

Overview of the CIRAL Track at FIRE 2023: Cross-lingual Information Retrieval for African Languages

Mofetoluwa Adeyemi¹, Akintunde Oladipo¹, Xinyu Crystina Zhang¹,
David Alfonso-Hermelo², Mehdi Rezagholizadeh², Boxing Chen² and Jimmy Lin¹

¹University of Waterloo

²Huawei Noah's Ark Lab

Abstract

This paper provides an overview of the first CIRAL track at the Forum for Information Retrieval Evaluation 2023. The goal of CIRAL is to promote the research and evaluation of cross-lingual information retrieval for African languages. With the intent of curating a human-annotated test collection through community evaluations, our track entails retrieval between English and four African languages which are Hausa, Somali, Swahili and Yoruba. We discuss the cross-lingual information retrieval task, curation of the test collection, participation and evaluation results. Analysis of the curated pools is provided, and we also compare the effectiveness of the submitted retrieval methods. The CIRAL track did show and encourage the research prospects that exist for CLIR in African languages, and we are hopeful for the direction this takes.

Keywords

Cross-lingual Information Retrieval, African Languages, Ad-hoc Retrieval, Passage Ranking, Community Evaluations

1. Introduction

Cross-lingual information retrieval (CLIR) is a specific category under multilingual retrieval, which retrieves documents in a language different from the given query. It plays a huge role in obtaining information mostly available in the document's language. Efforts in CLIR go as far back as the early 90s starting with Text Retrieval Conference (TREC) [1] and has been migrated to other conferences including CLEF [2], FIRE [3], and NCTIR [4], which have specific focus on European languages, Indian languages, and East Asian languages. Tracks dedicated to cross-lingual information retrieval in these conferences, such as the NeuCLIR track [5] in TREC, are a venue to promote the participation and evaluation of these groups of languages in CLIR. However, there is a lag in such research involvement for African languages.


There are also few resources for studying CLIR in African languages, despite the growing research efforts on them in cross-lingual information retrieval [6, 7, 8, 9]. AfriCLIRMatrix [10]

Forum for Information Retrieval Evaluation, December 15-18, 2023, India

✉ moadeyem@uwaterloo.ca (M. Adeyemi); aooladipo@uwaterloo.ca (A. Oladipo); x978zhan@uwaterloo.ca (X. C. Zhang); david.alfonso.hermelo@huawei.com (D. Alfonso-Hermelo); mehdi.rezagholizadeh@huawei.com (M. Rezagholizadeh); boxing.chen@huawei.com (B. Chen); jimmylin@uwaterloo.ca (J. Lin)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

constructs the first CLIR dataset in African languages, built synthetically based on Wikipedia’s structure and covers 15 African languages. Other collections such as large scale CLIR [11], CLIRMatrix [12] and the IARPA MATERIAL test collection [13] which was curated solely for low-resource languages, only cover a minimal amount of African languages.

The sparsity of resources and the bid to promote participation in CLIR research for African languages calls out the construction of CIRAL, which stands for **Cross-lingual Information Retrieval for African Languages**. The CIRAL track hosted at the Forum for Information Retrieval Evaluation (FIRE) focused on cross-lingual passage retrieval covering 4 African languages: Hausa, Somali, Swahili, and Yoruba, which are some of the most widely spoken languages in Africa. Given the low-resource nature of African languages, even in widely used sources like Wikipedia, CIRAL’s collection is built with articles from the indigenous news domain of the respective languages. Similar to the passage ranking task in TREC’s Deep learning track [14], relatively few queries (80 to 100) are developed for the task. The queries and judgments are produced by native speakers who took the roles of query developers and relevance assessors. As is often the culture in community evaluations, CIRAL also set out to curate a test collection by pooling submissions from track participants.

In hosting CIRAL, we look out for: 1) The effectiveness of indigenous textual data in CLIR for African languages, 2) A comparison of how well different retrieval methods perform in CLIR for African languages, 3) The importance of retrieval and participation diversity. An overview of the curated collection, query development and relevance assessment process is provided and results from relevance assessment demonstrate the effectiveness of retrieving relevant passages from indigenous sources. Participation in the track and submissions for the respective languages are also discussed, comparing different retrieval methods employed in the task. Taking into consideration participation, submissions and other factors, we examine the test collection curated from the task, which informs future decisions in community evaluations for African languages.

We hope CIRAL fosters CLIR evaluation and research in African languages and in low-resource settings, and hence the development of retrieval systems that are well suited for such tasks. Details of the track are also available on the provided website.¹

2. Task Description

The focus task at CIRAL was cross-lingual passage ranking between English and African languages. For a kick-off, only four African languages were included this year: Hausa, Somali, Swahili, and Yoruba, which are selected according to the size of native speakers of the languages in East and West Africa. All four languages are in Latin script, two belonging to the Afro-Asiatic language family, and the other two to the Niger-Congo family. See details of the languages in Table 1. We choose English as the pivot language as it is an official language in countries where these African languages are spoken, with the exception of Somali whose speakers lean more towards Arabic than English.

Given English queries, participants are tasked with developing retrieval systems which return ranked passages in the African languages according to the estimated likelihood of

¹<https://ciralproject.github.io/>

Language	Language Family	Region	# Speakers	Script
Hausa	Afro-Asiatic	West Africa	77M	Latin
Somali		East Africa	22M	Latin
Swahili	Niger-Congo	East Africa	83M	Latin
Yoruba		West Africa	55M	Latin

Table 1
Details on the African languages in the CIRAL task.

Language	# of Passages	Median Tokens per Passage	Avg Tokens per Passage	# News Articles	News Sources
Hausa	715,355	144	135.29	240,883	LegitNG, DailyTrust, VOA, Isyaku, etc.
Somali	827,552	131	126.13	629,441	VOA, Tuko, Risaala, Caasimada, etc.
Swahili	949,013	129	126.71	146,669	VOA, UN Swahili, MTanzania, etc.
Yoruba	82,095	168	167.94	27,985	Alaroye, VON, BBC, Asejere, etc.

Table 2
Collection details for each language in CIRAL. The average and median number of tokens per passage also gives an idea of the distribution of passages in each language. The table also shows the number and some sources of news articles collected for each language.

relevance. Queries are formulated as natural language questions and passages are judged using binary relevance: 0 for non-relevant and 1 for relevant. The relevant passage is defined as the one that answers the question, whereas the non-relevant passage does not. To facilitate the development and evaluation of their retrieval systems, participants were provided with a training set comprising a sample of 10 queries for each language, their relevance judgments and the passage collection for the languages. Considering the nature of the task, we evaluate for early precision and recall using metrics such as nDCG@20 and Recall@100 and participants were also made aware of these in developing their systems. For evaluations, the test set of queries was provided for which submitted runs were manually judged to form query pools. Participants were also encouraged to rank their submitted runs in the order that they preferred to contribute to the pools. Details on the provided passage collection, development of queries, and pooling process are discussed in the following sections.

