applied sciences

MDPI

Article

How Good Are Large Language Models at Arithmetic Reasoning in Low-Resource Language Settings?—A Study on Yorùbá Numerical Probes with Minimal Contamination

Fiyinfoluwa Oyesanmi * and Peter O. Olukanmi *

School of Electrical and Electronics Engineering, University of Johannesburg, Johannesburg 2092 , South Africa
* Correspondence: oyesanmifiyinfoluwa@gmail.com (F.O.); polukanmi@uj.ac.za (P.O.O.)

Abstract: We study the performance of large language models (LLMs) in natural language understanding and natural language reasoning tasks in a low-resourced-language (LRL) setting. Using Yorùbá, an LRL, we curated a set of numerical probes with minimal contamination. The probes comprise three sets of questions—the first covers basic arithmetic, the second covers date and time (calendar system), and the last focuses on numerals and counting systems. Assessed in a zero-shot setup, three LLMs (ChatGPT, Gemini, and PaLM) were evaluated based on several metrics. The best-performing model, ChatGPT, generated some correct answers, showing logical steps in attaining the answers in Yorùbá (with an accuracy of 56% in set one, and 44% in set two). The second-best model (with an accuracy of 56% in set one, and 32% in set two) is Gemini. PaLM (with an accuracy of 16% in set one, and 8% in set two) showed the answers without logic. The three models performed poorly on the Yorùbá numerals question set (ChatGPT scored 8%, and Gemini and PaLM each had 0% accuracy). The study also revealed that there is significant room for improvement in the state of the art of LLMs when it comes to Yorùbá numerals.

Keywords: natural language understanding; Yorùbá numerical probes; natural language reasoning; low-resource languages; large language models

check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 11 March 2025
Revised: 10 April 2025
Accepted: 14 April 2025
Published: 17 April 2025

Citation: Oyesanmi, F.; Olukanmi, P.O. How Good Are Large Language Models at Arithmetic Reasoning in Low-Resource Language Settings?—A Study on Yorùbá Numerical Probes with Minimal Contamination. *Appl. Sci.* **2025**, *15*, 4459. https://doi.org/10.3390/app15084459

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Large language models (LLMs) have revolutionized natural language processing (NLP) and computing technologies with unprecedented potential for providing satisfactory answers to user inputs across different domains and in multiple languages. The ability of LLMs—such as the popular ChatGPT—to provide tangible answers to questions or even pass difficult examinations in various fields, including medicine [1,2] and mathematics [3,4], has amazed many and become a subject of investigation for others.

The increasing popularity and wide acceptance of LLMs are based on their seeming prowess and robustness built on their ability for self-awareness [5], self-learning [6,7], and processing tasks without prior proper fine-tuning for such specific tasks (zero-shot) [8,9]. This was, however, not always the case as these capabilities were not present in smaller models [10,11]. Refs. [12,13] submitted that LLMs find it hard to cope when dealing with tasks that require systematic and logical reasoning. Several models, such as the chain-of-thought (CoT) prompting method and zero-shot CoT [14], have been proposed to enhance such capabilities so that LLMs can properly handle arithmetic reasoning tasks [12]. Standing on the axiom that "numbers rule the universe", several research endeavors, including Refs. [15,16], have underscored the necessity of reasoning with numbers and training AI systems to do the same. In 2022, Ref. [17] submitted that optimal AI models are brittle at best

while dealing with arithmetic reasoning tasks when presented in different formats. While the ability of LLMs to handle tasks requiring systematic reasoning for mathematical challenges has improved through several approaches [15,18], most of the deductive reasoning tasks on which these models are benchmarked are in the English language.

In this work, we seek to contribute to the assessment of LLMs by evaluating their capabilities in providing zero-shot responses to arithmetic reasoning tasks in low-resourced languages (LRL), Yorùbá, in particular, using a dataset that has minimal contamination. Yorùbá is one of the most common languages in West Africa, with more than 40 million speakers globally. We chose arithmetic reasoning tasks for this assessment because they present an opportunity to analyze LLM's ability to understand and make systematic deductions from LRLs. We also present a set of Yorùbá numerical probes curated for this evaluation. The probes are made up of three categories of questions. The first set is made up of basic arithmetic comprising addition (ìsirò or àròpò), subtraction (àyokúrò), division (pípín), multiplication (ìsodipúpò), as well as a combination of these operations. The second set contains problems relating to time, day of the week, and month of the year, while the third set of questions deals with Yorùbá numerals and the counting system (ònkà).

To this end, we present the performance of three publicly available LLMs on our curated numerical probes and evaluate how the models perform in a zero-shot setup on the probes with minimal contamination. In zero-shot learning, a model is prompted to perform a task based on its internal knowledge without having been explicitly trained on examples related to that task [19]. While approaches such as few-shot learning or model fine-tuning can improve the performance of LLMs, we adopted a zero-shot approach to assess their inherent abilities and establish baseline capabilities without any augmentation or fine-tuning.

Our contributions in this study include (i) the introduction of Yorùbá numerical probes—the first, to the best of our knowledge—with minimal contamination, designed to evaluate the performance of LLMs on natural language understanding (NLU) and natural language reasoning (NLR) tasks for LRLs; (ii) an assessment of the NLR capabilities of three LLMs in a zero-shot setup using the same set of probes; (iii) identification of the strengths and weaknesses of these LLMs in NLU and NLR tasks for LRLs (Yorùbá in this case).

2. Review of Related Works

Understanding a language is fundamental before reasoning in such a language can occur. NLU is, therefore, aptly described in [20] as a machine's ability to grasp and make sense of a prompt in a natural language, either text or speech, and derive meaning from such input. NLR can be described as the ability of a machine to produce new statements, declarations, or actions based on existing knowledge without relying solely on its memorized facts, accumulated knowledge base, or explicit contextual input [21].

NLU in low-resource languages is a daunting challenge owing to the lack of sufficient data for training LLMs. There are several approaches to mitigate this challenge. The transfer of the capabilities of pre-trained language models from high-resource languages to low-resource languages, leveraging vocabulary matching, was proposed in [22]. This technique has been shown to be effective in improving the performance of BERT models, even with minimal training data in the target language. Another proposed approach involves retrieval-augmented language models. This model leverages an external voluminous collection of documents to aid the improved performance of knowledge-intensive tasks [23,24]. It has been demonstrated that the approach is effective in few-shot learning settings, with language models picking up knowledge-intensive tasks with very few training examples. The use of natural language descriptions of slots has also been proposed as a method to improve NLU models in LRLs [25].

The need for robust multilingual NLU benchmarks to evaluate the performance of NLU models on LRLs was addressed in [26] through the creation of SIB-200 for such an evaluation exercise. The benchmark, like many other multilingual benchmarks, is designed for the evaluation of the performance of NLU models in many languages, including LRLs, to facilitate a standardized comparison of the performance of different models and pinpoint areas for improvement. Ref. [27], however, submitted that a fundamental limitation of multilingual LLMs is their poor ability to generalize when it comes to NLU tasks in LRLs. LRLs, when included in multilingual benchmarks, are highly limited compared to other languages. Ref. [27] proposed the LM contamination index, which seeks to assess the impact of data contamination on the performance of LLMs as an alternative evaluation metric.

