROGUE-L (Lin and Hovy, 2003), chrF (Popović, 2015), chrF+, chrF++ (Popović, 2017), and BERTScore (Zhang et al., 2019) F1 computed with RoBERTaBase (Liu et al., 2019) [3] embeddings. Implementation is based on HuggingFace's evaluate library[4]. Overall performance in the NER task is computed in terms of precision, recall and F-1 scores using the CoNLL Evaluation Scripts[5], implemented in accordance with (Tjong Kim Sang and Buchholz, 2000). We obtain a final score per task and system by weighting the performance per language inversely by the total number of tokens in the test sets per language.

5.3 Submissions

The shared task received five submissions in the NER task, including `CUNI-LMU` (Charles University and LMU Munich) and `McGill` (McGill University) with system descriptions, and three submissions without descriptions, labeled as (`Ifeoma`, `Omkar`, `SandboxAQ`. RC task received three submissions in the multiple-choice QA subtask (RC-MC), from McGill, SandboxAQ and CUNI, and two submissions in the open-ended RC task by CUNI and McGill (RC-OE).

6 Results

We evaluate the overall system performance on the generative task using automatic metrics weighted by the number of articles in the test set containing individual context used for answering the RC questions Table 7 and Table 9. Detailed results per

[3] https://huggingface.co/roberta-base
[4] https://github.com/huggingface/evaluate
[5] https://github.com/sighsmile/conlleval

System	ChrF	ChrF+	ChrF++	RougeL	BERT F1
Claude 3.5 SonnetV2	0.51	0.50	0.47	0.42	0.89
GPT-4	0.45	0.44	0.42	0.36	0.87
Gemini 1.5 Pro	0.42	0.41	0.38	0.40	0.86
Llama 3.2 90B	0.45	0.43	0.41	0.41	0.87
CUNI	**0.48**	**0.46**	**0.45**	**0.42**	**0.88**
McGill	0.33	0.32	0.31	0.36	0.84

Table 7: RC-OE system evaluation. Results indicate weighted average of the metrics over 6 languages. Results are weighted by the number of paragraphs in the testset.

system and language for the open-ended RC task are presented in Table 8. We also present NER results for the system submission in Table 10.

NER The winning system in the NER task is **McGill University** system which deploys an ensemble of XLM-R-Large (Conneau et al., 2020), AfroXLMR (Alabi et al., 2022), and AfroXLMR-76L (Adelani et al., 2024) models fine-tuned on the collection of NER data sets, if we consider the median performance, winning 4 (out of the 5 languages).

RC-OE The RC-OE task is a competitive challenge and both McGill and CUNI, although CUNI has a slightly better performance. In this case, McGill system is comprised of fine-tuned mt5-large (Xue et al., 2021) and AfriTeVA V2 large (Oladipo et al., 2023) models, fine-tuned as ensemble on the publicly available multilingual QA data sets. CUNI system , on the other hand, uses an ensemble of LLAMA models and Aya-101 (Üstün et al., 2024). In the overall evaluation, we find **CUNI** system performs best across languages.

RC-MC The winning team for the multi-choice QA is **SandboxAQ** achieving an average performance of 95% accuracy score. The performance of the CUNI team is competive with only -2.0 point less than that of the winner. On the otherhand, McGill team came third with worse overall result especially for ALS.

7 Conclusion and Future Work

We presented a new multi-lingual multi-task benchmark on information retrieval from Wikipedia in five languages from typologically-diverse and low-resourced language families in the open-ended or multiple-choice QA and NER tasks. We organized a shared task to call for system development on this challenging benchmark where we conducted a detailed analysis on how state-of-the-art LLMs perform in language understanding and generation under low-resourced settings. In addition to finding strong evidence on fall backs in both understanding and generation capabilities of LLMs in low-resourced languages, we also find it crucial to invest in better automatic evaluation metrics for generation in different languages. While we do not find this task to be solved, we plan to keep the competition open and promote more investment into the progress of information retrieval for languages with non-prominent and low-resourced characteristics.

Limitations

We have presented a multilingual evaluation benchmark for information retrieval which was created relying on Wikipedia articles in different languages. Using Wikipedia has inherent limitations such as limitations in variety of content and styles across languages making it challenging to ensure a uniform difficulty level for comprehension questions. Additionally, relying solely on Wikipedia may introduce biases, as certain languages might have more comprehensive or detailed articles than others. Moreover, evaluating language models on Wikipedia-centric benchmarks may not fully reflect their generalization abilities, as the models might excel at leveraging the more structured and well-formulated information found on Wikipedia but may struggle more with more diverse and unstructured text from other sources. These limitations underscore the need for diverse and contextually rich benchmarks to provide a comprehensive assessment of LLMs across multiple languages.