3. Passage Collection

CIRAL’s passage collection is curated from indigenous news websites and blogs for each of the four languages. These sites serve as a source of local and international information and as shown in Table 2, are a huge source of text for their languages. The articles are collected using a web scrapping framework called *Otelemuye*² and combined into monolingual document sets. The collected articles date from as early as was available on the website (which was the early 2000s for some languages) up until March 2023. Passages are generated from the set by chunking each news article on a sentence level using a sliding-window segmentation [15]. To ensure natural discourse segments when chunking the articles, a stride window of 3 is used with

²<https://github.com/theyorubayesian/otelemuye>

a maximum of 6 sentences per window. The resulting passages are further filtered to remove those with less than 7 or more than 200 words. Table 2 shows the median and average number of tokens per passage in each language, providing more insight into their passage distribution. To ensure each passage is in its required language, we filter using the language’s list of stopwords, hence removing passages in a different language; a minimum of 3 to 5 stopwords was used to ascertain if a passage was in its African language. The resulting number of passages is shown in Table 2.

The curated passages are provided in JSONL files, each line representing a JSON object with details about a passage. Passages have the following fields:

- `docid`: Unique identifier
- `title`: The headline of the news article from which it was obtained.
- `text`: The passage body.
- `url`: The link to the news article from which it was gotten.

The unique identifier (`docid`) for each passage is constructed programmatically to have the format `source#article_id#passage_id` providing information on the news website and specific article number the passage was extracted from in the monolingual set. This is also helpful as there are a few news articles without titles, hence leaving the respective passages without a text in the `title` field. The passage collection files were made publicly available to participants in a Hugging face dataset repository.³

4. Query Development

Queries for the task are formulated as natural language questions, modelling that of collections such as MS MARCO [16] which is used in TREC’s Deep learning track [14], the MIRACL dataset [17], among others. Considering the passage collection for a language was curated from its indigenous websites, language queries had to have topics either of interest to the speakers of the language or with information that can be easily found in the language’s news. We term these queries *language/cultural-specific queries*,⁴ which is a combination of queries with both generic and indigenous topics depending on the language. The *language-specific* queries are developed as factoids to ensure answers are direct and unambiguous.

The process of query development involved native speakers generating questions with answers in the language’s news. For this task, articles from the MasakhaNews dataset [18] are used as a source of inspiration for the query formation. The MasakhaNews dataset is a news topic classification dataset and covers 16 African languages. It serves as a good starting point given that the documents in the dataset have been classified into categories namely: *business*, *entertainment*, *health*, *politics*, *religion*, *sport* and *technology*, providing a more direct approach for generating diverse queries. Using the same passage preprocessing implemented in generating the passage collection, articles in MasakhaNews are chunked into various passages but with an additional `category` field to jointly inspire queries. Query developers (interchangeably

³<https://huggingface.co/datasets/CIRAL/ciral-corpus>

⁴The generated queries also include some which are generic, but we term the queries language-specific due to the news data also capturing events which are mostly of interest in the language.

The image shows a web interface for searching passages. It consists of several input fields and a submit button. At the top left is an 'Annotator' dropdown menu. To its right is a 'Language' dropdown menu with 'Somali' selected. Below these is a 'Query' text input field with the placeholder 'Type your query here...'. Underneath the query field are two more text input fields: 'Translation' with the placeholder 'Translate your query to English...' and 'Inspiration' with the placeholder 'Enter the docid of the passage that inspired your query'. At the bottom center is a grey 'Submit' button.

Figure 1: The interface of the Hugging Face space used to search for relevant passages in the query development process. Annotators are able to select their names, input the in-language query, its English translation and docid of the passage that inspired it. This is the interface for the Somali space.

called annotators) are native speakers of the languages with reading and writing proficiency in both English and their respective African languages. To generate these queries, annotators are given the MasakhaNews passage snippets and tasked with generating questions inspired but not answerable by the snippet to ensure good-quality queries. The questions are generated in the African language and then translated into English by the annotator.

To ensure that generated queries had relevant passages, the annotators checked if an inspired question had passages answering the question in the CIRAL collection. This was done via a search interface developed as a Hugging Face space for each language using Spacerini.⁵ As shown in the Figure 1, the annotators provide the query in its African language, its English translation and the `docid` of the passage that inspired it and search was monolingual i.e. using the query in its African language. Using a hybrid of BM25 [19] and AfriBERTa DPR indexes,⁶ the top 20 retrieved documents were annotated for relevance with selections for either `true` or `false`. Relevance annotation was done as follows:

- Relevant (True): The annotator selected `true` if the passage answered the question or implied the answer without doubt.
- Non-relevant (False): The annotator selected `false` if the passage didn't answer the question.

Instances where the passages gave partial or incomplete answers to the question also occurred and depending on the level of incompleteness, the annotators judged the passages as non-relevant. Passages annotated as `true` in the interface were assigned a relevance of 1 and those annotated as `false` a relevance of 0. Queries retained and distributed in the task had at least 1 relevant passage and no more than 15 relevant passages to avoid way too simple queries for the systems. Ambiguous or incomprehensible queries were also filtered out from the collection. A set of 10 queries for each language was first developed and released along with the corresponding judgments as train samples. Subsequently, the test queries for which

⁵<https://github.com/castorini/hf-spacerini>

⁶<https://huggingface.co/castorini/afriberta-dpr-ptf-msmarco-ft-latin-mrtydi>

Language	# Dev Queries	# Dev Judg.	# Test Queries	# Test Judg.
Hausa	10	165	85	1,523
Somali	10	187	100	1,728
Swahili	10	196	85	1,656
Yoruba	10	185	100	1,921

Table 3

Statistics of CIRAL’s queries and judgements. 10 queries were released for each language as train samples along with their judgements.

Date	Event
13th July 2023	Hausa and Yoruba Training Data Released
6th Aug 2023	Somali and Swahili Training Data Released
21st Aug 2023	Test Data Released
10th Sep 2023	Run Submission Deadline
26th Sep 2023	Distribution of Results

Table 4

Track timeline showing the release dates of datasets, submission of runs and result distribution.

the pooling process was to be carried out were released: 85 for Hausa, 100 for Somali, 85 for Swahili and 100 for Yoruba as presented in Table 3.

Judgments obtained during the query development process are referred to as shallow considering they are few. The number of shallow judgments obtained through the query development process for the test queries is also shown in Table 3, and these judgments are reserved in the pool formation during relevance assessment. The different timelines for which each set was released, along with the run submission and result distribution dates are provided in Table 4.

5. Relevance Assessment

As often practised in community evaluation, runs submitted for the test set are manually judged to form the test collection’s qrels via pooling. A total of 84 submissions were made by the participating teams, 21 for each language. Using the ranked list of runs provided by the teams, query pools were formed for each language and we provide details on the relevance assessment process and analysis of pools in this section.

5.1. Pooling Process

The top 3 ranked submissions from participating teams contributed to the pooling process, with subsequent additions depending on available time and assessment resources. A total of 40 runs contributed to the pools across the four languages, depending on the model type; however dense models made up more of the contributing runs. The pool depth for submissions was kept

	Hausa	Somali	Swahili	Yoruba
Minimum across queries	47	53	58	43
Maximum across queries	117	117	126	125
Total pool size	7,288	9,094	8,079	8,311

Table 5

The minimum and maximum pool size per query across the languages. Certain queries do not get any more contributing passages to their pools and plateau at pool sizes 40 to 60, while some queries have pool sizes more than 100.

