

Leveraging Large Language Models for a Swahili Mathematics ITS in Tanzania: Designing Effective Prompts

Edger P. Rutatola^{1,2}[0009–0000–2401–3672], Koen Stroeken³[0000–0003–3409–444X],
and Tony Belpaeme¹[0000–0001–5207–7745]

¹ IDLab-AIRO, Ghent University – imec, Ghent, Belgium
<https://airo.ugent.be/>

² Department of Computing Science Studies, Faculty of Science and Technology,
Mzumbe University, Morogoro, Tanzania

³ CARAM, Ghent University, Ghent, Belgium
<https://research.flw.ugent.be/en/caram>

Abstract. The advancement of Large Language Models (LLMs) has significantly enhanced intelligent tutoring systems, enabling them to engage learners through natural dialogues. This interaction boosts learner engagement but presents challenges for low-resource languages, such as Swahili — Tanzania’s national language. By design, LLMs rely on patterns learned during training to predict subsequent words, making them more suited for conversational tasks than factual computations and reasoning tasks, such as solving mathematics problems. This study investigates the suitability of GPT-4 in generating Swahili-language mathematics content for teaching geometry to primary school students, assessing both contextual and factual accuracy. Using nine varied prompts, we generated 621 different topic introductions, which were evaluated by primary school mathematics teachers. Results reveal that GPT-4 can generate contextually relevant content but struggles with complex mathematical computations. Additionally, the prompt variations provided valuable insights into designing effective prompts for similar tasks.

Keywords: GPT-4 · Large Language Models · mathematics education · intelligent tutoring systems · prompt engineering · Swahili

1 Introduction

The United Republic of Tanzania (URT) has Swahili and English as official languages, with most of the population being more conversant in Swahili [15]. These two languages also serve as the medium of instruction in primary schools, with the Swahili language being used in the majority of public (state-owned) schools. At the same time, English is dominant in private schools [17]. Furthermore, as of 2022, 89.2% of the primary schools in the country were public schools [19].

A look into the results of the URT’s national primary school leaving examination (PSLE), the passing of which is a requisite to joining secondary education,

reveals a significant failure in the mathematics subject. Specifically, 46.72% and 51.17% of the students failed in mathematics in 2022 and 2023, respectively [3]. This is, in part, caused by acute poor teacher-student ratios which limit personalisation, pupils’ engagement, and overall class management [18,20].

One of the ways to ensure pupils’ inclusion, engagement, and personalisation in the learning process —despite the class size— is through the use of intelligent tutoring systems. Globally, following the advancement and increased accessibility of technology, different intelligent tutoring systems are being developed and adopted [4,14]. While these systems are effective, they used to lack in interaction with learners through robust conversations, which has been shown to enhance their engagement [24].

Recently, leveraging Large Language Models (LLMs) has proven to be one of the more effective ways to make intelligent tutoring systems more interactive, autonomous, and engaging [5]. Among the various LLMs available today, OpenAI’s ChatGPT stands out as the most popular, thanks to its open accessibility and wide-ranging capabilities [8]. By the time of conducting this research, GPT-4 was OpenAI’s most current model. Furthermore, the quality of the content produced by LLMs greatly depends on the availability of training data. Even powerful LLMs such as GPT-3.5 and GPT-4 underperform when dealing with low-resource languages [23]. Unfortunately, Swahili only contributes to about 0.009% of GPT’s training data, making it a low-resource language [10].

Several studies highlight ChatGPT’s ability to perform adequately in providing professional content in low-resource languages. For example, a study by Maginga et al. [16] evaluating GPT’s performance as a Swahili agricultural extension tool reports 89% of content correctness and 69% of good language use. While this holds promise, it would be naive to just assume that a similar level of correctness would be achieved in subjects like mathematics, which require step-by-step problem solving rather than the simple prediction of the next word. In addition to factual correctness, the resourcefulness of a language — culture, to an extension — leads to biases and affects the contextual correctness of the generated content by an LLM [1]. Contextual correctness is equally important, especially in teaching and learning, as the contextual correctness and relevance of examples given are crucial to understanding and retention.

These limitations further complicate the use of LLMs in education, underscoring the need for careful prompt engineering. Although powerful, LLMs depend heavily on prompts to guide the content they produce [28]. In addition, LLM APIs charge per token, which includes both the prompt and the output [21]. This makes it essential to obtain the desired content with as few trials as possible, highlighting the importance of effective prompt design. One approach is few-shot prompting, where examples of the desired output are incorporated into the prompt [6]. Although often effective, few-shot prompting can be inefficient and less ideal in dynamic learning environments that demand heterogeneous outputs [7]. Also, approaches such as Chain-of-Thought prompting have proven effective for tasks that require step-by-step reasoning [11,27], but it remains important to consider issues like contextual correctness and relevance. Hence, there is a need

for prompt design techniques that are both effective and efficient, incorporating only the necessary components to yield correct and relevant outputs.