Overall, natural language understanding in low-resource languages is a challenging task that requires the development of new approaches and models that can leverage limited resources and data [23].

3. Method

3.1. The Yorùbá Numerical Probes

The Yorùbá numerical probes used in this study are made up of simple arithmetic problems in the Yorùbá language with no translation in English or options for the LLMs to choose from. Carefully curated to ensure orthographic normalization, we ensured that all the questions in the prompt had appropriate tonal marks. We also used only the standard Yorùbá writing version (àkoto Yorùbá) in all the questions to ensure consistency. In addition, we ensured that the probes maintained the same structure in the three models to facilitate consistency.

There are three categories of questions in the probes. The first set is made up of 50 basic arithmetic problems, comprising addition (ìsirò or àròpò), subtraction (àyokúrò), division (pípín), multiplication (ìsodipúpò), and a combination of these operations that requires a multi-layered approach to answering the questions. The second set contains 25 mathematical problems relating to time, day of the week, and month of the year in Yorùbá since the calendar system is an integral part of arithmetic. The third set also contains 25 questions that deal with counting in the Yorùbá numerical system (ònkà).

Table 1 provides an example of the first set of questions. The first 20 questions in the probe are in Appendix A.

Table 1. Types of questions and examples.

Irú Ìbéèrè (Question Type)	Ìbéèrè (Question)
àròpò	Dupe gba naira marun fun owó oúnje kúrò ní ilé. Ó pàdé Iya Gbonke tó fún un ní naira mesan àti Baba Alatise tó fún un ní naira mejila bí ó ṣe ń lo sí ilé-iwe. Èló ni owó tí Dupe ní je lápapo?
àyokúrò	Baba Sade fun Sade ni eko mewa. Sade fun Faderera ni eko meji. Mélò ni eko owo Sade kù?
ìsodipúpò	Baba Jiginni gba Abeke niyanju lati je ataare kookan lojoojumo fun ose meji gbako. Ataare melo ni o ye ki Abeke ra?
pípín	Oluko ri wipe Aarinola ko fi eti sile ninu yaara ikeko, o si so fun wipe ki o lo ka iye eranko ti o wa ninu ogba ile ikẹko naa. Ti ẹse eranko ti o wa ninu ogba naa ba je merindinlogota, ki ni idahun Aarinola yoo je?

The first question in the table translates to the following:

"Dupe was given 5 naira for lunch before heading to school. She met Iya Gbonke and Baba Alatise, who gave her 9 naira and 12 naira, respectively. How much does Dupe have altogether?"

This typical question is adjudged to be simple enough for LLMs to understand and draw an inference or give the right answer before training them on such tasks.

In Table 2, we briefly show the types of questions in sets two and three. The full questions in the probes for sets two and three are in Appendix A.

Table 2. Example of questions in sets two and three.

Irú Ìbéèrè (Question Type)	Ìbéèrè (Question)
set 2	Tì òní báa jẹ ojo ajé, ojo wo ni oojo mewa òní yóò je?
set 2	Ojo mélòó ló wà nínú oṣù Èrèlé
set 3	Ónkà wo lo tele Ogúnlélúgba?
set 3	Kinni ogojo ni onka Yorùbá

The first question in Table 2 translates to *"If today is Monday, what day will it be ten days from today?"* The third question (set 3) translates to *"what is the next number after 220?"*.

Steps Taken to Avoid Dataset Contamination

Dataset contamination—a paradigm used to check whether evaluation data have been previously exposed to language models during training—aims to uphold the integrity and trustworthiness of LMs. We, therefore, ensured that the Yorùbá numerical dataset we curated had minimal contamination, taking the following steps. First, all the questions in the dataset were manually constructed. We did not use the translation of common mathematics questions in English to Yorùbá since this has been reported (in [28]) to contribute to contamination. We also chose to make the probes unique by not providing answer options (as is common in multiple-choice questions), so that the answers generated by the models would not come as mere guesses. Finally, we did not adopt any examination questions for this task, as is common in some datasets.

3.2. Models Used

In this study, we used three popular and publicly available models, namely, ChatGPT, Gemini, and PaLM. These models were chosen because they are currently regarded as state-of-the-art multipurpose LLMs. They have also been trained on diverse datasets and generalize well to tasks—such as question answering, named entity recognition, and text summarization—even in a low-resource language like Yorùbá. We considered other models, such as LLaMA-2-7B-Chat, but did not obtain results for our Yorùbá prompts and, therefore, decided to stick with those three.

- ChatGPT is arguably the most popular LLM. Having been trained on a wide range of multilingual datasets that include Yorùbá, its strong multilingual capabilities, accessibility, reasoning proficiency, and adaptability to generalize in LRL tasks like text generation and translation [29] make it an ideal choice for this assessment. The model we used was ChatGPT-4-turbo, which was accessed between January and April 2025.
- Gemini was developed by Google DeepMind; it is the second model we chose to assess NLR on Yorùbá arithmetic probes. Gemini has advanced multimodal capabilities, a fine NLR architecture, and robust support for LRL [30], including Yorùbá. The model we used was Gemini 2.0 Flash, which was accessed between January and April 2025.
- PaLM (Pathways Language Model) [31] has an architecture that supports complex reasoning and understanding of context. Its ability to solve multistep problems and support Yorùbá informed our choice of this model for the assessment. The model used was PaLM 2, accessed between January and April 2025.

3.3. Evaluation Metrics

The performance of LLM in arithmetic reasoning or "common sense" tasks for LLR was evaluated in [32] with a combination of quantitative and qualitative metrics. In the quantitative approach, the accuracy of task-specific multiple-choice questions was assessed, along with exact matches for subjective or "fill-in-gap" questions. Ref. [32] also proposed the use of human evaluators to assess the accuracy of responses generated by LLMs, as they can provide in-depth analyses—evaluating the models' understanding of language semantics, ability to acquire knowledge, reason accordingly, and generalize properly. This forms the qualitative evaluation approach. For LRLs, Ref. [33] submitted that it is important to consider the limitations of LLMs in terms of their ability to generalize to new languages and domains. Multilingual language models like GPT-3 and PaLM have demonstrated quality performance in attending to arithmetic reasoning tasks in English, but their performance may degrade when applied to LRLs. It may, therefore, be necessary to fine-tune the language model on a specific language or domain to address such challenges.

The metrics we adopted for this model are accuracy, reasoning completeness, and language understanding.

3.3.1. Accuracy

Accuracy compares the proportion of correct answers to the total number of questions, thus evaluating how the generated answer aligns with the expected or model answer [34]. It provides a general overview of the overall performance of a model. Accuracy is given as follows:

$$\frac{\text{Number of Correct Answers}}{\text{Total Number of Tasks}} \times 100$$

3.3.2. Reasoning Completeness

Ref. [35] submitted that reasoning completeness is an assessment of how far a generated answer covers the inherent details of a question. That is, it is an evaluation of the proportion of tasks in which the various models show all the reasoning steps taken in obtaining the final solution. It is given as follows:

$$\frac{\text{Number of Fully Completed Reasoning Tasks}}{\text{Total Number of Tasks}} \times 100$$

3.3.3. Language Understanding

This metric measures the linguistic and cognitive faculties of a model [36] (in this context, a measure of the proportion of tasks in which the model correctly interprets the problem statements expressed in Yorùbá).

$$\frac{\text{Number of Correct Interpretations}}{\text{Total Number of Tasks}} \times 100$$

4. Results

The performance of the three models was evaluated based on their zero-shot answers. The scoring (model answers) to the questions, against which the model's answers were compared, was done by humans who were both native speakers of the Yorùbá language and the curators of the numeric probes. Being domain experts, the model answers provided were the basis for the evaluation of the performance of the three LLMs.