	Hausa	Somali	Swahili	Yoruba
Minimum	1	1	1	1
Maximum	81	65	71	61
Mean	24	21	28	14
Median	20	19	29	10
Total	1,918	2,030	2,386	1,397

Table 6

Statistics of relevant passages in curated pools.

at a constant of $k = 20$, however, there was no restricted size for the pools. Judgments were carried out by two assessors for each language, where an assessor judged the full pool of a given query; the test set was split into distinct halves and each assigned to an assessor. Assessors provided judgments on a binarized scale using the following description:

- **Relevant:** The passage answers the query, or the answer can be very easily implied from the passage.
- **Non-relevant:** The passage doesn't answer the question at all, or is related to the question but doesn't answer the question.

Relevant passages are given a judgment of 1 and non-relevant a judgment of 0.

5.2. Pool Analysis

The total pool size obtained for each language from the relevance assessment is presented in Table 5. This includes the sparse judgments obtained during query development, which were also re-assessed during the relevance assessment phase. Across the languages, the minimum number of judgments per query ranges from 40 to 60 while some queries have up to over 120 judgments. 3 queries in Hausa, 4 in Somali, 2 in Swahili and 12 in Yoruba have pool sizes of less than 60 passages, indicating that contributing runs retrieved similar sets of passages for these queries in their top 20 ranks. Runs which contributed to the pooling process also retrieved more relevant passages across the four languages as seen in Table 6. However, certain queries were found to have no relevant passages and were discarded. This was as a result of wrongly annotated passages from the query development phase, or grammatical errors which affected retrieval results. This left Hausa with 80 test queries as opposed to the initial 85 queries

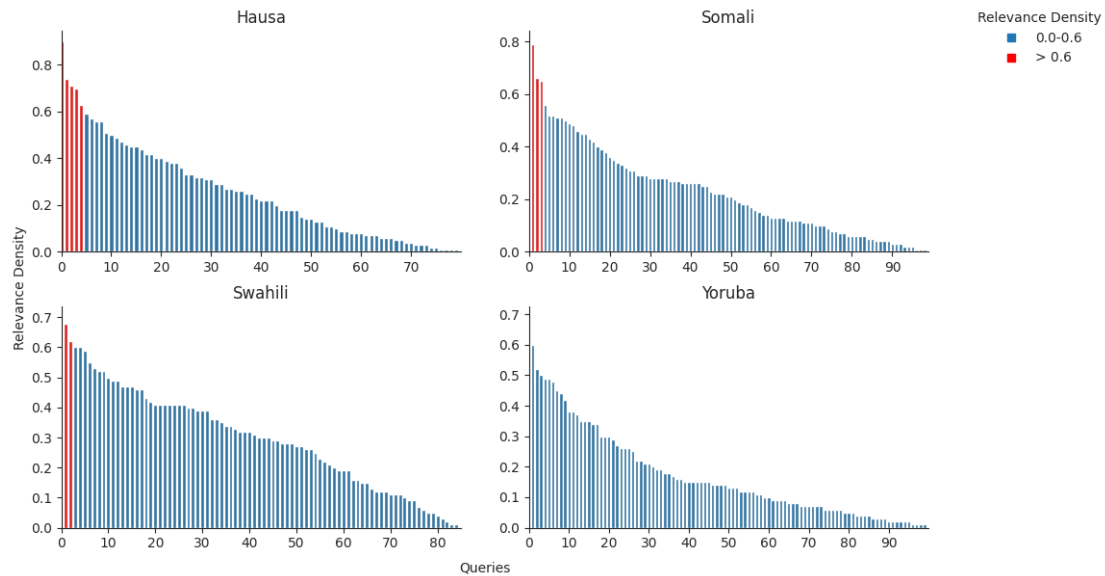


Figure 2: Distribution of relevance densities among the queries in each language. Across the languages, queries with densities of 0.6 and above are relatively few, with Yoruba having 0 queries.

and Somali with 99 as opposed to 100. There were also a few queries with just 1 relevant passage across the languages with Yoruba having the highest of 5 queries. The increased amount of relevant passages obtained from the pooling process is a good indication that African indigenous websites are a great source for retrieval, especially coupled with queries of interest to the language speakers, which could also include generic topics.

Table 6 also indicates that a large number of relevant passages were obtained for certain queries. Considering the minimal number of runs that contributed to the pools, this raises the concern that more relevant passages might remain unjudged, especially for runs that didn't contribute to the pooling process or are evaluated after the track. We analyse the number of queries with the highest tendency of having unjudged relevant passages using relevance densities. The relevance density of a query is its number of relevant passages compared to its pool size, and we adopt a rule of thumb that queries with relevance densities 0.6 and higher very likely still have unjudged relevant passages. Figure 2 gives a distribution of the relevance densities for each language and we find that the number of queries with densities higher than 0.6 is less than 5 across the languages. There is also a higher amount of queries having densities between 0.6 and 0.4, with Swahili and Hausa having up to 22 to 25% of queries. Although this approach to analysing the completeness of judgment isn't holistic, it provides some insight on the quantity of queries in each language that would most certainly have unjudged relevant passages from new systems.

	nDCG@20		MRR@10		Recall@100		MAP	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Hausa	0.2690	0.5700	0.4230	0.6952	0.3598	0.5902	0.1624	0.3611
Somali	0.2403	0.5118	0.4115	0.7102	0.3265	0.6436	0.1483	0.3567
Swahili	0.2644	0.5232	0.4537	0.7222	0.3249	0.5956	0.1406	0.3117
Yoruba	0.3115	0.5819	0.4486	0.6211	0.5091	0.8057	0.2135	0.4512

Table 7
Mean and Maximum scores across all runs.

6. Results and Analysis

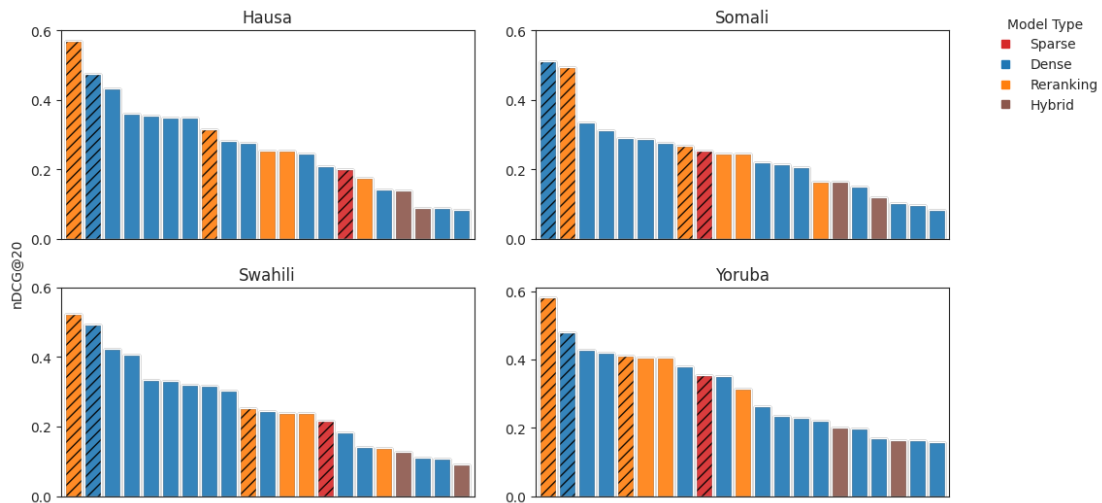
An overview of participants’ submissions and the results obtained from evaluating submitted runs on the pooled qrels is provided in this section. Results are also analysed at the query level to identify query difficulty, as well as the effectiveness of the submitted runs and model type.

