Findings of the 1st Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2023

Francesco Tinner
University of Amsterdam
14497425@uva.nl

David Ifeoluwa Adelani
University College London
d.adelani@ucl.ac.uk

Chris Emezue
TU Munich
chris.emezue@gmail.com

Mammad Hajili
Microsoft
mammadhajili@microsoft.com

Omer Goldman
Bar-Ilan University
omer.goldman@gmail.com

Muhammad Farid Adilazuarda
Institut Teknologi Bandung
faridlazuarda@gmail.com

Muhammad Dehan Al Kautsar
Institut Teknologi Bandung
faridlazuarda@gmail.com

Aziza Mirsaidova
Northwestern University
azizakhon@u.northwestern.edu

Müge Kural
Koç University
mugekural@ku.edu.tr

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

Chiamaka Chukwuneke
Lancaster University, UK
chukwunekechiamaka3@gmail.com

Chinedu Mbonu
Nnamdi Azikiwe University
ce.mbonu@unizik.edu.ng

Damilola Oluwaseun Oloyede
Federal University of Agriculture, Abeokuta
oloyededo.19@student.funaab.edu.ng

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

Jonathan Atala
Anglia Ruskin University
Olaatala7@gmail.com

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

Saksham Bassi
New York University
sakshambassi@nyu.edu

Rahul Aralikatte
MILA
rahul.aralikatte@mila.quebec

Najoung Kim
Boston University
najoung@bu.edu

Duygu Ataman
New York University
ataman@nyu.edu

Abstract

Large language models (LLMs) excel in language understanding and generation, especially in English which has ample public benchmarks for various natural language processing (NLP) tasks. Nevertheless, their reliability across different languages and domains remains uncertain. Our new shared task introduces a novel benchmark to assess the ability of multilingual LLMs to comprehend and produce language under sparse settings, particularly in scenarios with under-resourced languages, with an emphasis on the ability to capture logical, factual, or causal relationships within lengthy text contexts. 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, Igbo, Indonesian, Swiss German, Turkish, Uzbek and Yorùbá, which previously lacked annotated resources in information retrieval tasks. Our evaluation of leading LLMs reveals that, despite their competitive performance, they still have notable weaknesses such as producing output in the non-target language or providing counterfactual information that cannot be inferred from the context. As more advanced models emerge, the benchmark will remain essential for supporting fairness and applicability in information retrieval systems.

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 310–323
December 7, 2023

1 Introduction

Access to information on diverse subjects, recent events, or historical occurrences is of paramount significance in bolstering educational, media, and economic applications. Recent advancements in organizing online knowledge facilitated by Large Language Models (LLMs) have fundamentally reshaped the way we approach information retrieval. Extensive analysis of models have shown promising capabilities in competitive natural language processing (NLP) tasks, such as question answering (Mao et al., 2023), machine translation (Garcia and Firat, 2022; Hendy et al., 2023), and different types of reasoning (Zhou et al., 2021; Wei et al., 2022; Liu et al., 2023).

LLMs, or foundation models, are typically trained on extensive multilingual data sets, thereby enhancing their accessibility across a spectrum of languages (Floridi and Chiriatti, 2020; Touvron et al., 2023a; Muennighoff et al., 2022; Anil et al., 2023). However, this performance is limited in low-resources languages which lack representation in the public space (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.

By organizing the 1st Shared Task on Multilingual Multi-task Information Retrieval (MMIR), we aim to provide a common means where multilingual LLMs can be evaluated in terms of their applicability and fairness in providing access to users speaking languages from different regions across the world. As the evaluation resource we use Wikipedia which we find representative of the inclusion of languages online. We pick 7 languages with varying degrees of resources and linguistic typology from 4 different language families: Azerbaijani, Turkish and Uzbek (Turkic), Igbo and Yoruba, (Niger-Congo), Indonesian (Austronesian), 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: MT-0 (Muennighoff et al., 2022) and GPT-4 (OpenAI, 2023a), in addition to the system submissions. We also release this benchmark on CodaBench (Xu et al., 2022), where we provide a possibility to obtain the test sets and evaluate future submissions[1] until MRL 2024 .