Therefore, three research questions are addressed in this paper, which are: (1) Can GPT-4 generate Swahili content with contextually relevant examples to help a Tanzanian pupil understand mathematics concepts?, (2) can GPT-4 correctly and clearly handle the step-by-step mathematical calculations to enable a pupil to understand the process?, and (3) in teaching mathematics in the Swahili language, how does the variation of the prompts improve the quality of content generated?

2 Methodology

2.1 Study Design

The study was conducted in the United Republic of Tanzania. Being the domain experts, and based on our research questions, primary school mathematics teachers were the main respondents. We involved teachers from both private and public schools so as to fully capture the contextual diversification. Moreover, we sought data from both groups to be able to compare the results and shed more light on the disparities between high- and low-resource languages in LLMs.

We chose geometry as the topic of intervention, focusing on the calculations of areas and perimeters of various geometrical shapes. This selection allowed us to incorporate problems of varying complexity, enabling us to effectively evaluate the content generated by GPT-4. To that end, we formulated 16 sub-topics: (1) perimeter of a square; (2) area of a square; (3) perimeter of a rectangle; (4) area of a rectangle; (5) circumference of a circle with a diameter that is divisible by 7; (6) circumference of a circle with a diameter that is not divisible by 7; (7) circumference of a circle with a radius that is divisible by 7; (8) circumference of a circle with a radius that is not divisible by 7; (9) area of a circle with a diameter that is divisible by 7; (10) area of a circle with a diameter that is not divisible by 7; (11) area of a circle with a radius that is divisible by 7; (12) area of a circle with a radius that is not divisible by 7; (13) surface area of a cylinder closed on one side, having a diameter that is divisible by 7; (14) surface area of a cylinder closed on one side, having a diameter that is not divisible by 7; (15) surface area of a cylinder closed on one side, having a radius that is divisible by 7; and (16) surface area of a cylinder closed on one side, having a radius that is not divisible by 7.

Perimeter and area of rectangles and squares were assumed to be simple topics as their calculations only involve basic additions and multiplications. Area and circumference of circles were categorised as having medium complexity and interesting diversity, as they involve fractions (if radius/diameter is a factor of 7, thus $\pi = \frac{22}{7}$) or decimals (if radius/diameter is not a factor of 7, thus $\pi = 3.14$). Lastly, the surface area of a cylinder closed on one side was labelled as complex, because it involves relatively more complex calculations and needs attentive step-by-step solving.

2.2 The application

We created a comma-separated values (.csv) file with the following details for each topic, in both English and Swahili as separate columns: (1) the name of the shape (2) the calculation to be done (area or perimeter); (3) the formula for the calculation; and (4) the full name of the topic. We then created a Python script containing an API call to the GPT-4 model. Though GPT-4 is not specifically known for its mathematical precision, our decision to have it handle the computations was motivated by its demonstrated ability to accurately understand and solve relatively complex math problems compared to other models [22]. Additionally, GPT-4 has proven capable of comprehending and solving undergraduate-level math[11]. Given the simpler level of math in our study and the need for dialogue-based explanations, GPT-4 was the most suitable choice. The script loaded the CSV file in each runtime and consequently requested GPT-4 to generate introductory content for each topic. To set the context, we defined some variables (in both English and Swahili, where applicable) such as the model’s name and role (Zuhura and teacher, respectively); the pupil’s name and age; the country, Tanzania; and the languages, English and Swahili, to be used in the generated output as specified in the respective prompt. We also provided a flattering persona for the model:

You are a helpful and knowledgeable mathematics tutor, who performs calculations accurately, and explains concepts clearly, step by step, and in an easy way.

Nine (9) different prompts were designed to generate the content, as elaborated in the supplementary material⁴. In a nutshell, the variations in the prompts were as follows: an English prompt demanding an English response (prompt 1); a Swahili version of prompt 1 demanding a Swahili response (prompt 2); an English prompt (same as prompt 1) requiring an answer in Swahili (prompt 3); an English prompt demanding a Swahili response, but it has a custom dictionary attached for correct contextual translations of some words (prompt 4); a Swahili version of prompt 4 (prompt 5); prompt 4, but this time we also added the correct formula for the calculations (prompt 6); a Swahili version of prompt 6 (prompt 7); prompt 6, with the addition of the expected correct answer (prompt 8); and, a Swahili version of prompt 8 (prompt 9). Thus, for each topic, nine different introductions were generated based on each prompt. Moreover, the script rendered the introductions in a PDF file, with each topic having an independent file.