A total of 3 teams participated in the CIRAL track with 84 runs submitted. Considering that cross-lingual passage ranking was the major task, participants weren’t given any specifications on the retrieval type to employ and submissions comprised dense (52), reranking (20), hybrid (8) and sparse (4) methods. All submissions covered the four languages hence there is an equal number of runs among the languages. The retrieval methods employed by participating teams are properly discussed in their working notes.

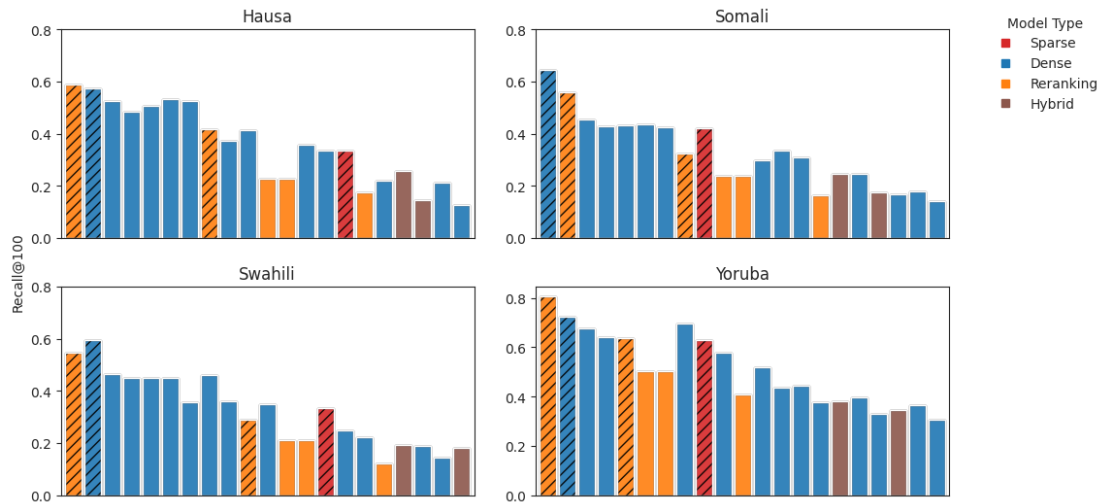
6.1. Overall Results

We present the results statistics of all languages in Table 7, and the detailed results of all submitted runs in Tables 8, 9, 10 and 11. The nDCG@20, MRR@10, Recall@100, and MAP@100 scores for each submission are reported and the average and maximum scores can be found in Table 7. The main metric in the task is nDCG@20 and a cut-off of $k=20$ is used considering a decent number of queries had above 10 relevant passages during query development. Dense models make up 62% of submissions for each language and have the highest average scores across the metrics. Most submissions employ end-to-end cross-lingual retrieval with a few document translation methods represented as DT in the table. However, the top 2 performing submissions across the languages employ document translation at one stage or the other in their systems and have the highest scores for all metrics.

The effectiveness of model types is better visualized in Figure 3. Runs are ordered by the nDCG@20 scores, and though dense runs make up most of the top runs, there is a variation in effectiveness across the dense models. The effectiveness of reranking methods also varies widely across the languages, with the exception of Yoruba where reranking models have the top nDCG@20 scores as seen in Figure 3a. Given there wasn’t a specific task on reranking, submitted runs employ different first and second-stage methods which has an impact on the varying degree of output quality. However, the best reranking run outperformed the best dense run across the languages with the exception of Somali. The submission pool has a very minimal number of hybrid and sparse runs, giving insufficient room for comparison of the model types on the task. The sparse run, however, outperforms some of the dense and reranking runs and achieves competitive nDCG scores, especially in Somali and Yoruba.



(a) nDCG@20



(b) Recall@100

Figure 3: Distribution of nDCG@20 and Recall@100 among the various run types, ordered by nDCG@20 in both images. Hatched bars represent runs that implement document translation at any stage in their methods, hence most submitted runs employ end-to-end CLIR retrieval.

Dense models achieve higher recall@100 across all languages as seen in Figure 3b. Maintaining the same order by nDCG@20, runs not having a high nDCG@20 retrieved more relevant passages in their top 100 candidates. With the exception of Yoruba and the best reranking model, reranking generally achieved lower recall@100, with even the sparse run achieving a better score across the languages. These results indicate that many of the submitted systems have relevant passages at deeper depths, however, due to the nature of the task, we optimize for early rankings using nDCG@20.

6.2. Query-level Results

Figures 4, 5, 6 and 7 provide query level effectiveness using nDCG@20 and queries are ordered by the median scores across evaluated runs. The median nDCG score for a good percentage of queries is greater than 0, indicating that most submissions do not perform too badly on individual queries across the languages. Certain queries, such as 41 in Hausa, also have quite a gap between the maximum score obtained and the scores by the rest of the runs, indicating specific runs perform better on these queries compared to other runs. The same can be said for queries like 81 in Swahili, where only a few runs identify the relevant passages of the query. This implies that these runs understand the semantics of the query and such queries could boost the scores of systems that are able to retrieve its relevant documents.

We also analyse the query difficulty across the languages, as queries that are too easy or difficult are not ideal in distinguishing systems' effectiveness. Examples of these are queries 72 in the Hausa language and 433 in Yoruba where the median nDCG@20 score is 1.0 across submitted systems, making them very easy queries and problematic for evaluation. There are also quite a number of difficult queries across the languages, with Somali having the highest, where only a few outliers score higher than 0 nDCG@20. However, a good number of queries such as 21 in Swahili and 161 in Somali have a decent spread of scores and are ideal for evaluation.

7. Conclusion

The CIRAL track was held for the first time at the Forum for Information Retrieval Evaluation (FIRE) 2023, with the goal of promoting the research and evaluation of cross-lingual information retrieval for African languages. The task covered passage retrieval between English and four African languages and test collections were curated for these languages via community evaluations. Submissions from participating teams comprise mostly dense single-stage retrieval systems, and these make up most of the best-performing systems on the task. Some limitations faced this year include a minimal number of participants and less diversity in submitted retrieval systems. Despite the limitations, we hope the CIRAL track evolves and the curated collection matures into its most reliable and reusable version.

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. We would like to thank the Masakhane community⁷ for their contributions in the query development phase of the project. We also appreciate John Hopkins University HLTCOE, organizers of the NeuCLIR track at TREC,⁸ for contributing the English translations of the passage collections to the track.