Figure 1 shows how each of the three models answered a given question. ChatGPT and Gemini presented the logic (reasoning completeness) followed to arrive at the final answer, while PaLM did not show the logic.

Ìbéèrè 1	**Aduke máa n lọ sí ibi isé ní ojoojúmó. Owó ọ̀yà re léyìn osù jé 60000 naira. Tí Aduke kò ba lọ sí ibi isé fún ọjọ́ marun ní osù yìí, èló ni owó osù re yóò jé l'ósù yìí?**		
Translation	Aduke goes to work every day and earns 60,000 Naira as her monthly salary. If Aduke did not go to work for 5 days this month, how much will her salary be?		
Model	ChatGPT	Gemini	PaLM
Logic	Ojọ osù = 30. Owo ojọ kan: 60,000÷30=2,000 Ejọ marun: 2,000×5=10,000 60,000−10,000=50,000	• Owó ọ̀yà Aduke léyìn osù = 60,000 náírà • Iye ọjọ́ tí Aduke kò lọ sí ibi isé = 5 ọjọ́ • A gbọ́dọ̀ kọ́kọ́ mọ owó tí ó ń gbà fún ọjọ́ kan: 60,000 náírà / (A gbà pé iye ọjọ́ isẹ́ nínú osù jẹ́ 30) = 2,000 náírà/ọjọ́ • Owó tí yóò pàdánù nítorí àìsí = 5 ọjọ́ * 2,000 náírà/ọjọ́ = 10,000 náírà • Owó osù tí yóò gbà = Owó ọ̀yà léyìn osù - Owó tí ó pàdánù = 60,000 - 10,000 = 50,000 náírà	
Ìdáhùn (Answer)	Naira 50,000	Owó osù Aduke yóò jẹ́ 50,000 náírà l'ósù yìí.	Owó ọ̀sù Aduke yoo jẹ́ 54,000 naira l'ọ̀sù yìí.

Figure 1. How the models answered a sample question showing the logic.

From the solution presented by the models shown in Figure 1, the answer provided by both ChatGPT and Gemini aligns correctly with the model/expected answer.

4.1. Question Set One: Basic Arithmetic Problems

Figure 2 shows the heatmap that reflects the distribution of scores across the fifty basic arithmetic questions; the green color represents correct answers while the red color shows incorrect answers. The heatmap captures how well the models perform in the questions that relate to simple addition, subtraction, multiplication, and division. ChatGPT and Gemini scored 28 questions apiece (56% accuracy), with PaLM scoring 8 questions (16% accuracy).

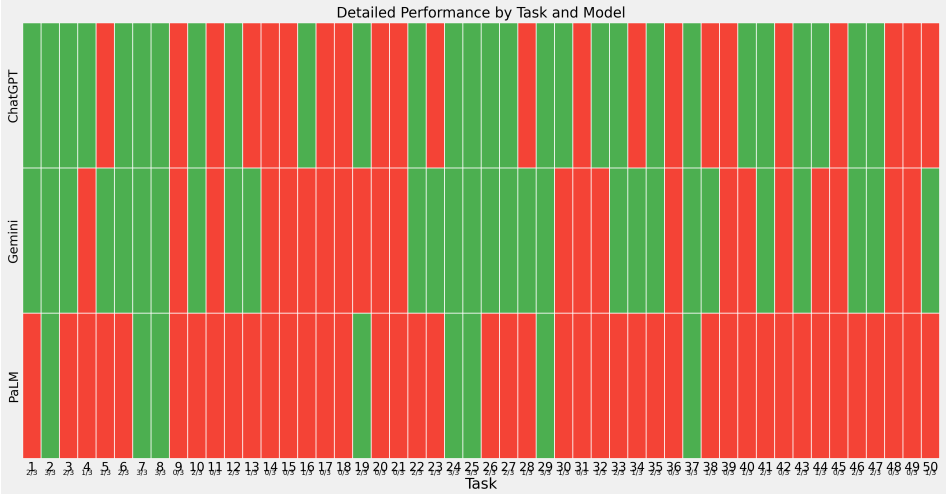


Figure 2. Performance heatmap of the three models on the first question set. The red colors represent the incorrect answers while the green colors represent the correct answers.

Figure 3 shows the distribution of how the responses from the three models align with the model answers for the first question set.

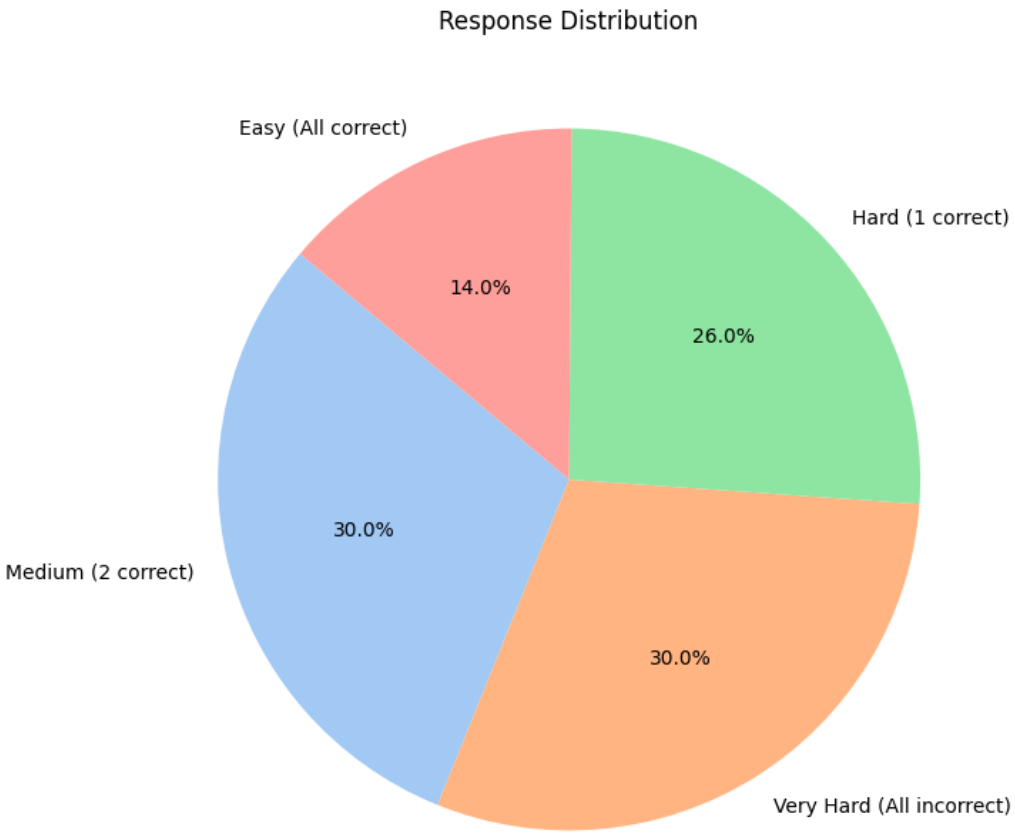


Figure 3. The distribution of how the answers of the three models map to the model answer on the first question set.

