# CONTEXTUAL EVALUATION OF LLM'S PERFORMANCE ON PRIMARY EDUCATION SCIENCE LEARNING CONTENTS IN THE YORUBA LANGUAGE

**Olanrewaju Lawal**
Research and Innovation
Data Science Nigeria
lawal@datasciencenigeria.ai

**Olubayo Adekanmbi**
Research and Innovation
Data Science Nigeria
bayo@datasciencenigeria.ai

**Anthony Soronnadi**
Research and Innovation
Data Science Nigeria
anthony@datasciencenigeria.ai

## ABSTRACT

In the rapidly evolving era of artificial intelligence, Large Language Models (LLMs) like ChatGPT-3.5, Llama, and PaLM 2 play a pivotal role in reshaping education. Trained on diverse language data with a predominant focus on English, these models exhibit remarkable proficiency in comprehending and generating intricate human language constructs, revolutionizing educational applications. This potential has prompted exploration into personalized and enriched educational experiences, streamlining instructional design to cater to students' needs. However, the inclusive effectiveness of LLMs in low-resource languages, like Yoruba, poses challenges. This research critically assesses the ability of LLMs, including ChatGPT-3.5, Gemini, and PaLM 2, to comprehend and generate contextually relevant science education content in Yoruba. The study, conducted across four tasks using a manually developed primary science dataset in Yoruba, reveals a comparative underperformance in various NLP tasks, emphasizing the need for language-specific and domain-specific technologies, particularly for primary science education in low-resource languages.

## 1 INTRODUCTION

The development of large language models, often referred to as pre-trained language models, has revolutionized the fields of (NLP) and artificial intelligence (AI). This advancement has notably enhanced the field of education Gamieldien et al. (2023).

LLMs are trained on vast quantities of general data. This has endowed LLMs with strong capabilities to generate and interpret natural language, resulting in their extensive deployment in the field of NLP, as described by Wu et al. (2023). The field of NLP has experienced substantial breakthroughs, including the emergence of word embeddings, the development of the seq2seq or encoder-decoder framework Sutskever et al. (2014), and the introduction of the Transformer architecture Vaswani et al. (2017), which serves as the foundational element for contemporary NLP models like BERT. These advancements have fundamentally altered the landscape of education.

In the AI era, the education sector has been confronted with several challenges, such as diminished student engagement, uneven allocations of educational resources, and high dropout rates due to low comprehension, especially in rural communities Nita et al. (2021). Conventional classroom methods often fall short of addressing the unique needs of diverse groups of students. Large language models, which are potent tools used in natural language processing, hold the promise of transforming traditional educational approaches. LLMs can facilitate personalized learning experiences and intelligent tutoring systems. Moreover, the big data era has led to the accumulation of an extensive array of learning data in the educational field Gan et al. (2023). LLMs have become indispensable in delivering personalized tutoring experiences.

AI-driven systems swiftly evaluate a student's performance and offer customized feedback, guidance, and educational resources that are optimally aligned with that individual's learning style and academic background İlçin et al. (2018). Large Language Models are trained on extensive datasets encompassing multiple languages and domains, with a predominant focus on English. This broad training base has led to their significant application in education and they have been widely used in primary science personalised learning content delivery. Nevertheless, concerns have been raised about their effectiveness and their conceptual understanding of learning contents in languages that are underrepresented in the training data, especially the Yoruba language; which is the mother tongue for students in Nigeria's western rural communities. This is particularly relevant in light of research indicating that learning, especially in primary science education in rural communities, is more effective when delivered in the learner's mother tongue.

Ethe et al. (2014). Recent studies have begun to assess the capabilities of LLMs capabilities to handle multilingual data, especially in low-resource languages. For example,Lai et al. (2023)conducted a comprehensive evaluation of ChatGPT across various NLP tasks and languages and focused on the LLM's performance with less-represented languages.Bang et al. (2023) examined ChatGPT's multilingual abilities in tasks that included language identification, sentiment analysis, and machine translation, although their scope was limited to a few languages and about 50 samples per language.Wang et al. (2023) explored ChatGPT's abilities in English, Chinese, and German, but their study was primarily concerned with grammatical error correction and cross-lingual summarization, and focused mainly on well-resourced languages.