2 Task Description

With the advancement of language models accessing and processing vast amounts of information in different formats and languages, it has become of great importance to be able to assess their capabilities to access and provide the right information useful to different audiences. In this shared task, we provide a multi-task evaluation format that assesses information retrieval capabilities of language models in terms of two subtasks: named entity recognition (NER) and Reading Comprehension (RC).

2.1 Named Entity Recognition (NER)

NER is a classification task that identifies phrases in a text that refer to entities or predefined categories (such as dates, person, organization and location names) and it is an important capability for information access systems that perform entity lookups for knowledge verification, spell-checking or localization applications. The XTREME-UP dataset (Ruder et al., 2023) contains processed data from MasakhaNER (Adelani et al., 2021b)) and MasakhaNER 2.0 (Adelani et al., 2022) in the following languages: 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).

2.2 Reading Comprehension (RC)

RC is an important capability that enables responding to natural language questions with answers found in text. 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). Examples can be found in Table 2.

[1] https://www.codabench.org/competitions/1672/?secret_key=c68a56e8-542b-4c85-b4f5-7ce6b65643c7

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 ṣẹgun 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).

The information-seeking question-answer pairs tend to exhibit less lexical and morphosyntactic overlap between the question and answer since they are written separately, which is a more suitable setting to evaluate typologically-diverse languages. Here, the system is given a question, title, and a passage and must provide the answer — if any — or otherwise return that the question has "no answer" in the passage. The XTREME-UP benchmark currently contains data only in Indonesian, Bengali, Swahili and Telugu (Ruder et al., 2023). The competing systems will therefore be required to infer information from different language annotations.

3 Languages

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

3.1 Azerbaijani (AZ)

Azerbaijani is a member of the Turkic language family, and spoken largely in Azerbaijan and Iran. Azerbaijani shares a high degree of linguistic characteristics with other Turkic languages, especially languages in the Western Oghuz subgroup such as Turkish, Gagauz and Turkmen. Azerbaijani has an agglutinative morphology, the language also uses a Subject-Object-Verb (SOV) word order, and does not have a gender in grammar. Azerbaijanis in Azerbaijan are using Latin script since its readoption in 1991. Arabic script is also used by Iranian Azerbaijanis. The data preparation for this study is done using text in Latin script.

3.2 Igbo (IG)

Igbo belongs to the Benue Congo group of the NigerCongo language family and is spoken by over 27 million people (Eberhard et al., 2021). It is native to the southeastern Nigeria, but also spoken in some parts of Equatorial Guinea and Cameroon. There are several Igbo dialects but the most used one is the central Igbo that was standardized in 1962 (Ohiri-Aniche, 2007). The standard Igbo consists 28 consonants and 8 vowels. There are two tones: high and low. High tone is marked with an acute accent, e.g., á, while low tone is marked with a grave accent, e.g, à. These are not normally represented in the orthography. Igbo along with other African languages have been include in several benchmarks by Masakhane such as MasakhaNER (Adelani et al., 2021b, 2022), AfriQA (Ogundepo et al., 2023), MasakhaPOS (Dione et al., 2023), AfriSenti (Muhammad et al., 2023) and so on.

3.3 Indonesian (ID)

Indonesian is a member of the Austronesian language family and official language in Indonesia. The language itself is well-standarized in terms of orthography and grammar through the country, however, it has high variety on usages, especially for registers and styles influenced by the cultural influences which creates dialect variances (Aji et al., 2022). In the colloquial setting, the language usage is more challenging due to new creative abbreviations and jargons created by the speakers, which is only popular for a particular generation. The research progress on Indonesian has been tremendously improved due to the recent advancement on benchmarks (IndoNLU (Wilie et al., 2020), IndoNLG (Cahyawijaya et al., 2021), NusaCrowd (Cahyawijaya et al., 2023a), IndoLEM (Koto et al., 2020)) and datasets (NusaX (Winata et al., 2023), NusaWrites (Cahyawijaya et al., 2023b)).