From the response distribution, there were seven questions that all the models answered correctly, sixteen questions that two of the models answered correctly, twelve questions that only one model answered correctly, and fifteen questions that none of the models answered correctly. ChatGPT and Gemini showed different but comparable capabilities, while PaLM seemed to have the most difficulty in answering the questions. For instance, there were seven questions that only Gemini answered right, five questions that only ChatGPT answered right, and one question that only PaLM answered right. We tagged the fifteen questions (30% of all questions) that none of the models could answer correctly as "very hard". These questions all had seemingly complex Yorùbá numerals, showing that the models struggled when dealing with questions involving "complex" Yorùbá numerical values, most of which were wrongly represented.

We equally assessed the performance of the models based on their reasoning completeness, language understanding, and numerical representation accuracy. We checked whether there was a logical breakdown in the step-by-step computations provided by the models, assigning a score of 1 or 0 for each complete or incomplete reasoning step to assess the reasoning completeness metric. We assessed the language understanding metric in the same manner, checking if the model "gets the gist" of the question or exhibited misinterpretation or misrepresentation of the question's context. The numerical representation accuracy checked if the figure used by the model in its calculation was the right equivalence of the Yorùbá numeral given in the question. Instead of awarding 1 or 0 for right or wrong in this case, we considered the three possibilities that might have happened and awarded corresponding scores to each. Correct representation of all numbers = 2, one or two misrepresentations = 1, and completely misrepresented the figures = 0. Since PaLM did not give the logic used to arrive at its answers, we only assessed ChatGPT and Gemini.

Table 3 provides a clear perspective of the underlying factors that influenced the accuracy of both models in terms of the number of questions answered correctly. While both have the same number of correctly answered questions (56%), there is a great difference in their reasoning completeness, language understanding, and numerical representation accuracy. Gemini outperformed ChatGPT in these three metrics, accounting for the 56% disagreement between the answers generated by both models. The numerical representation accuracy level of both models is, however, low, affecting their final computation.

Table 3. Performance of the models based on other criteria.

	Other Assessment Metrics		
Model	**Reasoning Completeness (%)**	**Language Understanding (%)**	**Numerical Representation Accuracy (%)**
ChatGPT	86	74	61
Gemini	98	92	66

4.2. Question Set Two: Time and Dates

This question set, which tests understanding of the time and calendar systems in the Yorùbá language, contains 25 probes. Figure 4 shows the distribution of the performance of the three models on the 25 questions. ChatGPT recorded 11 correct scores (44% accuracy), Gemini recorded 8 correct scores (32% accuracy), and PaLM answered 2 out of 25 tasks (8% accuracy).

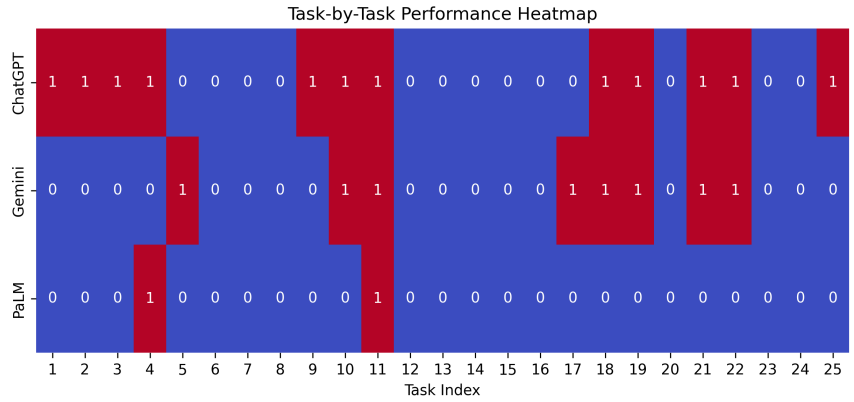


Figure 4. Performance heatmap of the three models on the second question set. The red color represents the correct answers, while the blue color represents the wrong answers.

The NLU ability of the three models was the focus of the second set of questions. It was obvious through the performance heatmap in Figure 4 that ChatGPT was able to understand the system through which the time, day of the week, and month of the year in Yorùbá operated, and was able to generalize more effectively on related questions compared to the other models.

Figure 5 shows that all the models answered 12 of the 25 (44%) questions wrong, and all answered only 1 question (4%) right. While Gemini's performance was closer to that of ChatGPT than to PaLM, it is clear that these LLMs do not have sufficient understanding of the time and calendar system in the Yorùbá language.

We did not assess the performance of the models based on their reasoning completeness, language understanding, and numerical representation accuracy for this set of questions. This is because most of the questions were simple and straightforward, and did not require a logical, step-by-step approach to derive answers. For instance, one of the questions in the set, "Ojo mélòó ló wà nínú oṣù Agemo" means "how many days are in July?".

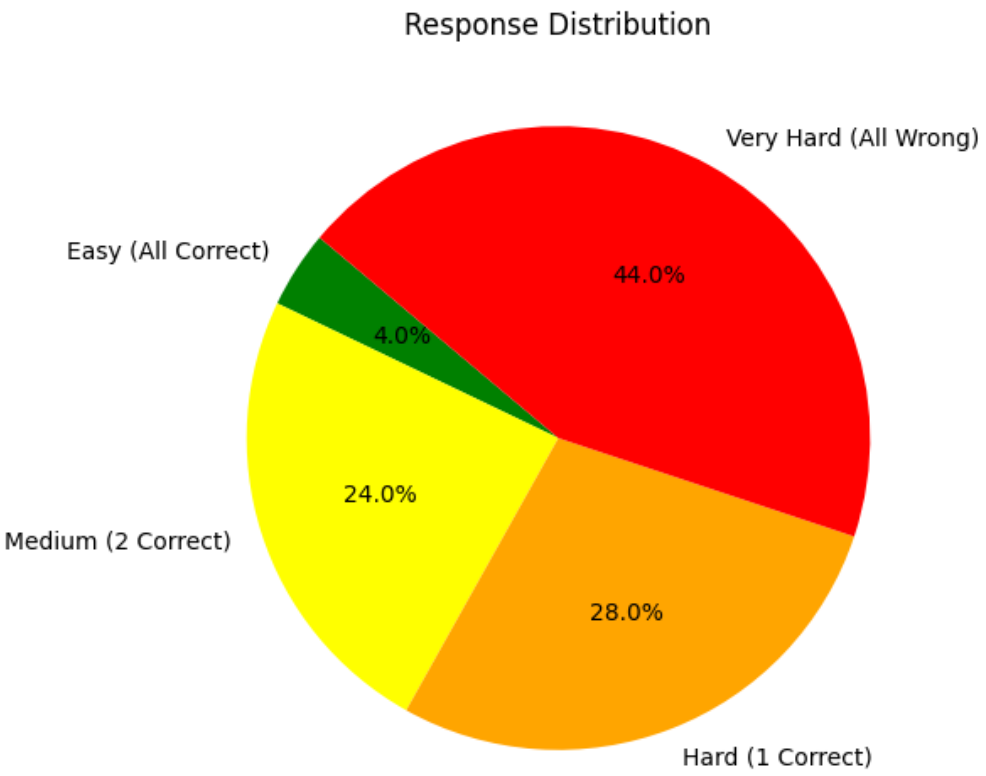


Figure 5. Response distribution of the three models on the second question set.