Ethics Statement

All annotators were provided with clear instructions and guidelines to ensure the responsible and unbiased annotation of the data. We ensured eth-

System	Language	ChrF	ChrF+	ChrF++	RougeL	BERT F1
CUNI	ALS	0.37	0.37	0.34	0.24	0.85
CUNI	AZ	0.55	0.55	0.52	0.51	0.92
CUNI	IG	0.63	0.63	0.61	0.62	0.91
CUNI	TR	0.48	0.48	0.45	0.43	0.90
CUNI	YO	0.38	0.38	0.36	0.35	0.86
McGill	ALS	0.32	0.31	0.30	0.32	0.84
McGill	AZ	0.29	0.27	0.26	0.33	0.85
McGill	IG	0.35	0.35	0.34	0.39	0.83
McGill	TR	0.24	0.24	0.23	0.26	0.83
McGill	YO	0.34	0.34	0.33	0.39	0.84
Claude 3.5 SonnetV2	ALS	0.33	0.34	0.31	0.20	0.84
Claude 3.5 SonnetV2	AZ	0.59	0.58	0.55	0.50	0.91
Claude 3.5 SonnetV2	IG	0.68	0.68	0.66	0.65	0.92
Claude 3.5 SonnetV2	TR	0.51	0.51	0.47	0.41	0.89
Claude 3.5 SonnetV2	YO	0.42	0.41	0.39	0.36	0.86
Gemini 1.5 Pro	ALS	0.36	0.35	0.32	0.29	0.84
Gemini 1.5 Pro	AZ	0.51	0.50	0.47	0.48	0.90
Gemini 1.5 Pro	IG	0.45	0.44	0.42	0.48	0.87
Gemini 1.5 Pro	TR	0.42	0.41	0.37	0.35	0.87
Gemini 1.5 Pro	YO	0.38	0.37	0.35	0.36	0.86
Llama 3.2 90B	ALS	0.41	0.40	0.37	0.32	0.86
Llama 3.2 90B	AZ	0.52	0.51	0.48	0.49	0.91
Llama 3.2 90B	IG	0.45	0.45	0.44	0.48	0.86
Llama 3.2 90B	TR	0.47	0.46	0.43	0.42	0.90
Llama 3.2 90B	YO	0.44	0.43	0.41	0.43	0.87

Table 8: RC-OE system evaluations for all languages.

ical practices by providing clear guidelines and obtaining informed consent. We appreciate their contributions, and ethical treatment remains a key focus in our research.

Acknowledgements

We thank Google Deepmind Gemini Academic Program for the compute provided for conducting our analysis. We also thank HumanSignal for providing us access to Label Studio's Enterprise version which allowed us execute the large-scale collaboration to perform human annotations in multiple tasks. Lastly, we thank Damilola Oluwaseun Oloyede, Nagihan Lule, Ece Yurtseven, Mehmet Can Dadan, and Saliha Muradoglu for their contributions to the dataset creation.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, et al. 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.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–

System	ALS	AZ	IG	TR	YO	Avg.
SandboxAQ	92.0	98.0	98.0	97.0	92.0	**95.0**
CUNI	92.0	98.0	98.0	96.0	86.0	93.0
McGill	78.0	97.0	82.0	97.0	85.0	88.0
Claude 3.5 SonnetV2	91.0	98.0	95.0	95.0	92.0	94.0
Gemini 1.5 Pro	91.0	96.0	96.0	96.0	90.0	93.0
Llama 3.2 90B	91.0	97.0	96.0	95.0	89.0	93.0

Table 9: RC-MC system evaluation. Results indicate weighted average of the metrics over 5 languages. Results are weighted by the number of paragraphs in the test set.

System	ALS			AZ			IG		
	pre	**rec**	**F1**	**pre**	**rec**	**F1**	**pre**	**rec**	**F1**
CUNI	77.07	64.74	70.37	69.88	49.49	57.31	69.88	**79.86**	**73.97**
Ifeoma	65	1.18	0.84	1.6	2.75	2.02	1.74	2.44	2.03
McGill	**81.83**	**76.15**	**78.89**	**78.93**	**85.43**	**82.05**	**97.3**	4.86	9.27
SandboxAQ	65.8	48.6	55.9	63.7	42.6	51	51.3	39.7	44.8
Omkar	1	1.3	1.1	2.1	3.03	2.48	-	-	-

