Y-NQ:

English-Yorùbá Evaluation dataset for Open-Book Reading Comprehension with Open-Ended Questions

Anonymous ACL submission

Abstract

The purpose of this work is to share an English-Yorùbá evaluation dataset for open-book reading comprehension with open-ended questions to assess the performance of models both in a high- and a low-resource language. The dataset contains 358 questions and answers on 338 English documents and 208 Yorùbá documents. Experiments show a consistent disparity in performance between the two languages, with Yorùbá falling behind English for automatic metrics even if documents are much shorter for this language. For a small set of documents with comparable length, performance of Yorùbá drops by $x2.5$ times and this comparison is validated with human evaluation. When analyzing performance by length, we observe that Yorùbá decreases performance dramatically for documents that reach 1500 words while English performance is barely affected at that length. Our dataset opens the door to showcasing if English LLM reading comprehension capabilities extend to Yorùbá, which for the evaluated LLMs is not the case.

1 Introduction

This study explores the intersection of reading comprehension with open-ended questions, examining how models perform on a task requiring both in-context understanding (i.e., open-book model, where the model has access to the context document during inference to answer a particular question) and generative text production (i.e. the answer is free-text which has to be compared to a gold standard reference). We aim to investigate the performance of this task in two languages: a high-resource language (English) and a low-resource language (Yorùbá). For this, we introduce Y-NQ (Yorùbá Natural Questions) a comprehensive open-book question-answer dataset (Section 2). Y-NQ is sourced from NQ (Kwiatkowski et al., 2019) and provides a complete article context for informed answers, and parallel documents on the same topic for both high- and low-resource languages. The data set also includes the comparability of the responses in languages. As a result, we are increasing Natural Language Processing (NLP) resources in Yorùbá (Ahia et al., 2024). Our data set is benchmarked against state-of-the-art Large Language Models (LLMs). The results and analysis (Section 3) shows that responses in Yorùbá are more inaccurate than those in English. As a by-product of human annotations, we identify inaccuracies in the English-language version of some Wikipedia articles (26 incorrect answers out of 1,566 humanly analyzed questions in the English-language subset of articles), which confirms the existence of accuracy discrepancies across languages for the same Wikipedia topics, thus supporting, for example, the need to better interlink Wikipedia articles across languages (Klang and Nugues, 2016).

2 Dataset description

2.1 Requirements and Background

The performance of Reading Comprehension (RC) in LLMs has been explored in different settings. At the high level, RC tasks can fall under two main categories: open-book tasks, such as in SQuAD (Rajpurkar et al., 2016), and close-book tasks, such as in TriviaQA (Joshi et al., 2017). Response formats vary across RC tasks as well and include: true/false classification (e.g., BoolQ; Clark et al., 2019), multiple-choice questions (e.g., Belebele), span selection (e.g., SQuAD), and text generation (e.g., NQ or TriviaQA).

Since we are interested in exploring the intersection of reading comprehension with open-ended questions covering both a high- and a low-resource language, we can explicitly set our requirements to include for each of the two types of language: (a) long articles (>100s words), (b) question-answer pairs with lengthy answers (>10s words), and (c) equivalence annotations for cross-lingual answers. Since there are no existing data sets to this effect, we extend existing research by tailoring an established data set to our specific requirements. We justify our choice of data sets and low-resource language selection as explained in the following.

Dataset. Among the open-book reading comprehenstion with open-ended questions, one of the largest datasets with multilingual information available is NQ which is shared under the license Creative Commons Share-Alike 3.0 .