For dynamic examples, the dimensions of the figures were randomly generated in the script, with defined floor and ceiling values. Only when a value needed to be perfectly divisible by 7 was that criterion added to the random number generator. Furthermore, we utilised the Matplotlib library to generate illustrations of these figures, including their dimension labels, and incorporated them into the PDF files for their respective topics.

⁴ Annex 1 – The Prompts: https://drive.google.com/drive/folders/1djgFY-TSpN1AY-qoaEOJYK7CkgqyAG-?usp=drive_link

2.3 Data Collection

The study employed a mixed-methods approach, combining qualitative and quantitative data collection. A total of 621 topic introductions generated by the system were evaluated. Ten primary school mathematics teachers qualitatively assessed the content’s relevance, correctness, and contextual appropriateness for teaching. Each teacher reviewed and annotated six to seven documents (54–63 topic introductions).

For the second research question, the 621 contents were categorised into four groups: (1) correct flow, wrong answer; (2) correct answer, wrong flow; (3) correct flow and answer; and (4) wrong flow and answer. These results were quantitatively recorded and analysed.

2.4 Data Analysis

The qualitative results from the teachers were thematically analysed. We manually coded their feedback and grouped them into common themes. Moreover, the teachers’ demographic information and our quantitative evaluation of the content were both descriptively analysed.

3 Results

3.1 Teachers’ Demographic Information

Seven of the ten teachers worked in public schools (teaching in Swahili), while the rest were in private schools (teaching in English). 60% were aged 20-30, with 20% in both the 31-40 and 41-50 age groups. Five teachers had less than 5 years of experience, two had 6-10 years, and three had over 10 years, providing insights from both new and experienced educators. Regarding AI and LLM awareness, nine responded: six had heard of it but never used it, two had used it, and one was unaware of AI. This highlights AI’s novelty in Tanzanian education.

3.2 Qualitative Evaluation of the Content by the Teachers

Theme1: Contextual Appropriateness and Relevance – In general, all teachers agreed that the provided examples were appropriate for classroom learning. Nothing unethical or culturally inappropriate was noted in all the evaluated content. Moreover, the examples used were deemed contextually relevant, as they were drawn from the pupils’ everyday encounters, making learning easier. Some quotes to commend the appropriateness and relevance of the introductions are:

The introduction has prepared the child appropriately for the lesson.
The mention of a garden and circular fields makes the lesson real (Topic: circumference of a circle).

Appropriate examples have been used, for example, a book’s surface (Topic: area of a rectangle).

Nevertheless, it is important to note that three (3) teachers pointed out that the introductions were too wordy. One of the teachers even gave an introduction a perfect score just because it was short and straightforward.

Theme 2: Contextual Correctness – The introductions were appropriate and relevant but lacked contextual correctness. For example, the model used pizza to explain the area and circumference of circles. While pizza is familiar in many places, it is considered a luxury in Tanzania and unfamiliar to most public school pupils, as noted by five public school teachers. However, private school teachers approved of the example, highlighting cultural and lifestyle differences between public and private school pupils, which are likely to impact learning.

A notable issue was the incorrect use of SI units. In Prompt 3, “*cm*” (centimetre) was used instead of the correct Swahili form, “*sm*” (sentimita). Additionally, seven teachers noted that in Swahili, the SI unit precedes the number, e.g., *sm* 5 for 5 cm.

Theme 3: Factual Correctness – **(1) In the examples:** The model mostly provided relevant examples, such as soda cans, buckets, and cooking pots for cylinders; gardens, tables, walls, and books for rectangles and squares; and wheels, plates, clocks, and coins for circles. However, some errors occurred, such as using running around a football pitch to explain a circle’s circumference. A teacher who identified this error stated:

An example given states that knowing the circumference of a circle can help you determine the distance you have to run around a football pitch. Though this is a good example to understand the concept of circumference, a football pitch is not a circle.

Five teachers observed that some examples used the amount of paint needed to explain areas and perimeters, but they noted that paint quantity (measured in litres or cubic centimetres/metres) is typically associated with volume calculations, not areas or perimeters.

(2) In the calculations: The model performed well in calculations for certain topics, handled some adequately, and struggled with others. Specifics for this are presented in section 3.3. However, some teachers noted incorrect formulas, inaccurate calculations, and correct but not best use of equations. Incorrect equations were mainly observed in the calculations of the surface area of cylinders, where the two correct formulas are “ $A = \pi r^2 + 2\pi rh$ ” or “ $A = \frac{\pi d^2}{4} + \pi dh$ ”, given radius or diameter, respectively. The model often used these wrong equations, e.g. “ $A = \pi r(r + h)$ ”, which expands to “ $A = \pi r^2 + \pi rh$ ”; and “ $A = \pi r^2 h$ ”, which calculates volume. This was remedied in prompts 6 – 9, by explicitly providing the correct formulas in the prompts.