Ojo & Ogueji (2023) attempted to bridge the gap in understanding LLMs' performance with African languages. They evaluated two commercial large language models' performance using text classification and machine translation in African languages. Their findings indicate that these models are less effective with African languages, especially in machine translation. However, there has been limited exploration of how LLMs understand and process primary educational content in languages like Yoruba, despite the LLMs' potential role in transforming education. Our study seeks to fill this gap by conducting a detailed evaluation of LLM applications in Yoruba, particularly for primary science education in rural areas. We will assess the LLM's conceptual understanding using four NLP tasks on a manually generated dataset of Yoruba primary science educational content. Our work will examine the tasks of Named Entity Recognition (NER), Parts of Speech (POS), Question and Answering, and Paraphrasing, which have not been covered by previous multilingual evaluations of LLMs. To improve the reproducibility of the evaluations and better reflect the approach of general users, our current work will focus on the zero-shot learning setting for these LLMs, meaning that no human-provided examples will be presented to the model.

## 2 Related Work

### 2.1 Multilingual Analysis

Ojo & Ogueji (2023) conducted a comprehensive study examining the performance of commercial large language models in machine translation and text classification tasks across eight African languages, encompassing a variety of language families and geographical regions. Their findings highlighted the underperformance of these models in handling African languages. Interestingly, they observed a relatively better performance in text classification compared to machine translation. This study underscores the urgent need for better representation of African languages in commercial language models, especially considering their increasing global usage.

Wu & Dredze (2020) investigated the effectiveness of mBERT (multilingual BERT) across a broader range of languages, with a particular focus on low-resource languages. Their study, which encompasses Named Entity Recognition (99 languages), Part of Speech Tagging, and Dependency Parsing (54 languages each), reveals that while mBERT performs on par with, or better than, baselines in high-resource languages, its performance significantly declines for low-resource languages. This study notes that monolingual BERT models for these languages underperformed compared to mBERT. However, when paired with linguistically similar languages, the performance disparities lessened. This research highlights the necessity for more efficient pretraining techniques and augmented data for enhancing models in low-resource languages.

Wu & Dredze (2019),further explored the cross-lingual capabilities of mBERT, evaluating it as a zero-shot language transfer model across five NLP tasks and 39 languages from various language families. These tasks included natural language inference (NLI), document classification, named entity recognition (NER), part-of-speech (POS) tagging, and dependency parsing. Comparing mBERT with the leading methods for zero-shot cross-lingual transfer, they found mBERT to be competitive across all tasks.

However,Wu et al. (2022) cautioned that inherent linguistic differences could lead to syntactic discrepancies in the predictions of multilingual pre-trained models. In 2023,Lai et al. (2023) Viet et al. conducted a comprehensive evaluation of ChatGPT's capabilities across multiple languages and tasks, using large datasets to go beyond anecdotal evidence. This study aimed to determine whether ChatGPT and similar LLMs are effective in diverse languages or if there is a need for more language-specific technologies. This research fills a critical gap in the evaluation of ChatGPT and similar models by providing deeper insights into multilingual NLP applications. While the studies mentioned above have significantly contributed to understanding the general performance of LLMs in low-resource languages, there remains a notable gap in research focused on these languages within specific domains. For example, the effectiveness and conceptual understanding of these models' performance in domains like primary science education is understudied. Such an investigation is crucial, as it would greatly benefit the creation of personalized learning content in local languages and mother tongues in low-resource languages. It would also foster inclusivity in learning for students in underserved communities; such students perform better when taught science in their mother tongues and local languages Ethe et al. (2014).