Low-resource language. There is a large number of low-resource languages that could be explored here. We prioritize a low-resource language that has overall limited digital resources (in compliance with the definition of low resource), but has a high representation in Wikipedia (on the order of several thousands of entries) and a significant number of speakers (in the order of tens of millions), and makes use of the same script (Latin) as the high-resource language in which results are compared. One of the languages that complies with all these criteria is Yorùbá, in which we can also find works on comprehension of the language in the domain of language exams (Aremu et al., 2024), based on short passages and multiple choice answers. Another work is the AfriQA dataset (Ogundepo et al., 2023) for answering open-retrieval questions, with a primary focus on retrieving correct answers that are answerable on Wikipedia. However, this cannot be used as an open book. Finally, Bebebele (Bandarkar et al., 2024) also includes Yorùbá, although it uses short passages and multiple choice answers.

2.2 Dataset creation

NQ pre-selection. We looked at 315,203 examples and 231,695 unique English Wikipedia pages from the NQ training and validation datasets. We filter questions for only those where every long answer is contained in an html tag $< p >$ where $< p >$ is the first identified html tag in the long answer span. This filters out about 25 percent of the questions.

We extracted 2,855 Yorùbá Wikipedia pages that are actively associated with the above English pages. We removed documents with fewer than 500 characters, including formatting, and performed multiple cleaning procedures, such as removing html formatting, removing citation notations, and filtering out irrelevant sections in Wikipedia articles (e.g., references, tables). 664 Yorùbá documents and 1,566 questions were sent for human annotation. We tried a pre-annotation effort to automatically reduce the workload but since it did not work, we only report it in the Appendix A.

Annotation guidelines and requirements. We designed the annotation guidelines as follows. We provided context on the objective of the task together with the project context and description of the task. The guidelines are summarized in Table 1.

Finally, beyond the guidelines, we provided additional examples and requested that annotators should be native speakers of the language of the source documents and should have at least CEFR C2 level proficiency in English.

Annotator findings. We noticed that many articles have a significant amount of English content. Several documents also contained errors, such as incorrect spelling, ungrammatical sentences, and sentences that lacked clarity or meaning. We disregarded such articles and corrected articles that were contaminated with a small amount of English content. We also removed the entries where no answers could be found in the Yorùbá articles.

Following the guidelines, the annotators encountered the following: (a) questions with multiple correct answers, for which they annotated each correct answer for the question; (b) questions with correct answers in Yorùbá, but incorrect in English, where they annotated the Yorùbá appropriately, but flagged the English portion incorrect (there were 26 questions in the category); (c) unclear questions (5 questions) to which no annotations were assigned; (d) answers existing in multiple paragraphs in the document for which they annotated the row with all paragraphs. There were 456 Yorùbá documents that did not answer the question; therefore, we discarded those. Only eight incorrect English answers from the previous 26

Objective	Read an article and find a paragraph containing enough information to answer a specific question.
Project Context	Evaluate accuracy of large language models in finding long contexts and short answers; extend Natural Questions dataset to multilingual, non-English centric.
Task Components	• QUESTION: Simple question requesting information or explanation. • ARTICLE: Numbered paragraphs containing relevant information.
Task Steps	1. Read QUESTION carefully. 2. Read ARTICLE paragraphs until sufficient information is found. 3. Record findings by answering task questions.
Additional task steps	Discard questions that contain the answer in English in the Yorùbá document When possible, add Yorùbá questions, translate them into English, and find answers both in the Yorùbá and English documents.

Table 1: Linguistic guidelines and annotation

	ENG	YOR
#Q&A	358	358
#DOCS	338	208
AVG. DOC LEN	10363	430
MEDIAN DOC LEN	9272	172
AVG. QUESTION LEN	8.86	9.39
AVG. LONG ANSWER LEN	113.80	32.89

Table 2: Dataset Statistics. Length is in words.

	LAN	R-1	R-2	R-L
GPT4O	ENG	0.39	0.23	0.30
	YOR	0.34	0.19	0.27
O1MINI	ENG	0.45	0.22	0.30
	YOR	0.30	0.14	0.22
LLAMA	ENG	0.31	0.18	0.23
	YOR	0.20	0.15	0.18

Table 3: Results for 3 LLM in terms of Rouge computed for the entire set of questions.

remain in the final dataset, and we did not correct them since the English documents remained the same as in the original NQ.