Inaccuracies were observed in calculations involving fractions and decimals, particularly when using π . More complex calculations, such as for cylinders, increased the likelihood of errors. In some cases, the model avoided calculations altogether, as shown below (translated from Swahili):

Now, let us take an example of a cylinder with a diameter of 7 cm and a height of 17 cm.

Remember, $\pi = \frac{22}{7}$.

We will use this formula: $A = \frac{\pi d^2}{4} + \pi dh$

Step one is to take the diameter (d) and substitute it into the formula.

This will be $(\frac{22}{7} \times 7^2 \div 4)$.

Step two, take the height (h) and the diameter (d) and substitute them into the formula. This will be $(\frac{22}{7} \times 7 \times 17)$.

Then, add the results from the first and second steps to get the final answer.

This is a simple formula and should always be used. Remember, practice makes perfect.

Teachers noted that students are expected to use the equation corresponding to the given radius or diameter in class. However, the model often disregarded this, using correct but mismatched formulas, as emphasized by four teachers. The model showed a preference for formulas with radius, unless instructed otherwise. In hindsight, this might be because of being tasked to teach the pupil the simplest formula for the task. Computations with radius are simpler, especially for areas, as they do not involve division. For example, $A = \pi r^2$ is simpler than $A = \frac{\pi d^2}{4}$. This issue was addressed in prompts 6–9, where the correct formulas were incorporated. We also observed that the language of instruction, i.e. English in Prompt 1 and Swahili in the rest, did not impact the correctness of the calculations.

Theme 4: Teaching methods – (1) Use of Examples: Unanimously, teachers agreed that the model used sufficient examples that enhance learning. This was commended as good practice because it draws students to learning and stimulates them to think.

(2) Pupils’ Engagement: The model was praised for its examples but sometimes failed to engage pupils. Teachers suggested various engagement strategies, including: asking questions to assess prior knowledge (2 teachers), measuring objects related to the topic (5 teachers), confirming understanding (5 teachers), and having pupils solve arithmetic problems (1 teacher). Overall, all teachers emphasised the importance of pupil engagement, with one suggesting using a flipped classroom approach:

The introductions are one-sided. The teacher should not be the only source of knowledge in class. We believe that the pupils also have something to offer, so they should be given a chance.

(3) Teaching Aids: The model sometimes relied on imagined scenarios (e.g., football pitch, garden, pizza). Teachers stressed that effective learning requires examples of items readily available in class, such as books, tables, or walls. For topics involving uncommon objects, teachers either prepare them beforehand or

ask students to bring items like buckets or soda cans, which serve as effective teaching aids.

(4) Step-by-Step Approach: This aspect drew significant attention from teachers. In many instances, the model failed to explain the meaning and importance of each equation component, as noted by five teachers. A common example was the lack of explanation about π , its relation to diameter and circumference, and its role in calculations. Students were told its value and asked to memorise it and related equations. Additionally, six teachers observed that the model often skipped step-by-step demonstrations of arithmetic computations, e.g.:

Now, let's calculate the surface area step by step:

First, we calculate the curved surface area:

$$\pi \times \text{diameter} \times \text{height} = 3.14 \times 18\text{cm} \times 20\text{cm} = 1130.4\text{cm}^2$$

Next, we calculate the area of the base:

$$\pi \times (\text{radius})^2 = 3.14 \times (9\text{cm})^2 = 3.14 \times 81\text{cm}^2 = 254.34\text{cm}^2$$

Finally, we add these two areas together:

$$1130.4\text{cm}^2 + 254.34\text{cm}^2 = 1384.74\text{cm}^2$$

So, the surface area of the drum is 1384.74cm^2 .

That's the area you would need to cover if you were to paint the outside of the drum!

On the same note, one teacher further stressed that the math could have been easier if the model had simplified the fractions step-by-step:

Where the radius or diameter is divisible by 7, we normally use the π in fraction form ($\frac{22}{7}$). Then, we show the pupils how they can simplify the calculations by simplifying the fractions. The simplification was not explained in the calculations.

(5) Psychological Engagement: Teachers praised the model for preparing pupils' mindsets and building their confidence by emphasizing the subject's simplicity and approachability.