## 3 Methodology

Our research aims to assess the contextual understanding of large language models on Yoruba primary science texts, focusing on four NLP tasks. Following the approach of Lai et al. (2023), we will evaluate the LLMs in the areas of POS tagging, NER, question and answering (QA), and paraphrasing. The evaluation of model outputs was conducted manually and subsequently reviewed by domain experts in primary science education with proficient Yoruba language skills. This rigorous process ensures the accuracy of our comparative analysis.

#### 3.0.1 The Dataset

The dataset encompasses meticulously selected primary science sentences, along with questions and answers extracted from Nigerian science textbooks. These datasets underwent a rigorous translation process conducted by human translators fluent in the Yoruba language, subsequently being digitized to enhance accessibility. This comprehensive collection includes questions and their corresponding answers, key primary scientific terms paired with their definitions, and a variety of scientifically relevant sentences. In total, the dataset includes 4463 science sentences, 1720 key scientific terms, and 1720 question-answer pairs. Furthermore, the translated datasets underwent meticulous review by learning experts to ensure accuracy and contextual coherence, thereby enhancing their educational value.

In line with Lai et al. (2023) zero-shot learning methodology, an NLP task (T) is defined through a natural-language description (D). For a new data sample with input text (X), the task (T) involves concatenating D and X and inputting this into the LLM, (G), to elicit a natural-language response (R = G ([D; X]). This response is then processed using predefined task-specific rules (P) to produce an output (Y = P (R (G ([D; X]) in the desired format for T, such as a specific classification label. We will store the output (Y) from our evaluation dataset to gauge the LLM's performance on task T. We employ single-stage prompting, integrating only the task description D with each input X, ensuring clarity and simplicity for zero-shot application.

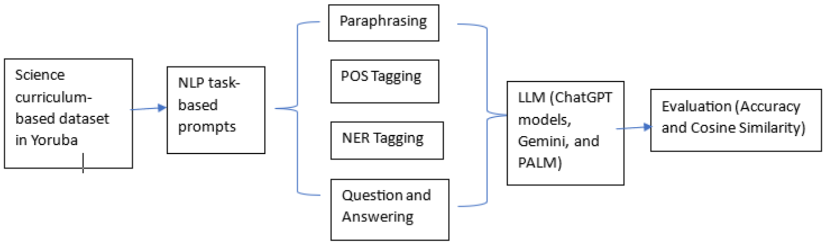
Science curriculum-based dataset in Yoruba
NLP task-based prompts
Paraphrasing
POS Tagging
NER Tagging
Question and Answering
LLM (ChatGPT models, Gemini, and PALM)
Evaluation (Accuracy and Cosine Similarity)

Figure 1: The Flow Process of the LLMs Contextual Evaluation

### 3.1 POS TAGGING

This task involves labeling the syntactic roles of words within sentences. Our approach for POS tagging with ChatGPT, similar to Lai et al. (2023), includes a prompt with a task description, an output format note, and an input sentence, in that sequence (PromptPOS = [task description; output format note; input sentence]). The tags include [”ADJ”, ”PROPN”, ”ADV”, ”NOUN”, ”NUM”, ”PRON”, ”VERB”]. We calculated and recorded the accuracy of the results in Table 1, using the formula:

The accuracy is determined using Equation 1, defined as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

### 3.2 TASK DESCRIPTION

Task Description: Please analyze the following Yoruba science sentence and provide the Part-of-Speech (POS) tags for each word. The output format should be a list of tuples, where each tuple consists of a word from the sentence and its corresponding POS tag label from the tag label set: [”ADJ”, ”PROPN”, ”ADV”, ”NOUN”, ”NUM”, ”PRON”, ”VERB”]

Note: Your response should include only a list of tuples, in the order that the words appear in the sentence, with each tuple containing the corresponding POS tag label for that word.