Statistics. Table 2 details the statistics of the data set.Our carefully curated selection contains 208 unique Yorùbá Wikipedia documents with an average word count of 430, and 358 questions. Only the questions are strictly comparable. English and Yorùbá documents are not comparable in number or length, but they are so in topic and domain. The answers are not comparable in length. Notice that English documents outnumber Yorùbá documents mainly due to multiple versions of the same English topic counted as different documents, while in Yorùbá we selected one version of the document and multiple topics in English that correspond to the same Yorùbá topic.

The fact that English documents are longer than those in Yorùbá makes the task easier for Yorùbá, since documents are significantly shorter within the same topic or domain. We identified a subset of six documents that are strictly comparable in length and topic for English and Yorùbá, which allows us to make a fair comparison. Table 5 in Appendix B shows the list of fields in Y-NQ and a sample entry.

3 Experiments

Baselines We evaluate our dataset with GPT-4o[1] (et al., 2024b), o1-mini[2], and LlaMA-3.1-8b (et al., 2024a), thereby covering both open and closed models, as well as models of different sizes. For each Y-NQ entry, we prompt the models with the following formatted instructions.

```
"""
Given the following passage and
a question,answer the question
in a single paragraph with
information found in the passage.

####
PASSAGE
{document}

####
QUESTION
{question}

####
ANSWER
"""
```

[1] gpt-4o version 2024-08-06
[2] o1-mini version 2024-09-12

Evaluation. We evaluate the results by comparing the generated text and the reference long answer using several Rouge (Lin, 2004) versions (Rouge-1, Rouge-2, Rouge-L).

Automatic metrics. Table 3 reports the results showing that Yorùbá consistently performs worse than English (e.g., losing 0.4 in Rouge-1). However, the Yorùbá task is much easier because the documents are much shorter, which means that answering the question becomes an easier task. Even if we prompt the model to only answer based on the in-context document, we can not discard the idea that English may get better results due to using the internal knowledge from the model.

Length analysis. Model performance changes with the length of the document, as shown in Figure 1. The dataset was split into equal size of documents in each length bucket. We can see a drop in performance when the Yorùbá documents reach 1,500 words, which shows the challenges that current models face in long-context understanding of low-resource languages.

Comparable documents. For a small portion of long-enough documents of comparable length between English and Yorùbá (only 4 documents that are over 900 words long), English performance demonstrates a significant edge (1.58X-2.56X), see Table 4.

Human evaluation. For the comparable documents, we performed a human evaluation. A bilingual proficiency speaker of English and Yorùbá evaluated the output of the models. Evaluation was performed by using a Likert scale from 1-3, being 3 a perfect response. On average, English responses across models scored 2.33, while Yorùbá responses scored 2. Appendix C reports a complete example with human evaluation for all translation outputs.

4 Conclusions

Y-NQ is a newly released dataset that enables to compare generative open-book reading comprehension between English and Yorùbá. The main contributions of our data set are to allow for the comparison of LLM results in a reading comprehension task across a high- and a low-resource language, showing what are the generalization capabilities of LLMs in this particular case. Moreover, our annotations confirmed variations in the accuracy of Wikipedia articles in all languages. In particular, we identify inaccurate English responses for Yorùbá language-specific content. Y-NQ allows us to evaluate how reading comprehension capabilities extend to Yorùbá. Y-NQ is not exactly comparable in its totality between languages. Given that Yorùbá has shorter documents than English, the reading comprehension task is easier for Yorùbá. Therefore, results on this language should be much better than in English to expect parity between languages. Our experiments show that the reading comprehension capabilities of current English LLMs do not extend to Yorùbá. Y-NQ is freely available[3].