3.3 Topic Evaluation

This section evaluates GPT-4's mathematical correctness in its solutions. As shown in Fig. 1, the model excelled in simple topics (e.g., area and perimeter of squares and rectangles) and performed satisfactorily in medium-complexity topics (e.g., area and circumference of circles). However, it struggled with complex topics, such as the surface area of a cylinder.

We further analysed the introductions, categorising them into four groups: correct flow, wrong answer; correct answer, wrong flow; correct answer and flow; and wrong answer and flow. As shown in Fig. 2, the most common cause of mathematical errors was the model failing to find the correct answer despite following the right flow. It rarely produced correct answers with the wrong flow. Notably, for the surface area of cylinders, the model often got both the flow and answer wrong, consistent with section 3.2 (Theme 3), where it was observed that incorrect formulas were repeatedly used unless explicitly provided in the prompt.

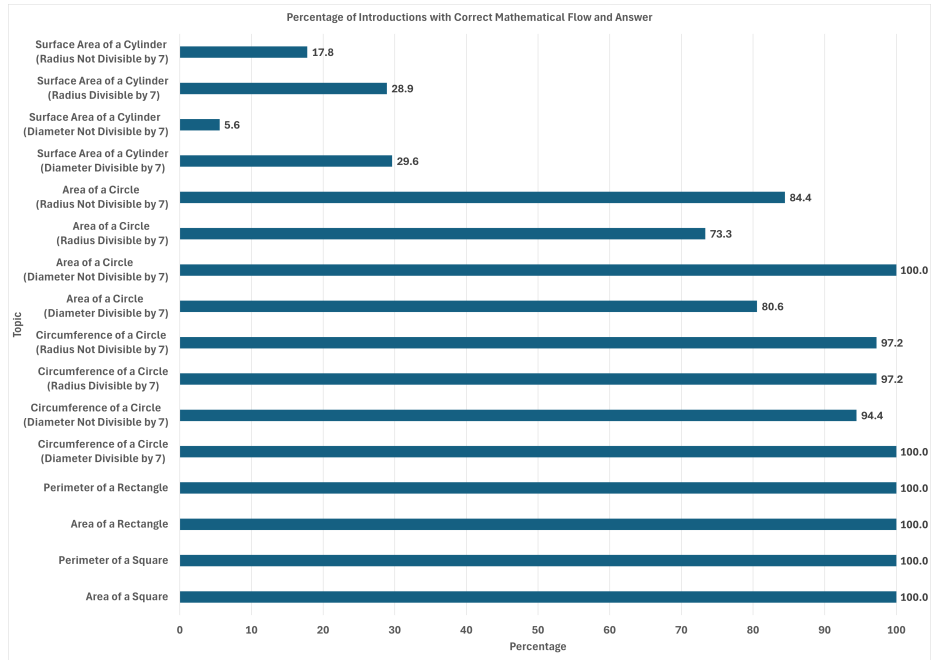


Fig. 1. Correctness of tutoring per topic

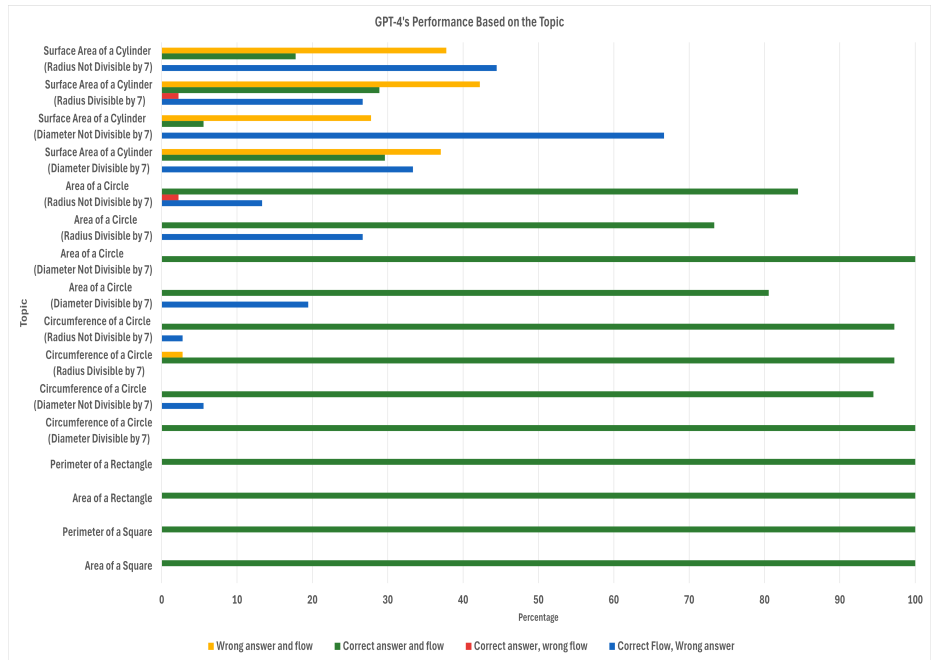


Fig. 2. Breakdown of failures per topic

3.4 Prompt evaluations

We quantitatively assessed how often each prompt produced mathematically correct computations, including flow and answers, as shown in Fig. 3. While no prompt was 100% accurate, prompts 8 and 9 performed best, with prompts 3 and 5 performing worst.