$InputSentence : yoruba_sentence$

Example Response Format:

Input Sentence:”Odìwon gígùn tabili ti Ade won je mita meta”
Expected Output: [(”Odìwon”, ”PROPN”), (”gígùn”, ”NOUN”), (”tabili”, ”NOUN”), (”ti”, ”PRON”), (”Ade”, ”PROPN”), (”won”, ”VERB”), (”je”, ”VERB”), (”mita”, ”NOUN”), (”meta”, ”NUM”)].

#### 3.2.1 RESULTS:

Table 1: Comparative Analysis of Various LLMs in POS tagging for Yoruba science content

| Model | Score |
|---|---|
| Gemini | 0.36 |
| ChatGPT-3.5 | 0.49 |
| PaLM 2 | 0.40 |

This table provides a comparative overview of the GEMINI, PALM, and ChatGPT 3.5 models' performance of the task of POS tagging in the context of Yoruba science education with ChatGPT having the highest accuracy. The table highlights the comparative performances of the LLMs in categorizing Yoruba science words in their respective parts of speech. The comparative analysis and accuracy of the models' output were manually done and calculated also and reviewed by learning experts.

### 3.3 NER

For the NER task, we focused on these entity types: PER (person), LOC (location), ORG (organization), and DATE (date). The prompt structure for LLMs in NER mirrors that used in POS Tagging (PromptNER = [task description; output format note; input sentence]), with acuary employed for evaluation. The findings are presented in Table 2.

Task Description: Your role is to perform named entity recognition on a Yoruba primary science text. Identify and categorize each named entity in the text using the following labels: PER (person), LOC (location), ORG (organization), and DATE (date). For multi-word entities, use the label "B" at the beginning and "I" for subsequent words within the entity. Label words not part of any named entity as "O".

Note: Please provide the output as a list of tuples. Each tuple should contain a word from the Yoruba text along with its corresponding named entity label.

Input Sentence: question

Example:

Input Sentence: ”Ade mu kalorimita dani sugbon se odiwon gígùn tabili naa ni Ile-Ife lana.”
Output: ”Ade mu kalorimita dani sugbon se odìwn gígùn tabili naa ni Ile-Ife lana.” will be [(”Ade”, ”PER”), (”mu”, ”O”), (”kalorimita”, ”O”), (”dani”, ”O”), (”sugbon”, ”O”), (”se”, ”O”), (”odìwn”, ”O”), (”gígùn”, ”O”), (”tabili”, ”O”), (”naa”, ”O”), (”ni”, ”O”), (”Ile-Ife”, ”LOC”), (”lana”, ”DATE”)]

#### 3.3.1 Results:

Table 2: Performance of Various LLMs in NER Tasks

| Model | Score |
|---|---|
| Gemini | 0.30 |
| ChatGPT-3.5 | 0.05 |
| PaLM 2 | 0.17 |

**Discussion** This table compares the effectiveness of the ChatGPT 3.5, Gemini, and PALM 2 models in identifying named entities within the Yoruba science education content. Gemini shows a moderate level of proficiency whereas ChatGPT-3.5 and Palm 2 exhibit lower performances, thus highlighting differences in the models' capabilities of understanding and categorizing key entities in Yoruba.

### 3.4 Question and Answering

In this segment of our study, we assessed the performance of LLMs using a set of science questions in Yoruba. Our objectives were to evaluate the accuracy and contextual relevance of the LLMs' responses in the Yoruba language. To quantify this, we utilized a method involving the calculation of accuracy scores between each model's contextual responses compared with the original answers provided in our datasets using Equation 2 defined as follows.

Equation 2

$$\text{Accuracy} = \frac{\text{Number of Correctly Generated Texts}}{\text{Total Number of Generated Texts}} \tag{2}$$

This approach allowed us to measure how closely the LLM-generated answers aligned with the expected responses, based on vector representations in a multidimensional space. A higher accuracy score indicates a greater degree of alignment, thus reflecting higher accuracy and contextual appropriateness of the LLM's answers to the original Yoruba science questions. This evaluation provides valuable insights into the LLM's capabilities of understanding and responding accurately in a specific non-English language context, particularly in the domain of science education.