4.3. Question Set Three: Yorùbá Counting and Numeral System

The last set of the numerical probes contained twenty-five questions, seeking to assess how well LLMs understand the Yorùbá counting and numeral system. ChatGPT was the only model that answered any of the questions correctly, scoring 2 out of 25 (8% accuracy), with the others scoring 0.

In Figure 6, the red color represents correct answers while the blue color denotes wrong answers. The heatmap clearly shows that the three models struggled greatly when dealing with Yorùbá numerals. The level of understanding of the Yorùbá counting and numeral systems is still some distance from a desirable level.

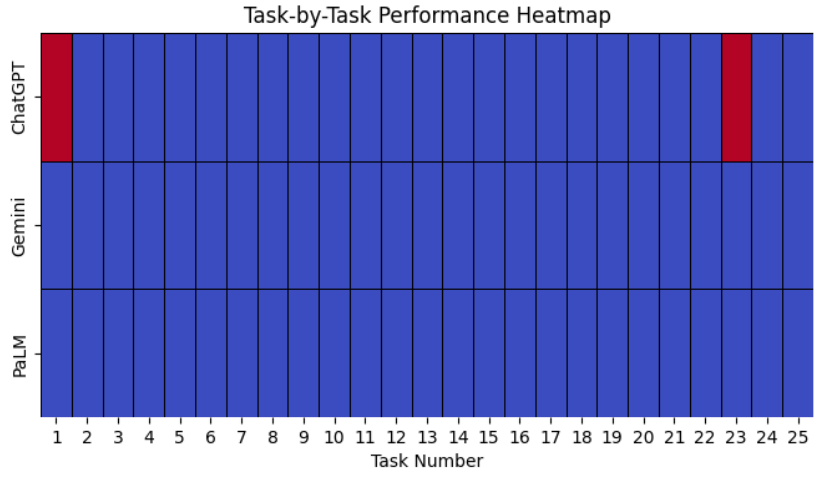


Figure 6. Performance heatmap of the three models on the onka Yorùbá question set. The red represents correct answers, while the blue represents the wrong answers.

4.4. Overall Performance of the Three Models

Table 4 presents a summary of the performance of the models based on their accuracy scores for the three tasks.

Table 4. Overall performance of the models based on their accuracy.

	Accuracy		
Model	**Basic Yorùbá Arithmetic Set (%)**	**Yorùbá Time, Days and Months (%)**	**Yorùbá Numerals (%)**
ChatGPT	56	44	8
Gemini	56	32	0
PaLM	16	8	0

The performance of the three models across the three question sets shows that none of the models achieved up to 50% accuracy. As seen in Figure 7, the best performance stood at 35.3% overall.

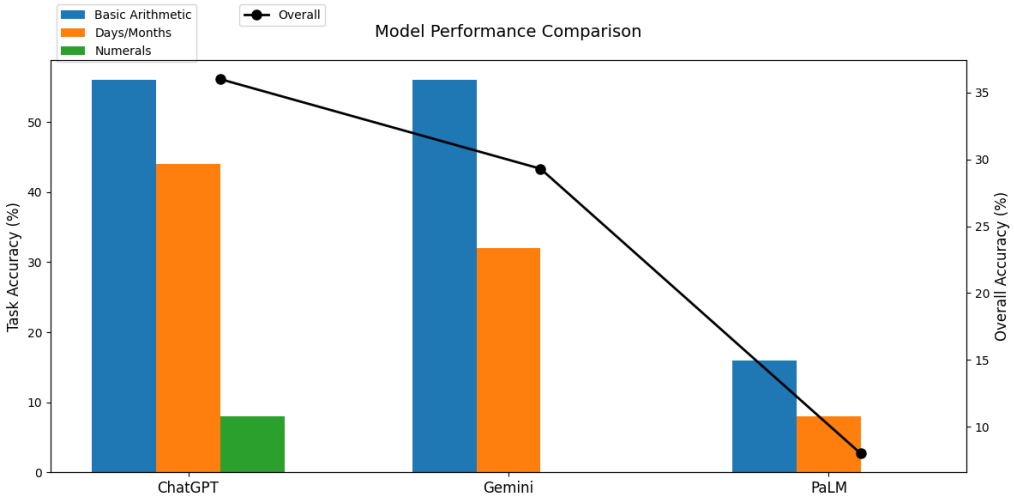


Figure 7. Overview of the overall performance of the three models across all question sets.

We equally evaluated the performance of the models based on their reasoning completeness, language understanding, and numerical representation accuracy, especially for the first question sets, where these metrics could be easily assessed. These metrics point to two types of error—the first involves the misrepresentation of Yorùbá numbers as figures, and the second is the misrepresentation of the problem context. Since the former is the most prominent of the two types of errors, we will provide examples of how numerals in some questions are misrepresented in Section 4.4.

Analysis of Numerical Representation Errors

Here, we provide three examples that show the question asked and the translation. We then present the expected representation of the numerals for each question alongside the numeral representations used by the models in their calculations in the following table.

Question 21

Yorùbá: *Èwo ló po jù nínú otàlélúgba ati àrúnlélotàlélúgba?*
Translation: Which is greater between 260 and 265?

Question 31

Yorùbá: *Ramo elélùbo rà goro èlùbo ogbon ní egbèsàn naira. Tí ó bá tà goro èlùbo kookan ní àádota naira, èrè àbí àdánù èló ní Ramo elélùbo je lórí ojà yìí?*

Translation: Ramo the cassava-flour seller bought 30 drums of cassava flour for N1800. If she sells each cassava drum at N60, how much profit or loss did she make?

Question 49

Yorùbá: *Balogun, ìyàwó rè àti àwon omo won méta fe lo wo Saheed Osupa níbi tí ó ti n seré. Owó ìwolé fún àwon omodé je èedegbejo naira, ti àwon àgbàlagbà sì je egbàáta naira. Èló ni owó ìwolé lápapo?*

Translation: Balogun, his wife, and three kids want to watch the live performance of Saheed Osupa. The gate fee for kids is N1500, and that for adults is N6000. How much is the gate fee altogether?

Tables 5–7 show the semantic gap in common LLMs for an LRL like Yorùbá, as reflected by the expected numerals in each question and their misrepresentations by the models.

Table 5. Numerical misrepresentation error analysis, example 1.

Numerals	Expected Representation	ChatGPT	Gemini
otàlélúgba	260	63	230
àrúnlélotàlélúgba	265	65	235

Table 6. Numerical misrepresentation error analysis, example 2.