System	TR			YO			Avg	Med
	pre	**rec**	**F1**	**pre**	**rec**	**F1**		
CUNI	**85.38**	71.46	77.8	78.61	82.55	80.53	**71.996**	73.97
Ifeoma	3.04	5.91	4.02	0.69	1	0.82	1.946	2.02
McGill	84.19	**81.12**	**82.62**	**85.81**	**85.56**	**85.69**	67.704	**82.05**
SandboxAQ	62.1	44.7	52.0	-	-	-	50.925	51.5
Omkar	3.8	5.5	4.5	1.7	1.7	1.7	2.445	2.09

Table 10: Test results for NER. Averages are weighted by number of tokens per language. Best results are in bold. *Avg: Average. Med: Median.*

English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021b. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Yesim Aksan, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk Demirhan, Hakan Yilmazer, Gülsüm Atasoy, Seda Öz, Ipek Yildiz, et al. 2012. Construction of the turkish national corpus (tnc). In *LREC*, pages 3223–3227.

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 Ataman. 2018. Bianet: A parallel news corpus in turkish, kurdish and english. In *LREC 2018 Workshop*, page 14.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, pages 261–268.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa

Bentivogli, and Marcello Federico. 2013. Report on the 10th iwslt evaluation campaign. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. Ethnologue: Languages of the world. twenty-third edition.

Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz. 2018. Constructing a wordnet for turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3):1–15.

Philip N. Garner, David Imseng, and Thomas Meyer. 2014. Automatic speech recognition and translation of a swiss german dialect: Walliserdeutsch. In *Interspeech*.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Djouhra Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Said Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino D'ario M'ario Ant'onio Ali, Davis C. Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Rabiu Gwadabe, Samuel Rutunda, Tadesse Destaw Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. *ArXiv*, abs/2302.08956.

Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Njoroge Kahira, Shamsuddeen H. Muhammad, Akintunde Oladipo, Abraham Toluwase Owodunni, Atnafu Lambebo Tonja, Iyanuoluwa Shode, Akari Asai, Tunde Oluwaseyi Ajayi, Clemencia Siro, Steven Arthur, Mofetoluwa Adeyemi, Orevaoghene Ahia, Aremu Anuoluwapo, Oyinkansola Awosan, Chiamaka Chukwuneke, Bernard Opoku, Awokoya Ayodele, Verrah Otiende, Christine Mwase, Boyd Sinkala, Andre Niyongabo Rubungo, Daniel A. Ajisafe, Emeka Felix Onwuegbuzia, Habib Mbow, Emile Niyomutabazi, Eunice

Mukonde, Falalu Ibrahim Lawan, Ibrahim Said Ahmad, Jesujoba O. Alabi, Martin Namukombo, Mbonu Chinedu, Mofya Phiri, Neo Putini, Ndumiso Mngoma, Priscilla A. Amuok, Ruqayya Nasir Iro, and Sonia Adhiambo. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages.

Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani. 2023. Ìròyìnspeech: A multi-purpose yorùbá speech corpus.

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.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Tiago Pimentel, Maria Ryskina, Sabrina J Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, et al. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. *ArXiv*, abs/2010.02810.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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. *arXiv e-prints*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, et al. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint arXiv:2305.11938*.

Bahadir Sahin, Mustafa Tolga Eren, Caglar Tirkaz, Ozan Sonmez, and Eray Yildiz. 2017. English/turkish wikipedia named-entity recognition and text categorization dataset. *Mendeley Data, V1*.

Gözde Gül Şahin and Eşref Adalı. 2018. Annotation of semantic roles for the turkish proposition bank. *Language Resources and Evaluation*, 52:673–706.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067. Association for Computational Linguistics.

Umut Sulubacak and Gülşen Eryiğit. 2018. Implementing universal dependency, morphology, and multiword expression annotation standards for turkish language processing. *Turkish Journal of Electrical Engineering and Computer Sciences*, 26(3):1662–1672.

Umut Sulubacak, Memduh Gökırmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal dependencies for turkish. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3444–3454.

Francesco Tinner, David Ifeoluwa Adelani, Chris Emezue, Mammad Hajili, Omer Goldman, Muhammad Farid Adilazuarda, Muhammad Dehan Al Kautsar, Aziza Mirsaidova, Müge Kural, Dylan Massey, et al. 2023. Findings of the 1st shared task on multilingual multi-task information retrieval at mrl 2023. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 310–323.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Francis M Tyers and Murat Serdar Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC workshop on exploitation of multilingual resources and tools for Central and (South-) Eastern European Languages*, pages 49–53.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

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.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.