Task Description: Answer the Yoruba science question in English based on your understanding or knowledge. The answer should be concise and relevant to the question.

Example:

Question: Kíni ìdí tí ohun abemí fi nilo oúnje àti omi?

Answer: Ohun abemí nilo oúnje fún okun, àti omi láti ní omi lára.

Note: Your answer should be in Yoruba and directly address the question. Question: question

### 3.4.1 Results:

Table 3: Comparative Analysis of Various LLMs in Question-Answering (QA) for Yoruba Science Content

| Model | Score |
|---|---|
| Gemini | 0.16 |
| ChatGPT-3.5 | 0.002 |
| PaLM 2 | 0.04 |

**Discussion** In the QA task, after the evaluation of the models' outputs by experts in the field, GEMINI scored 0.1575, while PaLM 2 scored 0.04 with ChatGPT-3.5 having the lowest score. Gemini's superior performance indicates a better understanding of Yoruba science education content in the context of answering questions. This could be due to the model's ability to comprehend complex queries and provide relevant answers, which is crucial in educational settings

## 3.5 Paraphrasing

We also evaluated LLMs' abilities to paraphrase content in the context of primary science education in Yoruba. Paraphrasing, in this sense, refers to the model's capability to restate the meaning of a text using different words and structures while preserving the original message and factual accuracy. some of the output of the models are shown in table 5 below.

The process involved providing the LLMs with sentences or passages from the Yoruba primary science dataset and then prompting them to rephrase the content.

Task Description: Your task is to perform paraphrase generation for Yoruba primary science sentences. Given an input science sentence in Yoruba, generate a semantically equivalent paraphrase. The paraphrase should convey the same scientific concept as the input sentence but may be expressed using different words or structures.

Note: Please provide the output as a paraphrased science sentence in Yoruba.

Input Sentence: Yoruba primary science sentence

Example: Input Sentence: ”Òdìwon gígùn, ìbú àti gíga.”
Output: ”Gígùn, ìbú àti gíga je odìwon”

Table 4: Paraphrased Sentences from the Different Models

| | Paraphrased | | |
|---|---|---|---|
| Original Sentences | ChatGPT-3.5 | Gemini | PALM |
| Ayika wa ni agbegbe wa ati gbogbo ibi to sun mo wa | Agbegbe wa ni o da gbogbo ibi to sun mo wa ni igbo, bi o se se pelu awon eranko miiran ni omi. | Eyin wa laarin awon ayika meji: Eyin to daju ati eyin to koko woye. | Ayika wa lo sun mo wa to si yi waka |
| Orisirisi ayika meji lo wa: Adayeba ati Atowoda | Nigbati awon eranko miran ba n gbe ni omi, awon eranko kan n gbe ni igbo ni agbegbe. | A kole lo lati ibikan si ibomiran Lai lo ona | Ayika pe orisirisi meji to wa: Adayeba ati Atowoda |

We evaluated the results using Equation 2 above.

Table 5: Efficacy of Various LLMs in Paraphrasing Yoruba Science Education Content

| Model | Score |
|---|---|
| Gemini | 0.07 |
| ChatGPT-3.5 | 0.40 |
| PaLM 2 | 0.44 |

**Discussion** This table 5 below contrasts the performances of ChatGPT-3.5, Gemini, and PALM 2 in paraphrasing tasks specific to Yoruba Science Education. The data reveal PALM's superior performance over ChatGPT-3.5 and Gemini, suggesting its better aptitude in rephrasing content while maintaining the original context and meaning, a crucial aspect in educational material interpretation. The evaluation of the models' outputs was conducted manually and reviewed by learning experts in the field using equation 2 defined previously, ensuring accuracy in the comparative analysis.

## 4 Discussion

The focus of our research on the application of Large Language Models (LLMs) such as ChatGPT-3.5 and Gemini in processing Yoruba primary science text holds significant implications for science education, particularly for students in rural areas. Our evaluation, which spanned the tasks of POS Tagging, NER, QA, and paraphrasing, offered crucial insights into these LLM's adaptability to the specific educational and linguistic contexts of Yoruba.