Numerals	Expected Representation	ChatGPT	Gemini
ogbon	30	40	30
egbèsàn	1800	9	900
àádota	50	15	50

Table 7. Numerical misrepresentation error analysis, example 3.

Numerals	Expected Representation	ChatGPT	Gemini
èedegbejo	1500	800	250
egbàáta	6000	3000	600

Overall, ChatGPT was able to accurately represent numerals completely correctly only in 21 out of 50 (42%) questions and answer the main gist of question 37 out of 50 (74%) times. Gemini, on the other hand, represented the numerals completely correctly in only 23 questions (46%) and was able to correctly interpret the questions 46 times (92%). The error analysis of the two models is presented in Figure 8.

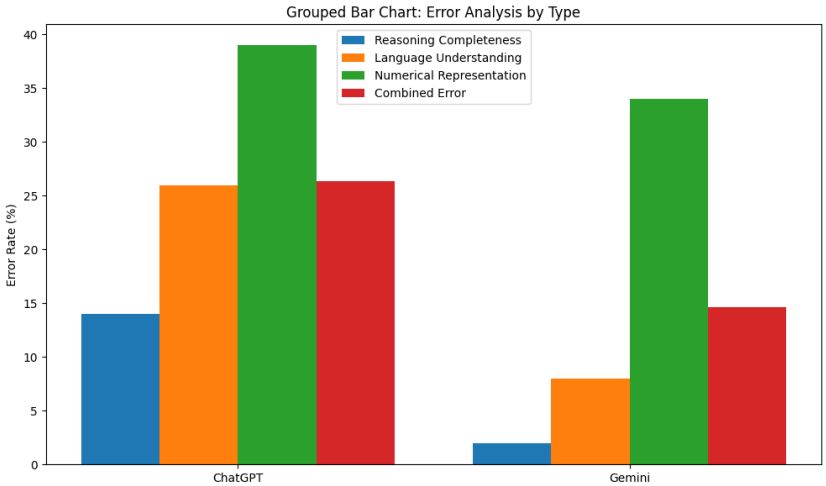


Figure 8. Overview of the error analysis of ChatGPT and Gemini across the 50 questions in set one.

While Gemini obtained better performance than ChatGPT across the three metrics in the first question set, its accuracy in the other two question sets was much lower than that of ChatGPT. PaLM did not show any logic in arriving at its answers and showed no logical flow.

5. Conclusions

In this paper, we investigated the ability of LLM to handle NLR tasks for LRLs. Set up in a zero-shot environment, three LLMs (ChatGPT, Gemini, and PaLM) were exposed to three sets of questions with limited contamination, covering basic arithmetic, time, days and months of the year, and the numeral systems in the Yorùbá language. The best-performing model of the three (based on the total number of questions correctly answered), ChatGPT, was able to generate correct answers, showing logical steps to attaining the answers in two of the three question sets (basic arithmetic, time, days, and months) but failed significantly in numeral systems. Our assessment of the models based on their reasoning completeness, language understanding, and numerical representation accuracy reveals the two fundamental errors that led to the sub-par results produced by the three models. The first is the misrepresentation of the Yorùbá numerals in figures used for calculation by the models. This was the case in 29/50 questions (58%) for ChatGPT and in 27 of 50 questions (54%) for Gemini in the basic arithmetic question set. The second error was shown by the language understanding metric, where we assessed if the models understood the main gist of the questions. Again, ChatGPT misinterpreted 13 of 50 questions (26%), with Gemini only misinterpreting the questions on 4 occasions (8%) in the basic arithmetic question set.

The tasks involving the accurate representation of Yorùbá numerals presented the greatest challenge for the three models. The reasons for this included the difference between the vigesimal setup of the Yorùbá numerals and the subtlety involved in deriving the various Yorùbá numerals, unlike most LLMs, which have been trained using the decimal system. In the Yorùbá numeral system, the numbers 20 (ogún), 200 (igba), 2000 (egbàá), and 20,000 (oke) are critical bases or reference points. They are multiplicative foundations from which other numerals are constructed [37]. Also, the insufficient or limited training on Yorùbá-specific data or numeral-reasoning tasks contributes glaringly to the results achieved. ChatGPT and Gemini showed better exposure to training on the Yorùbá dataset and could generalize better than PaLM.

The performance of LLMs in tasks involving LRLs will be significantly improved if LLMs can understand LRLs more and can perform NLR in LRLs. To this end, there is a need for the creation of benchmarks to assess how well LLMs understand, can reason, and generalize in specific LRL domains such as culture, tradition, technology, education, etc., in LRLs. Also, there is a need for the curation of datasets in the Yorùbá language that addresses the numeral and counting systems, as well as time, days, and months.

Author Contributions: Conceptualization, F.O. and P.O.O.; methodology, F.O. and P.O.O.; validation, F.O. and P.O.O.; investigation, F.O. and P.O.O.; resources, F.O. and P.O.O.; data curation, F.O. and P.O.O.; writing—original draft preparation, F.O.; writing—review and editing, P.O.O.; visualization, F.O. and P.O.O.; supervision, P.O.O.; project administration, P.O.O. funding acquisition, P.O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Details of the Numerical Probes

Appendix A.1. General Arithmetic—First 20 Questions

1. Dupe gba náírà márùn-ún fún owó oúnje kúrò ní ilé. Ó pàdé Ìyá Gbónke tó fún un ní náírà mesàn-án àti Bàbá Alátìṣe tó fún un ní náírà méjìlá bí ó ṣe ń lo sí ilé-ẹ̀ko. Èló ni owó tí Dùpé ní je lápapò?

2. Bàbá Sádé fún Sádé ní eko mewàá. Sádé fún Fáderèrà ní eko méjì. Mélò ni eko owó Sádé kù?

3. Bàbá Jíginni gbà Abeke níyànjú láti je ataare kóòkan lójojumo fún ose méjì gbàkò. Ataare mélò ni ó ye kí Abeke ra?

4. Olùko rí wípé Ààrínolá kò fi etí síle nínú yára ìkeko, ó sì so fún wípé kí ó lo kà iye eranko tí ó wà nínú ogbà ilé ìkeko náà. Tí ese eranko tí ó wà nínú ogbà náà bá je merìndínlógótà, kí ni ìdáhùn Ààrínolá yóò je?

5. Ajani maa n sun fun wakati meẹ̀ọ lojoojumọ. Ti o ba ji ni aago mẹsan ni àárọ̀ oni, aago mélò ni Ajani sun?

6. Bukola pinu láti fi 365 naira pa món ní ojoojúmó fún ọdun kan. Èló ni Bukola yóò ti fi pa món nígbà tí ọdún bá parí?

7. Aduke máa n lọ sí ibi isé ní ojoojúmó. Owó ọ́yà re léyìn osù jé 60000 naira. Tí Aduke kò ba lọ sí ibi isé fún ọjọ́ marun ní osù yìí, èló ni owó osù re yóò jé l'ósù yìí?

8. Àgbè kan máa n lọ sisé l'óko fún wákàtí méfà lójoojúmó. Ní òsè yìí, kò leè lọ sí oko fún ọjọ kan nítorí òjò alágbára kan tí ó rò sulè ọjọ náà. Wákàtí mélòó ni àgbè náà fi sisé l'óko lápapò ní ọ̀sè yìí?