3.4 Swiss German (ALS)

Swiss German is a member of the Germanic language family and the subgroup of Alemannic dialects. In contrast to Standard German, Swiss German provides a unique challenge for multilingual NLP methods, as it is a non-standardized dialect continuum with a great variety in terms of lexicon, phonetics, morphology and syntax. Especially challenging is that there exists no official orthography, and therefore each dialect variant and also each person tends to write words differently following their own interpretation of the phonetic spelling. As it is not one of Switzerland's official languages, it is mainly used in the spoken form and in informal contexts. Formal writing is done in Standard German.

Context	Question	Answer
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?	İndi "Zaqatala" qəzetində 5 nəfər işləyir.
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?	Will de Eurytos am Herakles nöd sis Töchterli - d Iole - het welle geh.
Jembatan Siak atau Jembatan Tengku Agung Sultanah Latifah adalah jembatan sepanjang 1.196 m yang terletak di kota Siak Sri Indrapura. Jembatan ini membentang di atas Sungai Siak dan diresmikan pada tanggal 11 Agustus 2007. Pembangunan jembatan ini dimulai sejak 27 Desember 2002 dan nama jembatan ini diambil dari nama gelar Tengku Syarifah Mariam binti Fadyl, permaisuri dari Sultan Syarif Kasim II, sultan terakhir di Kerajaan Siak.	Berapa panjang jembatan siak?	Jembatan siak membentang sepanjang 1.196 m yang terletak di kota siak sri indrapura
Bugünkü arokarya ağacının akrabası olan bulunmuş fosiller 50 milyon yaşındadır. Dolayısıyla dünyanın en eski ağaç familyalarından birinin üyesidir.	Arokarya ağacının dünyanın en eski ağaç familyasına ait olduğu neden düşünülmektedir?	Bulunan akraba fosillerinin 50 milyon yaşında olması sebebiyle Arokarya ağacının dünyanın eski ağaç familyasına ait olduğu düşünülmektedir.
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?	Aisha Adamu jẹ ọmọ fun Senator Adamu Baba Augie
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?	Aisha Adamu jẹ ọmọ fun Senator Adamu Baba Augie

Table 2: Examples from the RC validation data in different languages.

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

Table 3: List of languages and language families.

Consequently, very few textual resources are available. Most notably, Hollenstein and Aepli compiled a text corpus for PoS tagging using the following sources: Alemannic Wikipedia, the Swatch Group's annual report, novels of Viktor Schobinger, newspaper articles and blog posts (Hollenstein and Aepli, 2014). Further resources are available in the format of speech corpora, such as 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). Some of these also provide Swiss German transcriptions.

3.5 Turkish (TR)

As the highest-resourced language from the Turkic language family, Turkish is distinguished with its agglutinative morphology and employs an Subject-Object-Verb (SOV) word order. While lacking grammatical gender, it also features a rich case system. Verbs are inflected to indicate tense, mood, and person, while personal pronouns are used for person reference. The language incorporates vowel harmony and sound rules, with a significant number of palatalized consonants. Turkish has no definite or indefinite articles, relying on context for specificity. Additionally, it has phonemic vowel length, which affects word meaning. These properties collectively make Turkish a unique and complex language, distinct from many Indo-European languages, however its adoption of the Latin script allows meaningful comparison to representatives from the Indo-European family.

Corpus studies in Turkish include plenty monolingual (Aksan et al., 2012) and parallel resources (Tyers and Alperen, 2010; Cettolo et al., 2012; Ataman, 2018). Previous efforts also allowed the development of different 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). Turkish is now part of many public multilingual benchmarks including the mc4 corpus (Raffel et al., 2019), and it is recognized in different multilingual NLP benchmarks to create human-annotated resources, 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).

3.6 Uzbek (UZ)

The Uzbek language is spoken by over 44 million speakers globally, securing its position as the second most spoken language in the Turkic Languages group, following Turkish. It accommodates both Cyrillic and Latin scripts in its writing systems. Agglutination is a significant characteristic of Uzbek, where suffixes are appended to morphemes. It shares a high degree of agglutination with the Azeri language among Turkic languages.