POS Tagging and NER: The results of the POS Tagging and NER tests demonstrated notable but inconsistent performance by the LLMs when processing in the Yoruba language. This inconsistency is a matter of concern in educational contexts, where language accuracy and clarity are indispensable for effective learning. The models' struggles with accurately identifying and classifying linguistic elements in Yoruba indicate a gap in their understanding of this language's unique structures, which could lead to misunderstandings in an educational setting.

Question answering: The QA results highlighted a significant deficiency in the models' contextual comprehension, particularly when handling science questions posed in Yoruba. This shortfall is especially problematic in rural educationl settings, where the nuanced understanding of language plays a pivotal role in how students grasp and engage with scientific concepts. The LLMs' inability to accurately interpret and respond to questions in Yoruba could impede their application as educational aids in these regions.

Paraphrasing: While the LLMs showed a moderate ability to rephrase scientific content in Yoruba, maintaining the factual accuracy and educational integrity in their paraphrased content was a challenging task. This aspect is crucial in rural education environments; resources are often limited, and the reliability of the information presented is critical. The observed limitations in paraphrasing indicate the need for further refinement of these models to ensure they provide accurate and contextually relevant educational content.

## 5 Conclusion

This study highlights the need for more inclusive and culturally sensitive approaches in the development of LLMs, especially for educational applications in languages like Yoruba. While the potential of LLMs to enhance educational content and methodology is significant, their effectiveness in non-English, rural educational settings is yet to be fully realized.

To better serve rural educational needs, it's essential to develop LLMs with a deeper understanding of local languages, cultures, and educational contexts. This involves diversifying training datasets, improving contextual understanding, and ensuring cultural appropriateness. Future research should focus on the development of LLMs tailored to the needs of rural education, and exploring ways to make this technology more accessible and relevant to these communities.

In summary, our research contributes to understanding the role and capabilities of LLMs in rural education settings, emphasizing the need for targeted advancements to make these tools truly effective and inclusive for all learners, particularly those in underrepresented and linguistically diverse regions.

## REFERENCES

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Nathaniel Ethe, E Avbenagha Andrew, and O Akpojisheri Monday. The effect of using mother tongue in teaching and learning basic science in delta state, nigeria. In *Proceedings of INTCESS14—International Conference on Education and Social Sciences Proceedings*, pp. 1640–1645, 2014.

Yasir Gamieldien, Jennifer M Case, and Andrew Katz. Advancing qualitative analysis: An exploration of the potential of generative ai and nlp in thematic coding. *Available at SSRN 4487768*, 2023.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. Large language models in education: Vision and opportunities. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 4776–4785. IEEE, 2023.

Nursen İlçin, Murat Tomruk, Sevgi Sevi Yeşilyaprak, Didem Karadibak, and Sema Savcı. The relationship between learning styles and academic performance in turkish physiotherapy students. *BMC medical education*, 18(1):1–8, 2018.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*, 2023.

Andreea-Mihaela Nita, Gabriela Motoi, and Cristina Ilie Goga. School dropout determinants in rural communities: The effect of poverty and family characteristics. *Revista de Cercetare si Interventie Sociala*, 74, 2021.

Jessica Ojo and Kelechi Ogueji. How good are commercial large language models on african languages? *arXiv preprint arXiv:2305.06530*, 2023.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. Zero-shot cross-lingual summarization via large language models, 2023.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632*, 2023.

Linjuan Wu, Shaojuan Wu, Xiaowang Zhang, Deyi Xiong, Shizhan Chen, Zhiqiang Zhuang, and Zhiyong Feng. Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. 2022. doi: 10.48550/ARXIV.2204.00996. URL `https://arxiv.org/abs/2204.00996`.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 833–844, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL `https://aclanthology.org/D19-1077`.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*, 2020.