⁷<https://www.masakhane.io/>

⁸<https://neuclir.github.io/>

References

- [1] P. Schäuble, P. Sheridan, Cross-language information retrieval (CLIR) track overview, NIST SPECIAL PUBLICATION SP (1998) 31–44.
- [2] C. Peters, Information retrieval evaluation in a changing world lessons learned from 20 years of CLEF (2019).
- [3] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Pal, D. Modak, S. Sanyal, The FIRE 2008 evaluation exercise, *ACM Transactions on Asian Language Information Processing (TALIP)* 9 (2010) 1–24.
- [4] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, S. Hidaka, Overview of IR tasks at the first NTCIR workshop, in: *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition, 1999*, pp. 11–44.
- [5] D. Lawrie, S. MacAvaney, J. Mayfield, P. McNamee, D. W. Oard, L. Soldaini, E. Yang, Overview of the TREC 2022 NeuCLIR track, *arXiv preprint arXiv:2304.12367* (2023).
- [6] X. Zhang, K. Ogueji, X. Ma, J. Lin, Toward best practices for training multilingual dense retrieval models, *ACM Transactions on Information Systems* 42 (2023) 1–33.
- [7] M. Yarmohammadi, X. Ma, S. Hisamoto, M. Rahman, Y. Wang, H. Xu, D. Povey, P. Koehn, K. Duh, Robust document representations for cross-lingual information retrieval in low-resource settings, in: *Proceedings of Machine Translation Summit XVII: Research Track, 2019*, pp. 12–20.
- [8] S. Nair, P. Galuscakova, D. W. Oard, Combining contextualized and non-contextualized query translations to improve CLIR, in: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020*, pp. 1581–1584.
- [9] L. Zhao, R. Zbib, Z. Jiang, D. Karakos, Z. Huang, Weakly supervised attentional model for low resource ad-hoc cross-lingual information retrieval, in: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), 2019*, pp. 259–264.
- [10] O. Ogundepo, X. Zhang, S. Sun, K. Duh, J. Lin, AfriCLIRMatrix: Enabling cross-lingual information retrieval for African languages, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022*, pp. 8721–8728.
- [11] S. Sasaki, S. Sun, S. Schamoni, K. Duh, K. Inui, Cross-lingual learning-to-rank with shared representations, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018*, pp. 458–463.
- [12] S. Sun, K. Duh, CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020*, pp. 4160–4170.
- [13] C. Rubino, Machine translation for English retrieval of information in any language (machine translation for English-based domain-appropriate triage of information in any language), in: *Conferences of the Association for Machine Translation in the Americas: MT Users’ Track, The Association for Machine Translation in the Americas, Austin, TX, USA, 2016*, pp. 322–354.
- [14] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, Overview of the TREC 2019

- Deep Learning track, arXiv preprint arXiv:2003.07820 (2020).
- [15] M. S. Tamber, R. Pradeep, J. Lin, Pre-processing matters! Improved Wikipedia corpora for open-domain question answering, in: Proceedings of the 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 163–176.
 - [16] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human-generated MACHine Reading COMprehension dataset (2016).
 - [17] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, J. Lin, Making A MIRACL: Multilingual information retrieval across a continuum of languages, arXiv preprint arXiv:2210.09984 (2022).
 - [18] D. I. Adelani, M. Masiak, I. A. Azime, J. O. Alabi, A. L. Tonja, C. Mwase, O. Ogundepo, B. F. Dossou, A. Oladipo, D. Nixdorf, et al., MasakhaNews: News topic classification for African languages, arXiv preprint arXiv:2304.09972 (2023).
 - [19] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: BM25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

Run	Team	End-to-End	Model Type	nDCG@20	MRR@10	MAP	Recall@100
bm25-dt-mT5-pft-rerank	h2oloo	DT	Reranking	0.5700	0.6952	0.3610	0.5902
dt.plaid	HLTCOE	DT	Dense	0.4743	0.5846	0.3088	0.5733
plaid-xlmr.mlmfine.tt	HLTCOE	✓	Dense	0.4335	0.5625	0.2711	0.5256
plaid-xlmr.mlmfine.tt.jholo	HLTCOE	✓	Dense	0.3601	0.4886	0.2372	0.4829
plaid-xlmr.tt	HLTCOE	✓	Dense	0.3557	0.5693	0.2468	0.5083
plaid-xlmr.mlmfine.et	HLTCOE	✓	Dense	0.3488	0.5524	0.2461	0.5326
plaid-xlmr.et	HLTCOE	✓	Dense	0.3481	0.5153	0.2430	0.5237
bm25-dt-afrint5-rerank	h2oloo	DT	Reranking	0.3152	0.5306	0.1993	0.4186
afroxlmr_..._ft_ckpt2000	Masakhane	✓	Dense	0.2830	0.4288	0.1364	0.3732
afro_xlmr_..._sw_mrtydi_ft	Masakhane	✓	Dense	0.276	0.4921	0.1566	0.4148
splade.bm25-mt5-rerank	h2oloo	✓	Reranking	0.2530	0.4159	0.1384	0.2256
splade-mt5-rerank	h2oloo	✓	Reranking	0.2530	0.4159	0.1384	0.2256
afroxlmr_base_..._ckpt1000	Masakhane	✓	Dense	0.2451	0.3908	0.1232	0.3576
afriberta_base_ckpt25k	Masakhane	✓	Dense	0.2085	0.3567	0.1153	0.3342
dt.bm25-rm3	HLTCOE	DT	Sparse	0.2015	0.3571	0.1260	0.3359
splade-afrint5-rerank	h2oloo	✓	Reranking	0.1771	0.3687	0.095	0.1757
afriberta_base_..._mrtydi_ft	Masakhane	✓	Dense	0.1417	0.2408	0.0738	0.2187
hybrid_afriberta_dpr_splade	h2oloo	✓	Hybrid	0.1405	0.3143	0.0706	0.2567
hybrid_mdpr_msMarco_clir_splade	h2oloo	✓	Hybrid	0.0895	0.2174	0.0446	0.1436
afriberta_base_..._sw_miracl	Masakhane	✓	Dense	0.0895	0.1744	0.0418	0.2122
afriberta_..._sw_miracl_ft_ckpt100	Masakhane	✓	Dense	0.0846	0.2113	0.0376	0.1264

Table 8
Hausa Results.

Run	Team	End-to-End	Model Type	nDCG@20	MRR@10	MAP	Recall@100
dt.plaid	HLTCOE	DT	Dense	0.5118	0.7102	0.3567	0.6436
bm25-dt-mT5-pft-rerank	h2oloo	DT	Reranking	0.4935	0.6542	0.3387	0.5581
plaid-xlmr.mlmfine.tt	HLTCOE	✓	Dense	0.3366	0.5414	0.2115	0.4534
plaid-xlmr.mlmfine.tt.jholo	HLTCOE	✓	Dense	0.3117	0.5249	0.1892	0.4277
plaid-xlmr.et	HLTCOE	✓	Dense	0.2915	0.5513	0.1881	0.4332
plaid-xlmr.tt	HLTCOE	✓	Dense	0.2878	0.4783	0.1906	0.4373
plaid-xlmr.mlmfine.et	HLTCOE	✓	Dense	0.2760	0.4609	0.1911	0.4260
bm25-dt-afrint5-rerank	h2oloo	DT	Reranking	0.2676	0.4680	0.1718	0.3234
dt.bm25-rm3	HLTCOE	DT	Sparse	0.2550	0.3978	0.1789	0.4210
splade.bm25-mt5-rerank	h2oloo	✓	Reranking	0.2445	0.4440	0.1402	0.2372
splade-mt5-rerank	h2oloo	✓	Reranking	0.2445	0.4440	0.1402	0.2372
afroxlmr_..._ft_ckpt2000	Masakhane	✓	Dense	0.2209	0.3846	0.1119	0.2975
afro_xlmr_..._sw_mrtydi_ft	Masakhane	✓	Dense	0.2165	0.4004	0.1219	0.3337
afroxlmr_base_..._ckpt1000	Masakhane	✓	Dense	0.2067	0.3840	0.1087	0.3083
splade-afrint5-rerank	h2oloo	✓	Reranking	0.1647	0.3691	0.0862	0.1645
hybrid_afriberta_dpr_splade	h2oloo	✓	Hybrid	0.1644	0.3191	0.0928	0.2467
afriberta_base_ckpt25k	Masakhane	✓	Dense	0.1505	0.2845	0.0789	0.2459
hybrid_mdpr_msMarco_clir_splade	h2oloo	✓	Hybrid	0.1198	0.2730	0.0725	0.1765
afriberta_base_..._mrtydi_ft	Masakhane	✓	Dense	0.1034	0.2206	0.0436	0.1663
afriberta_base_..._sw_miracl	Masakhane	✓	Dense	0.0965	0.1713	0.0546	0.1780
afriberta_..._sw_miracl_ft_ckpt100	Masakhane	✓	Dense	0.0830	0.1589	0.0453	0.1407

Table 9
Somali Results.