Uzbek is enriched with a diversity of dialects influenced by East-Iranian (Tajik) and Turkish languages. However, the presence of multiple dialects across various regions in Uzbekistan, each with unique orthographic rules, make it challenging to standardize grammatical conventions across the language. Additionally, the Uzbek lexicon has been heavily influenced by the Russian language, resulting in a blend and substitution of words. This linguistic amalgamation poses substantial challenges in the realm of computational linguistics due to its complexity and variability.

There are few notable resources available in Uzbek. Such as (Gribanova, 2012-2020), who developed a dataset on morphological word formation involving copular and non-copular verbs including some regional and other dialectal variation of Uzbek. Further, (Gribanova, 2018-2020) compiled a dataset including native Uzbek speakers' assessment about sentences involving verb-stranding and argument ellipsis. Other resources include, Uzbek WordNET (Agostini et al., 2021), a collection of similar word pairs, (Salaev et al., 2022) and rule based Uzbek POS tagger (Sharipov et al., 2023).

3.7 Yorùbá (YO)

Yorùbá belongs to the Volta-Niger subgroup of the Niger-Congo language, native to the South-Western part of Nigeria, Benin and Togo. It is spoken by over 45 million speakers according to Ethnologue, making it one of the top-5 most spoken African language 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

We obtain the textual data for the generative task from the XML dumps provided on Wikimedia downloads[2] and sample 200 articles, which are split paragraph-wise for annotation. 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 [3].

[2] https://dumps.wikimedia.org/

[3] https://github.com/Fenerator/wikiDataProcessingForQAandNER

5 Experimental Methodology

5.1 Baseline Systems

MT0 is the open-source multi-lingual multi-task model developed by Big Science (Muennighoff et al., 2022). We use the mT0-large version of the model with 24 Transformer layers, which is based on the mT5 model that supports 101 languages. The model is finetuned on 46 additional languages with English and translated prompts.

GPT-4 OpenAI (2023b) is a Transformer-style large language model pre-trained to predict the next token similar to GPT-3 (Brown et al., 2020) followed by additional training to follow an instruction in a prompt and provide a response. The instruction training is based on Reinforcement Learning from Human Feedback (RLHF), similar to InstructGPT (Ouyang et al., 2022).

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., 2019a) [4] embeddings. Implementation is based on HuggingFace's evaluate library[5]. Overall performance in the NER task is computed in terms of precision, recall and F-1 scores using the CoNLL Evaluation Scripts[6], 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. We also perform human evaluation of the RC outputs (context-question-answer pairs) of all baselines, and the best performing submission. Two annotators judge whether the generated answer is correct, in a binary sense, and optionally add observations on the characteristics of the generated grammar, adequacy between the answer and the context, as well as any typical behavior from models related to strengths, fall backs and stylistic properties.

5.3 Submissions

The shared task received a valid submission from Charles University (CUNI) which was also the win-

[4] https://huggingface.co/roberta-base

[5] https://github.com/huggingface/evaluate

[6] https://github.com/sighsmile/conlleval

Lang	Task	# Paragraphs Val	Test	# Sentences Val	Test	# Tokens Val	Test
AZ	NER	-	-	126	124	7,774	8,200
IG	NER	-	-	711	143	54,526	11,668
ID	NER	-	-	0	0	0	0
ALS	NER	-	-	130	166	8,761	11,610
TR	NER	-	-	113	151	7,375	11,736
YO	NER	-	-	100	303	4,166	11,490
AZ	RC	38	64	116	220	2,138	3,618
IG	RC	100	175	240	469	6,263	12,175
ID	RC	100	175	230	488	4,789	10,293
ALS	RC	100	175	434	728	7,516	13,430
TR	RC	100	175	551	697	8,876	12,707
YO	RC	100	175	370	680	8,258	15,259

Table 4: Dataset statistics for the validation and test splits.

	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 answers from the systems.

ning system. In this section we describe notable details from the system developed by CUNI which aims to perform multi-lingual multi-task information retrieval by providing a pivoting approach where any input is translated into English to perform the end task, and translated back to the original language for final comparison.