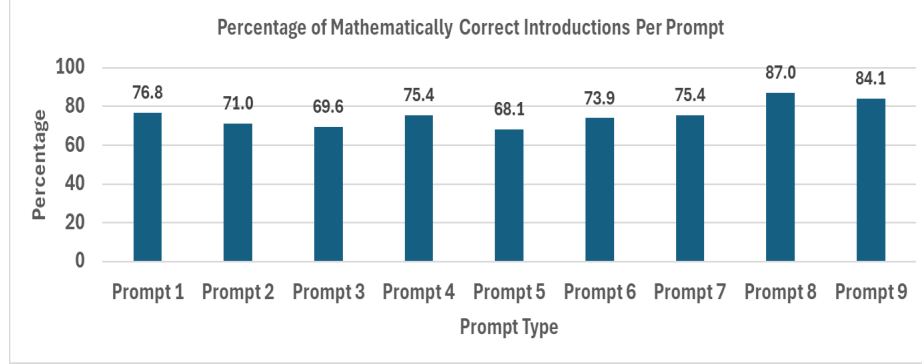


Fig. 3. Evaluation of prompts

4 Discussion

Our domain experts value **effective teaching methods** above all: encouraging active learner engagement, using demonstrations and examples effectively, offering step-by-step problem-solving, and offering encouragement for learners. These align with recommendations on wider effective teaching strategies [12,26]. Unsurprisingly, our domain experts believe that AI-generated content should incorporate these attributes for ITSs to be effective.

Furthermore, our findings demonstrate that, though not primarily developed to solve math, **GPT-4 for now can only handle simple mathematical computations**. This is directly tied to its nature as an LLM, where it simply generates content by predicting the next best word based on the patterns in the data it has been trained on. Thus, if the math calls for a simple calculation such as $Perimeter = 2 \times (Length + Width)$, it will get it right as it has encountered such addition and multiplication patterns in its training data. Unfortunately, when it comes to more complex math involving multi-step reasoning, such as the surface area of a cylinder, the accuracy declines as its predictive nature fails. To mitigate this, external libraries have to be engaged. This confirms the findings in the study by [25], where several GPT models were tested with mathematical modelling tasks of varying complexity. The future availability of neurosymbolic LLMs is expected to remedy this [9].

Additionally, our findings demonstrate that the **resourcefulness of a language and culture significantly impacts the output of LLMs**. This was evident in the cultural bias found in some of the generated examples. For instance, the model used 'pizza' to illustrate the area and circumference of circles – a concept that, while familiar globally, was considered unfamiliar and even luxurious to most public school pupils in Tanzania. This aligns with the study by [1], which similarly highlights ChatGPT’s limitations in cultural relevance and contextual awareness in a low-resource context.

The study by Frieder et al. [11] shows that **prompt engineering** and specifically “chain-of-thought prompting” (e.g. telling the model to solve a problem step by step) improves (Chat-)GPT’s performance in solving math. This was also observed in our study, where the later prompt versions (prompts 8 and 9), which had the correct formula and the expected answer, performed better than the rest. However, in teaching the pupils, getting the context right is also as important. Therefore, our study highlights the crucial details that need to be included in the prompts for both factual and contextual correctness, as described in Table 1.

The inclusion of this information will solve most of the contextual and factual issues observed in sections 3.2 and 3.3. Moreover, it should be noted that the custom dictionary is unnecessary if the expected output is in English. Furthermore, in three out of four cases, the versions with Swahili prompts requiring Swahili responses were outperformed by their counterparts having English prompts requiring Swahili responses. This could be because the model comprehends English instructions better than Swahili, demonstrating the gap in language resources ploughed into LLM training [2,13]. Nevertheless, for a more robust application, these parameters could be filled in by the user in a well-designed form, making each interaction unique.