9. Ó gba enìkan ní ọjọ mésàn-án láti gbẹ kànga kan. Ọjọ mélòó ló ma gba ènìyàn meta láti gbé kànja mérindínlógójì?

10. Sade n tún yÀra re se. Ó pinu láti kó asọ ogún géérégé sínú àpótí kòòkan. Ó ní orísìí asọ méji—ankara àti asọ àrán. Tí ó bá ní àpótí meta fún ankara àti àpótí méjì fún asọ àrán, asọ mélòó ni Sade ní lápàpoò?

11. Aduke wọ ọkọ̀ ojú irin láti Èkó lọ sí Sokoto. Ìrìnàjò náà yóò gba wákàtí márùn-dín-lógbón. Tí Aduke bá gbéra ní aago meje àárò ọjọ abameta, ọjọ wo ni Aduke yóò dé Sokoto?

12. Ajani bèrè isé ọde ní aago méje ìròlé, ó sì lo wákàtí méjj̀ilá gbáko. Aago mélòó ni Ajani padà sí ilé?

13. Faderera ní ẹgbèrún mewa lọwọ. Ó fi ìdá kan nínú ìdá marun owó náà ra oúnje, ó fi ìdá meji nínú marun ra ìwé, ósì fi ìda kan nínú merin ra asọ̀. Eló ni owó náà sékù?

14. Tí a bá fi ẹgbẹ̀rúnméjì lé ní ọ̀tÀlénígba naira kọ̀ ọgbà yí oko ẹ̀wà onígun merin ká, èló ni a ná láti kọ̀ abala kan nínú ọgbà náà?

15. Ade ra ẹyin méjìlá ní ẹgbẹrun marun naira. Ó ta ìkòòkan ní ẹgbẹta naira. Èrè èló ni Ade jẹ lórí ọjà náà?

16. Ajoke se ìgbéyàwó ní ọmọ ọdún márùnlélógún. Ó bí ọmọ àkọ̀kọ́ léyìn ọdún marun. Kíni ọjọ orí Ajoke nígbà tí ó bí ọmọ yìí?

17. Ẹ̀ro àgbéléwò kan jẹ̀ ẹgbẹ̀rún márùn-lé-lógójẹ naira. Èdínwó ìdá meje àti àabọ̀ nínú ọgọrun wà lórí ọjà náà tí a bá san owó náà lójú esè. Èló ní oníbàráà yóò san lórí ọjà yìí tí ó bá san owó náà lójú ẹsẹ̀?

18. Baba olówó kan ní òjìlénígba malu tí ó fẹ́ pín fún àwọn ọmọ rè mẹta ní ìwọ̀n meji sí mẹta sí marun. Mélòó ni ọmọ kòòkan yóò rí gbà?

19. Janduku kan lọ sí ilé ìrajà kan, ó sì jí ẹgbẹrún meji naira. Ó fi owó tí ó jí náà ra ìpápánu tí ó jé ẹgbẹrunkan-lé-nígba naira, ó sì gba àpò owó merin padà. Èló ni ilé ìrajà náà pàdánù?

20. Alahji Malami ní malu ọgọ̀òrún, ogójì jé ako malu, mewa nínu abo malu náÀ ti bí ọmọ meta meta. Ọgbọn nínu abo malu náà ti bí ọmọ méjì méjì. Gbogbo awọn abo malu tókù ti bí ọmọ merin, merin. Mélòó ni gbogbo malu Alahji lápàpoò?

Appendix A.2. Time, Days, and Months—First 10 Questions

1. Tí òní bá jẹ ojo ajé, ojo wo ni ojo mewa òní yóò je?
2. Ojo àkókó nínú oṣù Ope odún yìí bo sí ojo Ojorú. Ojo wo ni ètàdínlógún òní yóò bo sí?
3. Ojo merin merin ni a máa n ná ogà DùgbèṪí a bá ná ojà náà ní ojo Etì ní osè tí ó kọjà ojo wo ni ojà náà yóò bo sí ní osè yìí?
4. Ojo mélòó ló wà nínú oṣù Agemọ?
5. Oṣù mélòó nínú odún ni o ní ogbòn ojo géérégé?
6. ọjọ ọjÀ Owódé tí ó kọjá ni a bí Abebi. ọjọ wo ni ọjọ tí a bí Abebi bọ sí tí òní ọjọ ọjọ̀bọ bá jé ọjọ ọjÀ Owode mírÀn, tí a sì n ná ọjÀ náÀ ní ọjọ méfÀ mẹ̀fÀ?
7. Ọjọ mẹlòó ló wà nínú oṣù Èrèlé
8. Ti ọjọ àkọ̀kọ́ ọdun yìí bá bọ sí ọjọ Àbámẹ̀ta, kínni ọjọ tí ó parí oṣù àkọ̀kọ́ yóò jé?
9. Ti ọjọ àkọ̀kọ́ ní oṣù Ògún b abọ si ọjọ Ìsẹ̀gun, ọjọ wo ni ọjọ àkọ̀kọ́ ní Oṣù Owewe yóò bọ sí?
10. Ti ọjọ àkọ̀kọ́ nínú ọdún bá bọ sí ọjọ Àìkú, ọjọ Àìkú mélòó ni yóò wà nínú ọdún náà? dahun awon ibeere wonyii

Appendix A.3. Onka—First Twenty Questions

1. Kinni numeri yii: 660, ni oro?
2. Onka wo lo tele Ogúnlélúgba?
3. Kinni Òjìlénírinwó ni onka?
4. Kinni Ẹ̀ẹ́dẹ́gbẹ̀rin ni onka?
5. Kinni numeri yii: 130, ni ọrọ
6. Ẹ̀ẹ́dẹ́gbÀáfa tunmo si kinni ni onka Yoruba?
7. Kinni ÀádọwÀálélúgba ni onka Yoruba?
8. Kinni Ẹ̀sánlélójìlélúgba ni onka Yoruba?
9. Kinni ỌkÀnlélọ́gbọ̀nlélúgba ni onka Yoruba?
10. Kinni Ẹ̀sándínlógúnlélúgba ni onka Yoruba?
11. Kinni Ọgọ́jọ ni onka Yoruba?
12. Onka wo ló tèlé Àádọ̀rin?
13. Kinni numeri yii: 290, ni oro
14. Kinni numeri yii: 480, ni oro
15. Onka wo ló tèlé Àádọ̀jọ?
16. Kinni numeri yii: 570, ni oro
17. Kinni numeri yii: 350, ni oro
18. Kinni numeri yii: 920, ni oro
19. Kinni numeri yii: 710, ni oro
20. Kinni numeri yii: 840, ni oro

References

1. Kasai, J.; Kasai, Y.; Sakaguchi, K.; Yamada, Y.; Radev, D. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv* **2023**, arXiv:2303.18027.
2. Wu, J.; Wu, X.; Qiu, Z.; Li, M.; Lin, S.; Zhang, Y.; Zheng, Y.; Yuan, C.; Yang, J. Large language models leverage external knowledge to extend clinical insight beyond language boundaries. *J. Am. Med. Inform. Assoc.* **2024**, *34* , 2054–2064. [CrossRef] [PubMed]
3. Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A.R.; Zhang, S.; Sun, Y.; Wang, W. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv* **2023**, arXiv:2307.10635.