Run	Team	End-to-End	Model Type	nDCG@20	MRR@10	MAP	Recall@100
bm25-dt-mT5-pft-rerank	h2oloo	DT	Reranking	0.5232	0.7222	0.3110	0.5473
dt.plaid	HLTCOE	DT	Dense	0.5118	0.7102	0.3567	0.6436
plaid-xlmr.mlmfine.tt	HLTCOE	✓	Dense	0.4230	0.6175	0.2500	0.4645
plaid-xlmr.mlmfine.tt.jholo	HLTCOE	✓	Dense	0.4081	0.6002	0.2362	0.4477
plaid-xlmr.tt	HLTCOE	✓	Dense	0.3347	0.5615	0.2041	0.4510
plaid-xlmr.mlmfine.et	HLTCOE	✓	Dense	0.3301	0.5820	0.2093	0.4487
afroxlmr_..._ft_ckpt2000	Masakhane	✓	Dense	0.3215	0.4983	0.1378	0.3543
plaid-xlmr.et	HLTCOE	✓	Dense	0.3182	0.5541	0.1977	0.4616
afroxlmr_base_..._ckpt1000	Masakhane	✓	Dense	0.3037	0.5190	0.1302	0.3585
bm25-dt-afritm5-rerank	h2oloo	DT	Reranking	0.2542	0.4729	0.1351	0.2899
afro_xlmr_..._sw_mrtydi_ft	Masakhane	✓	Dense	0.2447	0.4621	0.1118	0.3499
splade.bm25-mt5-rerank	h2oloo	✓	Reranking	0.2395	0.4764	0.1055	0.2105
splade-mt5-rerank	h2oloo	✓	Reranking	0.2395	0.4764	0.1055	0.2105
dt.bm25-rm3	HLTCOE	DT	Sparse	0.2178	0.4172	0.1353	0.3340
afriberta_base_ckpt25k	Masakhane	✓	Dense	0.1833	0.3968	0.0792	0.2486
afriberta_base_..._mrtydi_ft	Masakhane	✓	Dense	0.1426	0.3143	0.0532	0.2207
splade-afritm5-rerank	h2oloo	✓	Reranking	0.1378	0.3043	0.0537	0.1215
hybrid_afriberta_dpr_splade	h2oloo	✓	Hybrid	0.1277	0.3000	0.0565	0.1930
afriberta_base_..._sw_miracl	Masakhane	✓	Dense	0.1109	0.2141	0.0463	0.1897
afriberta_..._sw_miracl_ft_ckpt100	Masakhane	✓	Dense	0.1068	0.2046	0.0395	0.1442
hybrid_mdpr_msmarco_clir_splade	h2oloo	✓	Hybrid	0.0909	0.2338	0.0435	0.1807

Table 10
Swahili Results.

Run	Team	End-to-End	Model Type	nDCG@20	MRR@10	MAP	Recall@100
bm25-dt-mT5-pft-rerank	h2oloo	DT	Reranking	0.5819	0.6211	0.4512	0.8057
dt.plaid	HLTCOE	DT	Dense	0.4793	0.6036	0.3657	0.7240
plaid-xlmr.mlmfine.tt.jholo	HLTCOE	✓	Dense	0.4297	0.5438	0.3044	0.6748
plaid-xlmr.mlmfine.tt	HLTCOE	✓	Dense	0.4189	0.5434	0.2985	0.6394
bm25-dt-afritm5-rerank	h2oloo	DT	Reranking	0.4103	0.5014	0.3162	0.6377
splade.bm25-mt5-rerank	h2oloo	✓	Reranking	0.4071	0.5822	0.2808	0.5037
splade-mt5-rerank	h2oloo	✓	Reranking	0.4071	0.5822	0.2808	0.5037
plaid-xlmr.mlmfine.et	HLTCOE	✓	Dense	0.3804	0.5520	0.2707	0.6950
dt.bm25-rm3	HLTCOE	DT	Sparse	0.3555	0.4909	0.2696	0.6273
plaid-xlmr.tt	HLTCOE	✓	Dense	0.3522	0.4731	0.2505	0.5784
splade-afritm5-rerank	h2oloo	✓	Reranking	0.3138	0.4996	0.2130	0.4100
plaid-xlmr.et	HLTCOE	✓	Dense	0.2627	0.4214	0.1863	0.5176
afriberta_base_ckpt25k	Masakhane	✓	Dense	0.2357	0.3763	0.1423	0.4369
afro_xlmr_..._sw_mrtydi_ft	Masakhane	✓	Dense	0.2296	0.4080	0.1435	0.4418
afroxlmr_..._ft_ckpt2000	Masakhane	✓	Dense	0.2210	0.4009	0.1186	0.377
hybrid_afriberta_dpr_splade	h2oloo	✓	Hybrid	0.2015	0.3454	0.1205	0.3786
afroxlmr_base_..._ckpt1000	Masakhane	✓	Dense	0.1981	0.3500	0.1071	0.3947
afriberta_base_..._mrtydi_ft	Masakhane	✓	Dense	0.1688	0.2999	0.0880	0.3278
hybrid_mdpr_msmarco_clir_splade	h2oloo	✓	Hybrid	0.1642	0.2921	0.0968	0.3452
afriberta_base_..._sw_miracl	Masakhane	✓	Dense	0.1634	0.2612	0.0922	0.3658
afriberta_..._sw_miracl_ft_ckpt100	Masakhane	✓	Dense	0.1594	0.2712	0.0865	0.3063

Table 11
Yoruba Results.