CUNI Question Answering (CQA) system uses the RoBERTa model (Liu et al., 2019b) fine-tuned on the question answering task using XTREME-UP (Ruder et al., 2023) and span matching based on the label projection approach by Chen et al. (2023).

CUNI Contrastive (CCo) In order to generate more naturalistic language and overcome issues related to domain mismatch, CUNI provided also contrastive generations (*i.e.*) in the RC task where they compared their output quality on the validation sets with the LLAMA-2 (Touvron et al., 2023b) model and make an additional experimental submission, which we also include in our evaluation.

CUNI NER also deploys multi-lingual fine-tuning including the MasakhaNER (Adelani et al., 2021b) data in order to increase robustness of the model to domain mismatch.

w. score	CQA	CCo	mT0	GPT-4
ChrF	0.23	0.27	0.26	0.45
ChrF+	0.22	0.25	0.24	0.44
ChrF++	0.21	0.23	0.23	0.42
RougeL	0.25	0.20	0.28	0.36
BERT F1	0.83	0.84	0.82	0.87

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

6 Results

6.1 Automatic Evaluation

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 6. Detailed results per system and language are presented in Table 7. We also present NER results for the CUNI system submission in Table 8.

6.2 Human Evaluation

Table 11 provides an overview of the relative amount of times the system generated an answer judged as correct by the human annotators.

Pearson correlation coefficients between the automatic metrics and the human annotations can be

system	language	ChrF aut.	ChrF r	ChrF+ aut.	ChrF+ r	ChrF++ aut.	ChrF++ r	RougeL aut.	RougeL r	BERTScore F1 aut.	BERTScore F1 r
CQA	AZ	0.42	-	0.40	-	0.39	-	0.44	-	0.90	-
CQA	ID	0.37	-	0.34	-	0.32	-	0.39	-	0.84	-
CQA	IG	0.14	-	0.14	-	0.13	-	0.19	-	0.79	-
CQA	TR	0.15	-	0.15	-	0.14	-	0.19	-	0.82	-
CQA	UZ	0.44	-	0.43	-	0.42	-	0.47	-	0.89	-
CQA	YO	0.23	-	0.22	-	0.21	-	0.24	-	0.82	-
CQA	ALS	0.12	-	0.11	-	0.11	-	0.09	-	0.79	-
CCo	AZ	0.34	0.36	0.33	0.37	0.31	0.35	0.28	0.34	0.87	0.25
CCo	ID	0.39	-0.04	0.36	-0.02	0.33	-0.02	0.30	0.07	0.86	0.01
CCo	IG	0.24	0.38	0.24	0.39	0.22	0.37	0.24	0.30	0.85	0.23
CCo	TR	0.24	0.04	0.24	0.05	0.22	0.06	0.21	0.07	0.85	0.08
CCo	UZ	0.36	0.44	0.34	0.42	0.31	0.43	0.22	0.38	0.85	0.32
CCo	YO	0.19	0.39	0.18	0.41	0.17	0.41	0.17	0.28	0.81	-0.04
CCo	ALS	0.19	0.27	0.19	0.28	0.17	0.27	0.07	0.33	0.82	0.39
mT0 (1B)	AZ	0.33	0.67	0.32	0.67	0.31	0.68	0.37	0.59	0.86	0.35
mT0 (1B)	ID	0.48	0.38	0.44	0.37	0.42	0.36	0.48	0.16	0.88	0.25
mT0 (1B)	IG	0.14	0.34	0.14	0.37	0.14	0.38	0.20	0.51	0.79	0.22
mT0 (1B)	TR	0.12	0.09	0.12	0.10	0.11	0.12	0.15	0.26	0.80	0.02
mT0 (1B)	UZ	0.49	0.47	0.47	0.47	0.46	0.47	0.55	0.52	0.90	0.31
mT0 (1B)	YO	0.28	0.47	0.27	0.47	0.26	0.47	0.30	0.47	0.82	0.21
mT0 (1B)	ALS	0.12	0.46	0.11	0.47	0.11	0.46	0.09	0.47	0.78	0.39
GPT-4	AZ	0.41	0.42	0.41	0.44	0.39	0.44	0.31	0.32	0.86	0.27
GPT-4	ID	0.51	0.08	0.49	0.09	0.47	0.10	0.47	0.11	0.88	0.08
GPT-4	IG	0.52	0.28	0.52	0.28	0.49	0.28	0.45	0.21	0.89	0.17
GPT-4	TR	0.57	0.02	0.57	0.03	0.53	0.03	0.49	0.05	0.92	0.11
GPT-4	UZ	0.53	0.02	0.52	0.02	0.51	0.02	0.43	0.01	0.87	0.09
GPT-4	YO	0.28	0.52	0.27	0.52	0.26	0.53	0.21	0.59	0.82	0.48
GPT-4	ALS	0.34	0.26	0.34	0.27	0.30	0.26	0.19	0.26	0.85	0.30