	LANG	R-1	R-2	R-L	Hum
GPT4O	ENG	0.45	0.23	0.30	2.50
	YOR	0.32	0.09	0.19	2.75
O1MINI	ENG	0.43	0.17	0.27	2.50
	YOR	0.27	0.06	0.17	2.25
LLAMA	ENG	0.46	0.28	0.33	2.00
	YOR	0.09	0.05	0.07	1.00

Table 4: Results and human evaluation (Hum) for for comparable English and Yorùbá documents. English documents have an average length of 3299 and Yorùbá have an average length of 3070 words.

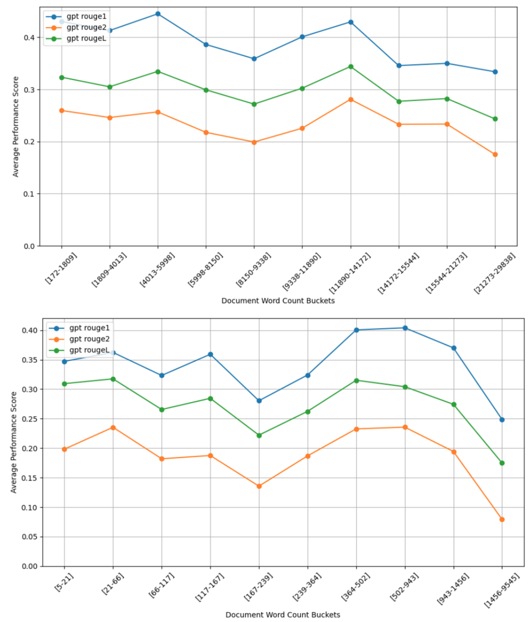

Figure 1: Impact of Document Length Buckets on Performance Scores for English (top) and Yorùbá (bottom) for GPT-4 outputs

[3] BLIND

Limitations and Ethical considerations

Y-NQ is limited in size, language, and domain coverage. The fact of using Wikipedia and extending an existing open-source dataset (NQ) may play in favor of having higher results in both languages due to contamination. Furthermore, the data set is not fully comparable between English and Yorùbá, since documents and answers vary in length.

Our experimentation is limited to models and automatic evaluation metrics, which is compensated for through a small-size human evaluation. Annotators were paid a fair rate and they gave consent to the use of the data that they were annotating. Annotators are included as authors of the paper.

References

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: NLP resources and models for Yorùbá regional dialects. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4392–4409, Miami, Florida, USA. Association for Computational Linguistics.

Anuoluwapo Aremu, Jesujoba O. Alabi, Daud Abolade, Nkechinyere F. Aguobi, Shamsuddeen Hassan Muhammad, and David Ifeoluwa Adelani. 2024. Naijarc: A multi-choice reading comprehension dataset for nigerian languages. *Preprint*, arXiv:2308.09768.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Clark, Matthew Gardner, Tom Fevry, and Robert Weischedel. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Abhimanyu Dubey et al. 2024a. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

OpenAI et al. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale question-answer dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Marcus Klang and Pierre Nugues. 2016. Pairing Wikipedia articles across languages. In *Proceedings of the Open Knowledge Base and Question Answering Workshop (OKBQA 2016)*, pages 72–76, Osaka, Japan. The COLING 2016 Organizing Committee.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics.*

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro, and Sonia Adhiambo. 2023. Cross-lingual open-retrieval question answering for African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

A Pre-annotation effort

In order to reduce the annotation workload, we automatically pre-selected Yorùbá sentences that could be good response candidates by computing a similarity score. If the answer to the question was in agreement with a high similarity score, the annotator would save time by looking through the document and only checking if the match was correct. We conducted a SONAR embedding similarity (Duquenne et al., 2023) analysis between Yorùbá documents and long English answers. We used the Stopes[4] sensitizers on all text extracted from $< p >$ elements for both the scraped Yorùbá Wikipedia articles downloaded from the previous step and the original NQ Wikipedia pages. We then created SONAR embeddings of each extracted sentence and identified those sentences in the Yorùbá pages which were most similar to sentences in the long English answers based on their cosine similarity scores. For a small set of samples, we asked the annotators to examine the entries in a small validation data set to identify a reasonable threshold indicating high similarity between Yorùbá/English sentences, which could then be applied to the rest of the data set. The analysis shows a low similarity matching rate, which is likely due to the low quality and short length of many Yorùbá articles and/or SONAR embeddings not being suitable for such a task. Given this low reliability, we abandoned this automatic pre-annotation, which would not reduce annotation efforts.