4. Imani, S.; Du, L.; Shrivastava, H. Mathprompter: Mathematical reasoning using large language models. *arXiv* **2023**, arXiv:2303.05398.
5. Peng, X.; Geng, X. Self-controller: Controlling LLMs with Multi-round Step-by-step Self-awareness. *arXiv* **2024**, arXiv:2410.00359.
6. Ji, K.; Chen, J.; Gao, A.; Xie, W.; Wan, X.; Wang, B. LLMs Could Autonomously Learn Without External Supervision. *arXiv* **2024**, arXiv:2406.00606.
7. Zhang, W.; Shen, Y.; Wu, L.; Peng, Q.; Wang, J.; Zhuang, Y.T.; Lu, W. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 12–14 August 2024.
8. Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; Yang, D. Is ChatGPT a general-purpose natural language processing task solver? *arXiv* **2023**, arXiv:2302.06476.
9. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv* **2022**, arXiv:2205.11916.
10. Li, L.; Wang, Y.; Zhao, H.; Kong, S.; Teng, Y.; Li, C.; Wang, Y. Reflection-Bench: Probing AI intelligence with reflection. *arXiv* **2024**, arXiv:2410.16270.
11. Fang, X.; Xu, W.; Tan, F.A.; Zhang, J.; Hu, Z.; Qi, Y.; Nickleach, S.; Socolinsky, D.; Sengamedu, S.H.; Faloutsos, C. Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding—A Survey. *arXiv* **2024**, arXiv:2402.17944.
12. Kadlčík, M.; Štefánik, M. Self-training Language Models for Arithmetic Reasoning. *arXiv* **2024**, arXiv:2407.08400.
13. Cai, C.; Zhao, X.; Liu, H.; Jiang, Z.; Zhang, T.; Wu, Z.; Hwang, J.N.; Li, L. The Role of Deductive and Inductive Reasoning in Large Language Models. *arXiv* **2024**, arXiv:2410.02892.
14. He, P.; Li, Z.; Xing, Y.; Li, Y.; Tang, J.; Ding, B. Make LLMs better zero-shot reasoners: Structure-orientated autonomous reasoning. *arXiv* **2024**, arXiv:2410.19000.
15. Akhtar, M.; Shankarampeta, A.; Gupta, V.; Patil, A.; Cocarascu, O.; Simperl, E.P.B. Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data. *arXiv* **2023**, arXiv:2311.02216.
16. Shao, Z.; Huang, F.; Huang, M. Chaining Simultaneous Thoughts for Numerical Reasoning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–9 December 2022.
17. Mishra, S.; Mitra, A.; Varshney, N.; Sachdeva, B.S.; Clark, P.; Baral, C.; Kalyan, A. NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022.
18. Zhou, F.; Dong, H.; Liu, Q.; Cheng, Z.; Han, S.; Zhang, D. Reflection of Thought: Inversely Eliciting Numerical Reasoning in Language Models via Solving Linear Systems. *arXiv* **2022**, arXiv:2210.05075.
19. Heston, T.F.; Khun, C. Prompt Engineering in Medical Education. *Int. Med. Educ.* **2023**, *2*, 198–205. [CrossRef]
20. Wong, W. Practical Approach to Knowledge-based Question Answering with Natural Language Understanding and Advanced Reasoning. *arXiv* **2007**, arXiv:0707.3559.
21. Yu, F.; Zhang, H.; Wang, B. Natural Language Reasoning, A Survey. *Acm Comput. Surv.* **2023**, *56*, 1–39. [CrossRef]
22. Rybak, P. Transferring BERT Capabilities from High-Resource to Low-Resource Languages Using Vocabulary Matching. In Proceedings of the International Conference on Language Resources and Evaluation, Turin, Italy, 20–25 May 2024.
23. Cahyawijaya, S.; Winata, G.I.; Wilie, B.; Vincentio, K.; Li, X.; Kuncoro, A.; Ruder, S.; Lim, Z.Y.; Bahar, S.; Khodra, M.L.; et al. IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *arXiv* **2021**, arXiv:2104.08200.
24. Joshua, A.M. Improving Question-Answering Capabilities in Large Language Models Using Retrieval Augmented Generation (RAG): A Case Study on Yoruba Culture and Language. In Proceedings of the 5th Workshop on African Natural Language Processing, Vienna, Austria, 11 May 2024.
25. Kale, M.; Rastogi, A. Few-Shot Natural Language Generation by Rewriting Templates. *arXiv* **2020**, arXiv:2004.15006.
26. Adelani, D.I.; Liu, H.; Shen, X.; Vassilyev, N.; Alabi, J.O.; Mao, Y.; Gao, H.; Lee, A.E.S. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023.
27. Sánchez-Salido, E.; Morante, R.; Gonzalo, J.; Marco, G.; de Albornoz, J.C.; Plaza, L.; Amig'o, E.; Fern'andez, A.A.; Benito-Santos, A.; Espinosa, A.G.; et al. Bilingual Evaluation of Language Models on General Knowledge in University Entrance Exams with Minimal Contamination. *arXiv* **2024**, arXiv:2409.12746.
28. Perelkiewicz, M.; Poswiata, R. A Review of the Challenges with Massive Web-mined Corpora Used in Large Language Models Pre-Training. *arXiv* **2024**, arXiv:2407.07630.
29. Liu, Y.; Xu, M.; Wang, S.; Yang, L.; Wang, H.; Liu, Z.; Kong, C.; Chen, Y.; Sun, M.; Yang, E. OMGEval: An Open Multilingual Generative Evaluation Benchmark for Large Language Models. *arXiv* **2024**, arXiv:2402.13524.
30. Wang, Y.; Zhao, Y. Gemini in reasoning: Unveiling commonsense in multimodal large language models. *arXiv* **2023**, arXiv:2312.17661.

31. Anil, R.; Dai, A.M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. Palm 2 technical report. *arXiv* **2023**, arXiv:2305.10403.
32. Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; Duan, N. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv* **2023**, arXiv:2304.06364.
33. Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H.W.; Tay, Y.; Ruder, S.; Zhou, D.; et al. Language models are multilingual chain-of-thought reasoners. *arXiv* **2022**, arXiv:2210.03057.
34. Xu, F.; Lin, Q.; Han, J.; Zhao, T.; Liu, J.; Cambria, E. Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 1620–1634. [CrossRef]
35. Huang, Y.; Tang, K.; Chen, M. A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry. *arXiv* **2024**, arXiv:2404.15777.
36. Nguyen, T.H.; Le, A.C.; Nguyen, V.C. ViLLM-Eval: A Comprehensive Evaluation Suite for Vietnamese Large Language Models. *arXiv* **2024**, arXiv:2404.11086.
37. Adetomiwa, A. Yorùbá Numeral System in 21st Century: Challenges and Prospects. 2023. Available online: https://www.researchgate.net/publication/375096095_YORUBA_NUMERAL_SYSTEM_IN_21ST_CENTURY_CHALLENGES_AND_PROSPECTS (accessed on 20 February 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.