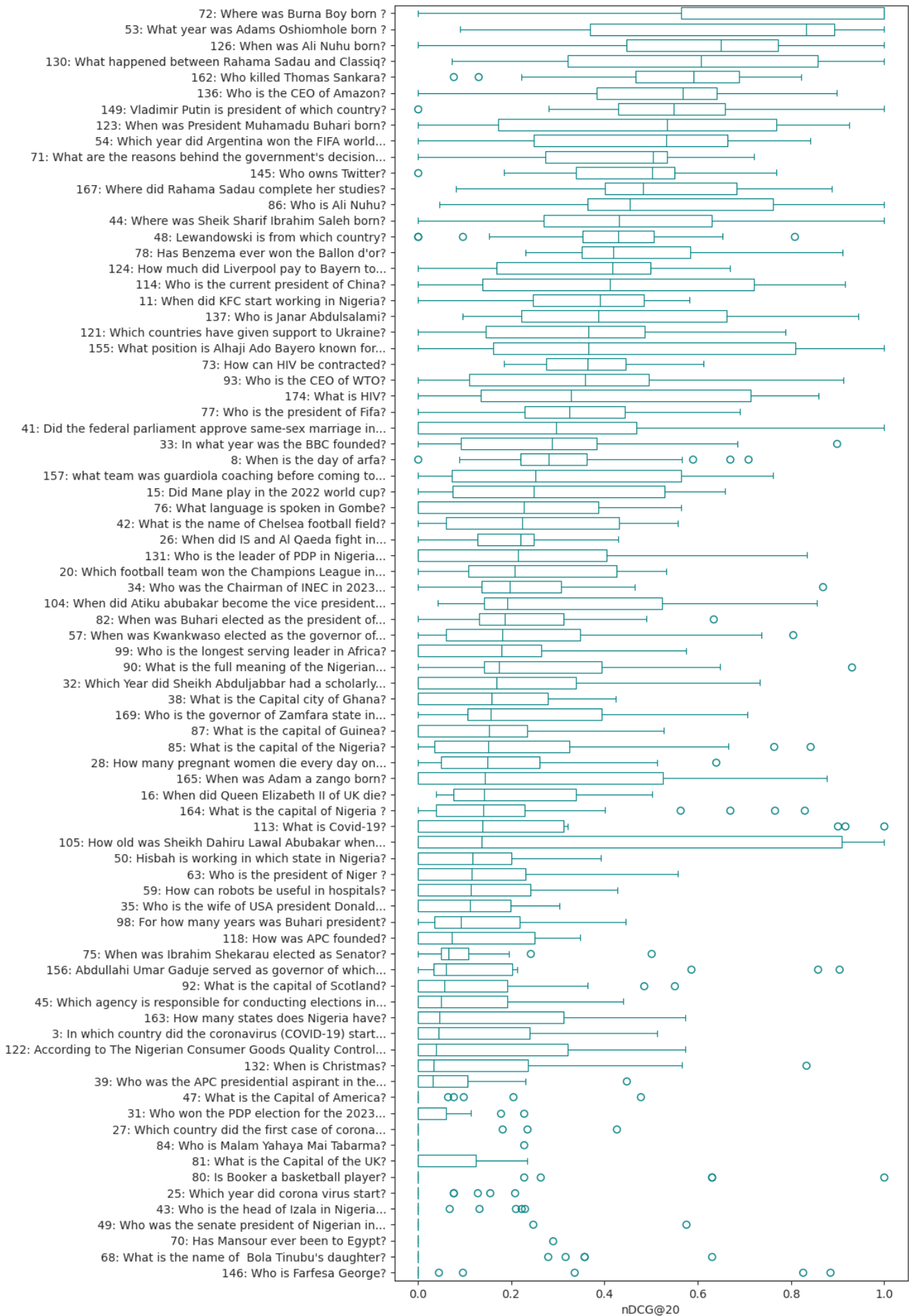


Figure 4: Boxplots showing nDCG@20 for Hausa Queries.