5 Conclusion and Limitations

The study answers the three research questions set. Firstly, yes, although not 100% accurate, GPT-4 can generate Swahili content with contextually relevant examples to help a Tanzanian primary school pupil understand mathematical concepts. As accuracy is of utmost importance in learning, we offer prompt engineering advice to increase factual and contextual correctness. The second research question has more than just a yes or no answer. Though GPT-4 correctly handled the math for simple and moderately complex topics, it underperformed on more complex problems. Moreover, while it demonstrated a step-by-step approach in its reasoning, it did not fully implement that in the calculations, where it simply provided the answers to computations without showing the underlying work out. Finally, the variation of the prompts improves both the contextual and factual quality of the content generated. Adding more contextual information enriches the relevance of the output, while incorporating solution-related specifics, such as formulas and expected answers, enhances mathematical accu-

Table 1. Prompt parameters.

Detail	Relevance	Description/Comments
Model’s name and role	Contextual	Provides the model’s persona for it to engage with the learner
Learner’s name and age	Contextual	To enable the model to personalise the interactions to the learner
Learner’s purpose	Contextual	To inform the model of the nature of the interactions (e.g educational)
The locale	Contextual	Could range from the country, region/city, to a more specific setting like a classroom. This enables the model to provide more specific and relevant examples.
Examples to be used	Contextual and Factual	Should be both, the things available in the learning environment that the learner can interact with (for engagement) and the things they encounter in their daily surroundings.
Custom dictionary	Contextual and Factual	Contains a few contextual translations of terms the model is likely to get wrong. Might be included as a key-value pair.
The formula to be used	Factual	To reduce the chances for errors and confusion for the learner
The expected correct answer	Factual	For dynamicity, it can be computed by a script. This enables the model to fact-check itself.
Step-by-step approach	Factual	To increase the chances of the model systematically arriving at the correct answer. Might also involve a call to an external library to handle this.
Words limit	Contextual	To ensure that the output is not unnecessarily too wordy
An example of the desired output	Contextual	To prevent the errors, such as 35cm instead of <i>sm.35</i> .
Level and means of engagement	Contextual	If needed, explain how and to what extent the model should engage the learner.

racy. Therefore, while GPT-4 holds a significant promise in Swahili-based mathematics tutoring, prompt engineering is yet crucial for its effectiveness.

Our study has several limitations that may have influenced the findings and their generalisability. For instance, we focused solely on GPT-4. Experimenting with other LLMs might yield different results. Moreover, we only engaged teachers as domain experts, but other stakeholders, such as pupils might also provide valuable insights. Finally, our study was limited to geometry, and therefore more research focusing on other mathematics topics of varying complexity should be done to comprehensively understand the model’s capabilities.

Acknowledgments. We are sincerely grateful to all the teachers who offered their valuable time to participate in the study. Their insights are highly treasured. Furthermore, we are extremely thankful to VLIR-UOS for funding the entire study.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmad, I.S., Dudy, S., Ramachandranpillai, R., Church, K.: Are Generative Language Models Multicultural? A Study on Hausa Culture and Emotions using ChatGPT. In: Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP. pp. 98–106. Association for Computational Linguistics (8 2024). <https://doi.org/10.18653/v1/2024.c3nlp-1.8>, <https://aclanthology.org/2024.c3nlp-1.8/>
2. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Holy, B.W., Ziwei, L., Tiezheng, J., Willy, Y., Quyet, C., Do, V., Xu, Y., Fung, P.: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. vol. 1, pp. 675–718. Long Papers (2023), <https://github.com/HLTCHKUST/chatgpt-evaluat>
3. Baraza la Mitihani la Tanzania: Taarifa ya Matokeo ya Mtihani wa Kumaliza Elimu ya Msingi (PSLE) Uliofanyika Septemba, 2023 (2023), https://www.moe.go.tz/sites/default/files/PRESS%20PSLE%202023%20final_231123_121622_1.pdf
4. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.: Social robots for education: A review. *Sci. Robot* **3**, 5954 (2018), <http://robotics.sciencemag.org/>
5. Cao, C.: Scaffolding CS1 Courses with a Large Language Model-Powered Intelligent Tutoring System. pp. 229–232. Association for Computing Machinery (3 2023). <https://doi.org/10.1145/3581754.3584111>
6. Chaaben, M.B., Burgueño, L., Sahraoui, H.: Towards using few-shot prompt learning for automating model completion. In: Proceedings - International Conference on Software Engineering. pp. 7–12. IEEE Computer Society (5 2023). <https://doi.org/10.1109/ICSE-NIER58687.2023.00008>, <https://ieeexplore.ieee.org/abstract/document/10173909>
7. CHANG, K., XU, S., WANG, C., LUO, Y., LIU, X., XIAO, T., ZHU, J.: Efficient Prompting Methods for Large Language Models: A Survey. In: IEEE International Conference on Program Comprehension. vol. 2022-March, pp. 36–47. IEEE Computer Society (2024). <https://doi.org/10.1145/nnnnnnn.nnnnnnn>, <https://arxiv.org/pdf/2404.01077>
8. Coello, C.E.A., Alimam, M.N., Kouatly, R.: Effectiveness of ChatGPT in Coding: A Comparative Analysis of Popular Large Language Models. *Digital* **4**, 114–125 (3 2024). <https://doi.org/10.3390/digital4010005>
9. Colelough, B.C., Regli, W.: Neuro-Symbolic AI in 2024: A Systematic Review. arXiv preprint arXiv:2501.05435 (2025)
10. Common Crawl Foundation: Statistics of Common Crawl Monthly Archives: Distribution of Languages (2024), <https://commoncrawl.github.io/cc-crawl-statistics/plots/languages>
11. Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P., Berner, J.: Mathematical Capabilities of ChatGPT (2023), <https://ghosts.friederr.org>
12. Ghafar, Z.N., Sawalmeh, M.H.M.: The Influence of Effective Teaching on Elementary School Students: A Review Paper. *Journal of Learning and Development Studies* **3**, 10–14 (2023). <https://doi.org/10.32996/jlds>, <https://creativecommons.org/licenses/by/4.0/>
13. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation (2 2023), <http://arxiv.org/abs/2302.09210>