B Dataset fields and example entry

Table 5 reports the dataset fields, descriptions and sample entry.

C Human Evaluation

Table 6 presents a complete sample and its human scores for all models outputs.

[4]https://github.com/facebookresearch/stopes

Field	Description	Example
1. Question ID	Unique identifier	3506772758530306034
2. English Document	English text document	
3. English Question	Question in English	what is the name of the first nigerian president
4. English Long Answer	Detailed answer in English	.ky is the Internet country code top-level domain (ccTLD) for the Cayman [..]
5. English Short Answer	Brief answer in English	Nnamdi Azikiwe
6. Yorùbá Document	Yorùbá text document	
7. Yorùbá Rewrite Flag	Was Yorùbá document rewritten? (0: no, 1: yes)	1
8. Yorùbá Question	Question in Yorùbá	kí ni ky dúró fún ní erékùṣù cayman
9. Yorùbá Short Answer	Brief answer in Yorùbá	Nnamdi Azikiwe ni Aare
10. Yorùbá Long Answer	Detailed answer in Yorùbá	Nnamdi Azikiwe ti o je Gomina Agba nigbana di Aare, ipo to je fun ayeye, [..]
11. Yorùbá Paragraph Info	Contextual information	P2
12. Answer Alignment	Semantic equivalence (0: not literal, 1: literal)	1

Table 5: Dataset Fields, Descriptions and Sample entry.