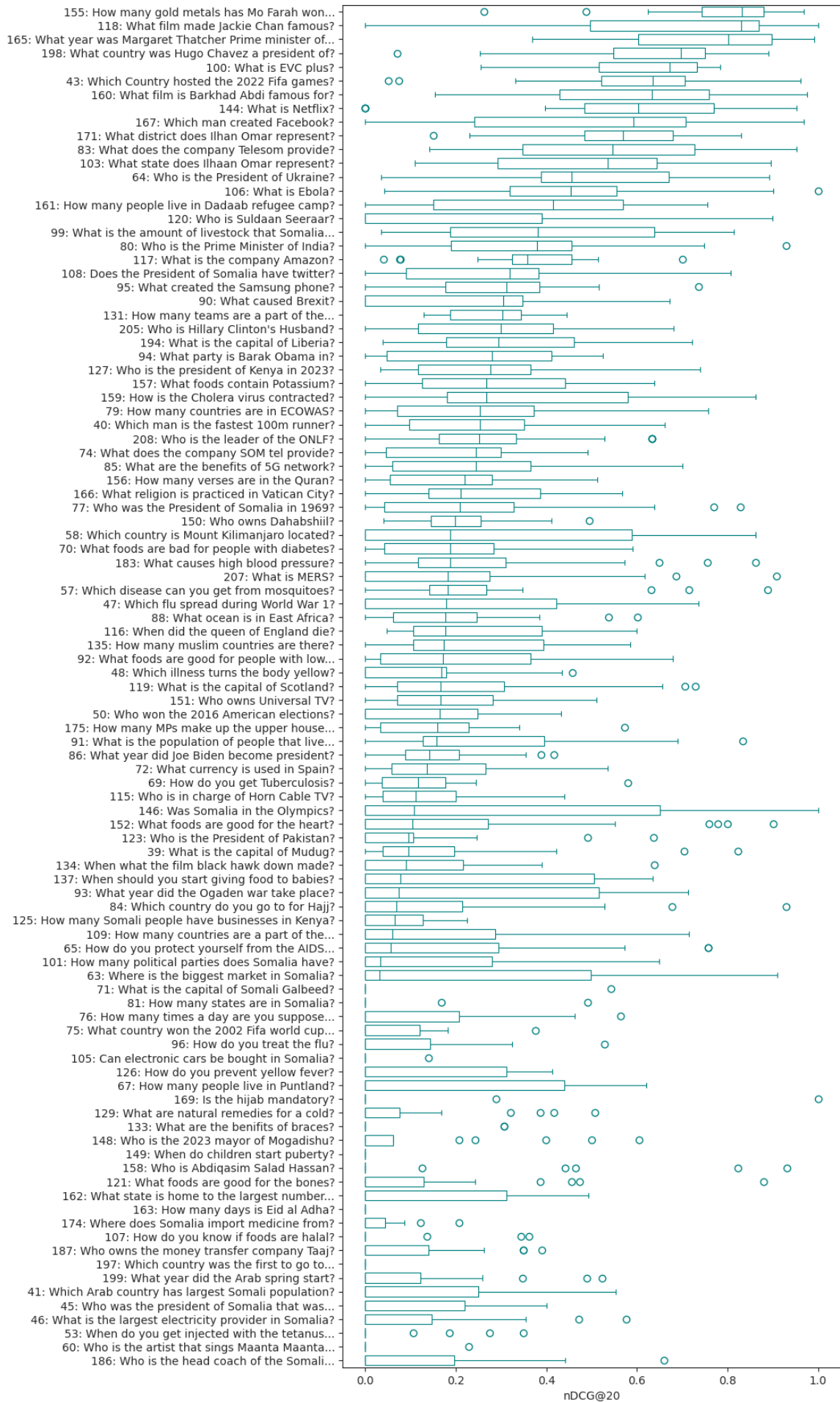


Figure 5: Boxplots showing nDCG@20 for Somali Queries

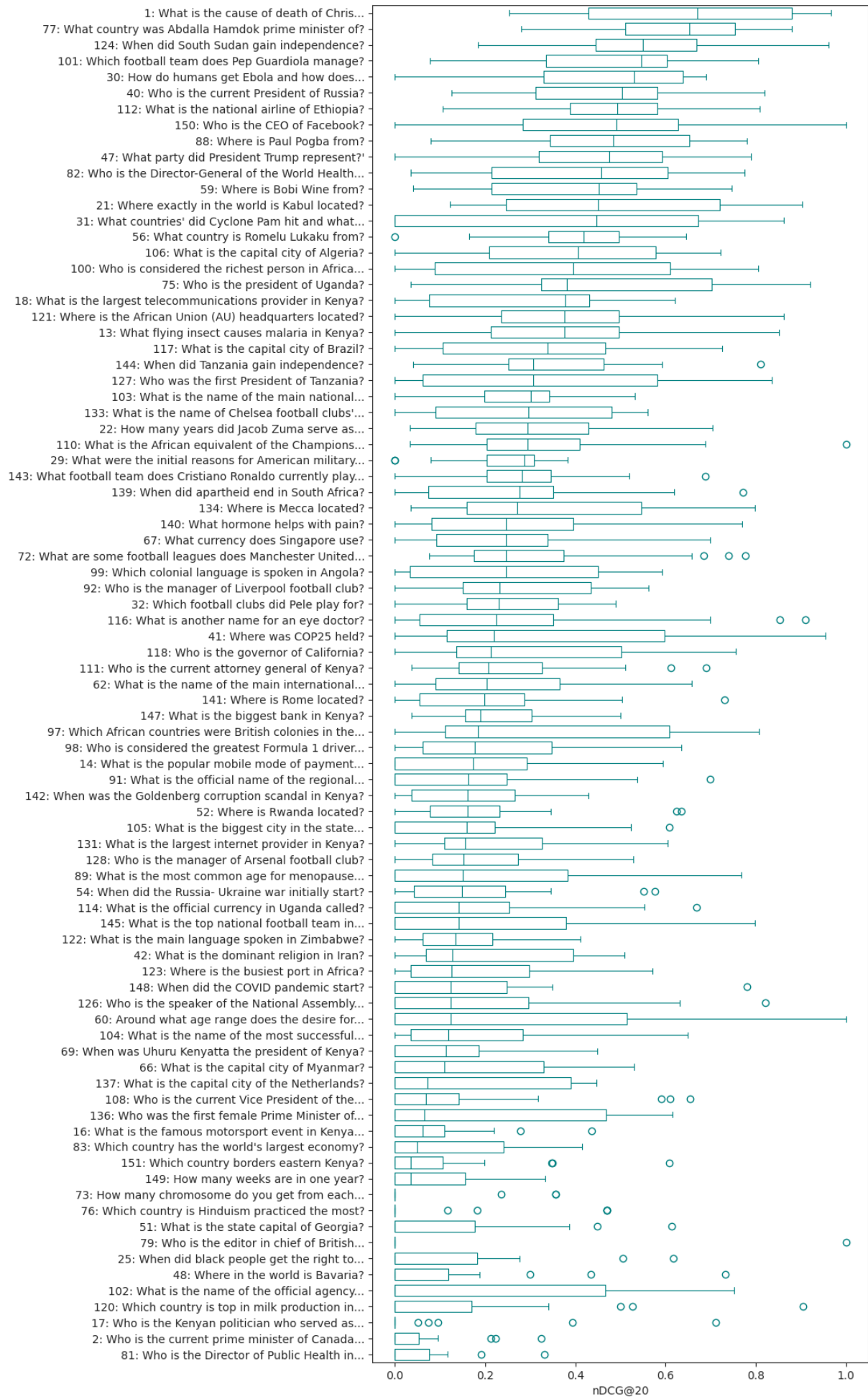


Figure 6: Boxplots showing nDCG@20 for Swahili Queries

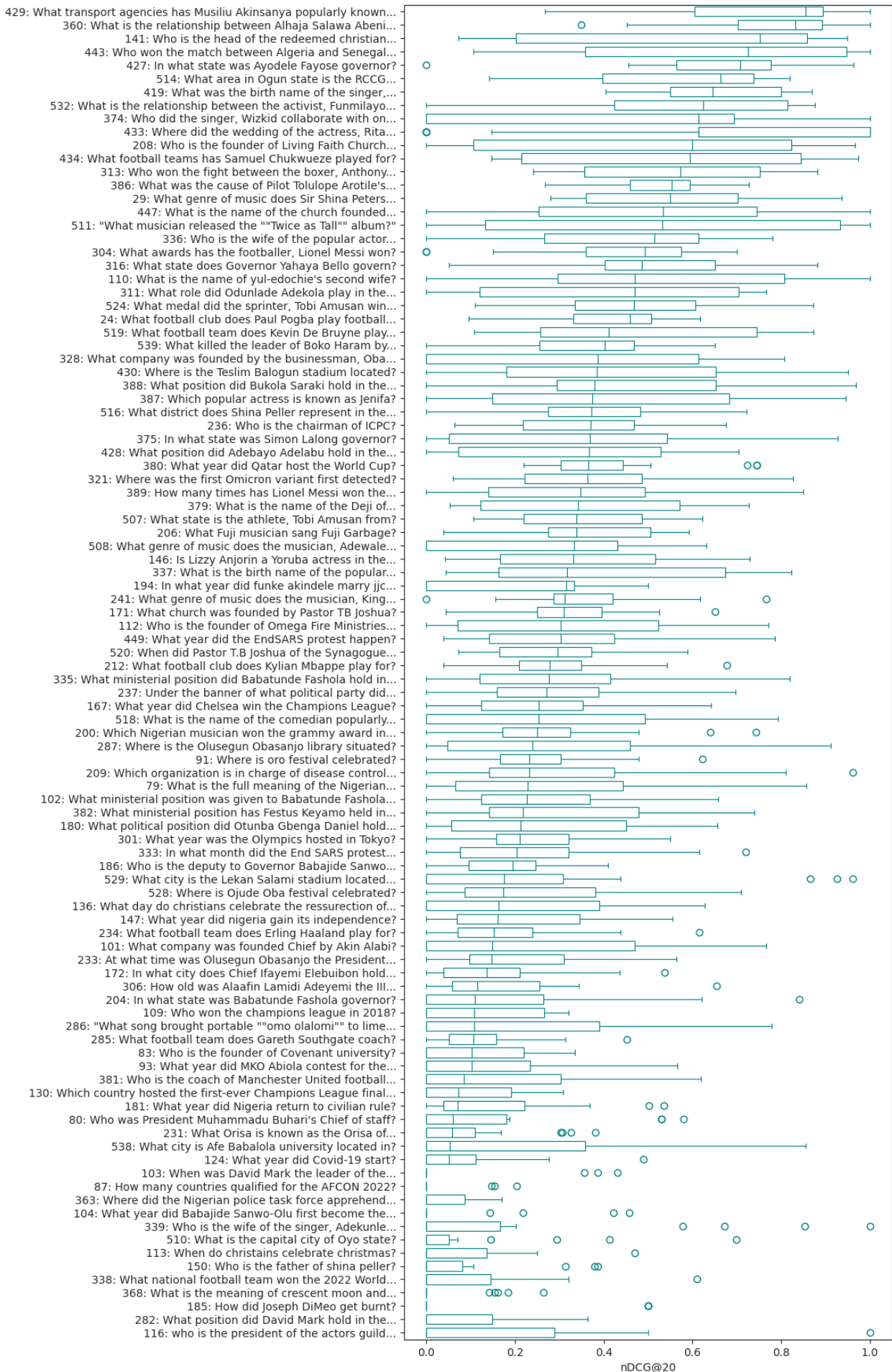


Figure 7: Boxplots showing nDCG@20 for Yoruba Queries