Table 7: Detailed RC results per system and language. "aut." denotes automatic evaluation results on the entire test set, r denotes the Pearson correlation coefficient between the respective metric and the binary human judgement on the annotated subset of the test data.

Lang.	All Tags acc	All Tags pre	All Tags rec	All Tags F1	LOC pre	LOC rec	LOC F1	ORG pre	ORG rec	ORG F1	PER pre	PER rec	PER F1
ALS	0.87	0.37	0.41	0.39	0.50	0.41	0.45	0.30	0.27	0.28	0.57	0.43	0.49
AZ	0.87	0.49	0.47	0.48	0.68	0.40	0.50	0.49	0.40	0.44	0.72	0.55	0.62
IG	0.89	0.46	0.58	0.51	0.67	0.51	0.58	0.33	0.34	0.33	0.78	0.68	0.72
TR	0.89	0.52	0.48	0.50	0.66	0.43	0.52	0.53	0.31	0.39	0.80	0.53	0.64
YO	0.84	0.52	0.63	0.57	0.73	0.44	0.55	0.49	0.51	0.50	0.85	0.81	0.83
w. average	0.87	0.47	0.52	0.49	0.64	0.44	0.52	0.42	0.36	0.39	0.75	0.60	0.66

Table 8: Test results for CUNI NER submission. Averages are weighted by number of tokens per language.

	$r(ChrF, h)$	$r(ChrF+, h)$	$r(ChrF++, h)$	$r(RougeL, h)$	$r(BERTF1, h)$
CCo	0.26	0.27	0.27	0.25	0.18
mT0 (1B)	0.41	0.42	0.42	0.43	0.25
GPT-4	0.23	0.23	0.24	0.22	0.21

Table 9: Pearson correlation r between metrics and human binary annotation (h) averaged over languages.

	$r(ChrF, h)$	$r(ChrF+, h)$	$r(ChrF++, h)$	$r(RougeL, h)$	$r(BERTF1, h)$
AZ	0.48	0.49	0.49	0.42	0.29
ID	0.14	0.15	0.15	0.11	0.11
IG	0.33	0.35	0.34	0.34	0.20
TR	0.05	0.06	0.07	0.13	0.07
UZ	0.31	0.30	0.31	0.30	0.24
YO	0.46	0.47	0.47	0.45	0.22
ALS	0.33	0.34	0.33	0.35	0.36

Table 10: Pearson correlation r between metrics and human binary annotation (h) averaged over systems.

Lang.	**mT0 (1B)**	**GPT-4**	**CCo**
AZ	0.42	0.78	0.68
ID	0.85	0.98	0.54
IG	0.44	0.92	0.42
TR	0.44	0.90	0.60
UZ	0.80	0.92	0.78
YO	0.52	0.64	0.36
ALS	0.48	0.92	0.48

Table 11: Relative amount of answers that were judged as correct by human annotators.

found in detail in Table 8. Table 10 provides an overview of the correlations by language, and Table 9 condenses the correlations per system.