Field	Content	Hum
English question	where does the vaal dam get its water from	–
Yorùbá question	níbo ni ìsédò vaal ti ń gba omi rẹ̀	–
English long-answer	The Vaal Dam in South Africa was constructed in 1938 and lies 77 km south of OR Tambo International Airport. The lake behind the dam wall has a surface area of about 320 square kilometres (120 sq mi) and is 47 meters deep. The Vaal Dam lies on the Vaal River, which is one of South Africa 's strongest - flowing rivers. Other rivers flowing into the dam are the Wilge River, Klip River, Molspruit and Grootspruit. It has over 800 kilometres (500 mi) of shoreline and is South Africa 's second biggest dam by area and the fourth largest by volume.	–
Yorùbá long-answer	Àwọn òdò mìíràn tí ńṣàn sínu ìdídò náà ni Odò Wilge,Odò Klip Molspruit ati Grootspruit. ', 'Ìdídò Vaal wà lórí Odò Vaal, èyí tí ó jẹ̀ ọkan nínú àwọn òdò tí ńṣàn tí ó lágbára jùlọ ní orílẹ̀-èdè South Africa.	–
English GPT4o-answer	The Vaal Dam gets its water from several sources, including the Vaal River, into which other rivers like the Wilge River, Klip River, Molspruit, and Grootspruit also flow. Additionally, the Lesotho Highlands Water Project contributes to the water supply by piping water from Lesotho into the Liebenbergsvlei and Wilge Rivers. The Sterkfontein Dam is also part of the water transfer scheme, releasing water to boost the levels in the Vaal River System when the Vaal Dam drops below 16%.	2
Yorùbá GPT4o-answer	Ìsédò Vaal ń gba omi rẹ̀ láti ọ̀pọ̀lọpọ̀ odò, tí wọ́n ń fi kún omi láàrin ìdídò náà. Àwọn odò jíjẹ̀ wọ̀nyí ní Odò Wilge, Odò Klip, Molspruit, àti Grootspruit. Pẹ̀lú ẹ̀yin odò wọ̀nyí, àwùjọ omi ètò omi Ilè-Okè ní Lesotho tun ń pèsè omi sínú ìdídò náà.	3
English o1mini-answer	The Vaal Dam receives its water from multiple sources, primarily the Vaal River itself along with several tributaries including the Wilge River, Klip River, Molspruit, and Grootspruit. Additionally, the Lesotho Highlands Water Project contributes to the dam's water supply by piping water from Lesotho into the Liebenbergsvlei and Wilge Rivers. Furthermore, the Sterkfontein Dam plays a crucial role in augmenting the Vaal Dam's water levels through the Tugela-Vaal water transfer scheme, which transfers water from the Thukela River in KwaZulu-Natal. These combined sources ensure a stable and substantial inflow of water to the Vaal Dam, supporting both domestic and industrial needs in the surrounding regions.	3
Yorùbá o1mini-answer	Ìsédò Vaal ń gba omi rẹ̀ láti Odò Vaal àti àwọn òdò tí ń darapọ̀ mọ́ra bí Odò Wilge, Odò Klip Molspruit, àti Odò Grootspruit. Pẹ̀lú náà, omi tó wá látinú Ilẹ̀ Lesotho nípasẹ̀ ètò gíráfítì ń fi omi ranṣẹ́ sí Odò Liebenbergsvlei àti Odò Wilge. Bakan náà, Dámù Sterkfontein, tó jẹ́ apá kan ti ètò gbígbé omi Tugela Vaal, ń gba omi láti Odò Thukela ní KwaZulu-Natal, tí ń kó ipa tó lágbára sí ìkún omi Vaal Dam. Àwọn orísun omi wọ̀nyí jọ ń fún Dámù Vaal ní agbára tó pọ̀ láti pèsè omi fún àwùjọ àwọn ìpínlẹ̀ àti àwọn iṣẹ́ amáyédẹrùn ní Gúusu Afrika.	2
English LlaMa-answer	The Vaal Dam gets its water from several rivers, including the Vaal River, the Wilge River, the Klip River, the Molspruit, and the Grootspruit. Additionally, the Lesotho Highlands Water Project feeds water into the system by gravity, and the Sterkfontein Dam contributes to the Vaal River System through the Tugela-Vaal water transfer scheme. Water is also piped from Lesotho into the Liebenbergsvlei and Wilge Rivers. The water from the Sterkfontein Dam is released once the Vaal Dam drops to below 16%. This stable supply of water is crucial for the people and industrial complex of Gauteng.	2
Yorùbá LlaMA-answer	Vaal Dam j?? ìsòro omi sí 2,609,799,000 cubic mítà (9.21642×1010 cu ft) àti síwájú síi 663,000,000 cubic mítà (2.34×1010 cu ft) tàbí ìdá m??rìndínl??gb??n (26%) le wà ní ìpam?? fún ìgbà dí?? fún ìdínkù i?àn omi. Ìdídò náà tún di gbígbé sókè ní ìb??r??p??p?? ?dún 50s sí gíga ìw??n 60.3 mítà (198 ft) èyí tí o fi kún agbara r?? láti di 2,188,000,000 cubic mítà (7.73×1010 cu ft). Ìgbéga kejì wáyé ní ?dún 1985 nígbà tí odi di gbígbé sókè nípas?? 3.05 mítà (10.0 ft) sí 63.4 mítà (208 ft) lékè ìpìl?? tí ó súnm?? il?? jùl?. Agbára ìdídò l??w??l??w?? j?? 2,609,799,000 cubic mítà (9.21642×1010 cu ft) àti síwájú síi 663,000,000 cubic mítà (2.34×1010 cu ft) tàbí ìdá m??rìndínl??gb??n (26%) le wà ní ìpam?? fún ìgbà dí?? fún ìdínkù i?àn omi.	1

Table 6: Example of Human Evaluation scores for all models.