14. Kulik, J.A., Fletcher, J.D.: Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research* **86**, 42–78 (3 2016). <https://doi.org/10.3102/0034654315581420>
15. Lupogo, I.: Language of Instruction: A Challenge for Secondary Schools and Tertiary Institutions in Implementing VET in Tanzania. *Journal of Educational Policy and Entrepreneurial Research* **1**, 26–30 (2014), <http://www.iiste.org/Journals/index.php/JEPER>
16. Maginga, T.J., Kutuku, S.J., Hamza, H.M., Mulokozi, G.G., Nsenga, J.: MkulimaGPT: Equitable AI Use Via a Swahili Chatbot for Maize Farming System in Tanzania. *African Journal of Agriculture and Food Science* **7**, 172–188 (10 2024). <https://doi.org/10.52589/AJAFS-VERNTB5I>
17. Mapunda, G., Gibson, H.: On the suitability of Swahili for early schooling in remote rural Tanzania: do policy and practice align? *Journal of the British Academy* **10**, 141–168 (2022). <https://doi.org/10.5871/jba/010s4.141>
18. Mazana, M.Y., Montero, C.S., Casmir, R.O.: Assessing Students’ Performance in Mathematics in Tanzania: The Teacher’s Perspective. *International Electronic Journal of Mathematics Education* **15**, em0589 (2020). <https://doi.org/10.29333/iejme/7994>
19. National Bureau of Statistics: Tanzania in Figures (2022), <https://www.nbs.go.tz/index.php/en/tanzania-in-figures/891-tanzania-in-figures-2022>
20. Ndijuye, L.G.: Early learning attainments of children of naturalized citizens of refugee backgrounds in the sub-Saharan region: Evidence from Tanzania. *Child Development Perspectives* **17**, 67–73 (3 2023). <https://doi.org/10.1111/cdep.12479>
21. Petrov, A., Malfa, L., Torr, P.H.S., Bibi, A.: Language Model Tokenizers Introduce Unfairness Between Languages (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/74bb24dca8334adce292883b4b651eda-Paper-Conference.pdf
22. Plevris, V., Papazafeiropoulos, G., Rios, A.J.: Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI (Switzerland)* **4**, 949–969 (12 2023). <https://doi.org/10.3390/ai4040048>
23. Robinson, N.R., Ogayo, P., Mortensen, D.R., Neubig, G.: ChatGPT MT: Competitive for High- (but not Low-) Resource Languages (9 2023), <http://arxiv.org/abs/2309.07423>
24. Schmucker, R., Xia, M., Azaria, A., Mitchell, T.: Ruffle&Riley: Towards the Automated Induction of Conversational Tutoring Systems (2023), https://gaied.org/neurips2023/files/38/38_paper.pdf
25. Spreitzer, C., Straser, O., Zehetmeier, S., Maaß, K.: Mathematical Modelling Abilities of Artificial Intelligence Tools: The Case of ChatGPT. *Education Sciences* **14** (7 2024). <https://doi.org/10.3390/educsci14070698>
26. Telaumbanua, F., Silalahi, H.R., Simanjuntak, J., Siregar, S.R., Simangunsong, M.I., Panjaitan, S., Hutauruk, A.J.: Psychological Encouragement for High School Students in Preparation for Online Learning During Covid-19. *Middle European Scientific Bulletin* **14** (2021)
27. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc. (12 2022), <http://arxiv.org/abs/2201.11903>
28. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT (2 2023), <http://arxiv.org/abs/2302.11382>