According to our analysis, we find the GPT-4 as a strong baseline in the RC task and it has competitive rephrasing and reasoning capabilities. We notice when GPT-4 generates an answer it often rephrases the question into a statement which might cause some grammatical errors if the case do not directly translate and may need additional inflectional changes. In general, we find although grammatical errors exist, they do not always lead to complete semantic loss in the sentence and might allow check the information.

An important remark is the factuality of the GPT-4 answers which we also approach skeptically. We find a small percentage of the time GPT-4 generates information that do not exist in the provided context.

Especially in dialects and low-resourced languages, we observe incorrect language in the output. The majority of these incorrect outputs are in Swiss German (ALS) and Azerbaijani (AZ). We also find this problem reciprocates in understanding the prompt, whereas observing in Swiss German similar words such as "zwei" (translation: two) and "zwor" (translation: hence) are misinterpreted. The ability to understand and generate output in the desired language might be limited by data availability and current observations state it is not trivial for GPT-4 to directly allow usage in low-resourced languages.

The second baseline, MT-0, was found to be relatively different in the style and characteristics of the language generated. Most answers were precise and rather short although, in light of our human evaluation results, majorly correct in some languages like Indonesian (ID) and Uzbek (UZ). We find MT-0 to be more prone to spelling errors which might lead to more semantic losses. For Igbo (IG), Turkish (TR) and Swiss German (ALS) we find the majority of answers are incorrect. We also observe multiple typographical errors, such as the way to write metrics (e.g., “k" instead of “km") in ID, although the values are correct.

The answers provided by CUNI were generally fluent and presented plausible language. The system tended more frequently to make up non-factual information or information that cannot be inferred from the given context. We also observed incorrect language in the output, which was at a significant level in Swiss German (ALS) and Uzbek (UZ).

7 Conclusion and Future Work

We presented a new multi-lingual multi-task benchmark on information retrieval from Wikipedia in seven languages from typologically-diverse and low-resourced language families. 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. Our leaderboard that will continue to promote open access evaluation of new submissions of specialized systems will be available until MRL 2024 on the competition website.

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

This research involved using human annotators to prepare data sets. All annotators were provided with clear instructions and guidelines to ensure the responsible and unbiased annotation of the data. We ensured ethical 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 our sponsors Google Deepmind and Bloomberg to make this shared task possible. 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.

References

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–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.

Alessandro Agostini, Timur Usmanov, Ulugbek Khamdamov, Nilufar Abdurakhmonova, and Mukhammadsaid Mamasaidov. 2021. UZWORDNET: A lexical-semantic database for the Uzbek language. In *Proceedings of the 11th Global Wordnet Conference*, pages 8–19, University of South Africa (UNISA). Global Wordnet Association.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249.

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.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, et al. 2023a. Nusacrowd: Open source initiative for indonesian nlp resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Maulana Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, et al. 2023b. Nusawrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. *arXiv preprint arXiv:2309.10661*.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898.

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*.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

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.

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.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv preprint arXiv:2202.11822*.

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

Vera Gribanova. 2012-2020. The combinatorics of the Uzbek verbal complex in polar questions: Stanford digital repository.

Vera Gribanova. 2018-2020. Argument ellipsis and verb-stranding ellipsis in Uzbek: Stanford digital repository.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv e-prints*, pages arXiv–2302.

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.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770.

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.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

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

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

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and gpt-4. *arXiv preprint arXiv:2308.12488*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv e-prints*, pages arXiv–2211.

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.

Chinyere Ohiri-Aniche. 2007. Stemming the tide of centrifugal forces in Igbo orthography. *Dialectical Anthropology*, 31(4):423–436.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. 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*.

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.

Ulugbek Salaev, Elmurod Kuriyozov, and Carlos Gómez-Rodríguez. 2022. Simreluz: Similarity and relatedness scores as a semantic evaluation dataset for uzbek language.

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.

Maksud Sharipov, Elmurod Kuriyozov, Ollabergan Yuldashev, and Ogabek Sobirov. 2023. Uzbektagger: The rule-based pos tagger for uzbek language.

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.

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. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, and Pascale Fung. 2023. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

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*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2021. Evaluating commonsense in pretrained language models.