ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/Ogheneruona-Maria-Esegbona-Isikeh/558

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 3 & Iss: 1

Research Article

Assessing Zero-Shot and Zero-Shot Chain-of-Thought Reasoning Abilities in JAMB Mathematics and Physics Exams: Do LLMs 'Know' JAMB?

Ogheneruona Maria Esegbona-Isikeh[1], Tewogbade Shakir Adeyemi[2], Oluwole Fagbohun[2], Patricia Chidubem Udorji[2], Grace Funmilayo Farayola[3] and Solalu Habeeb[2]

[1]Birmingham City University, UK

[2]GenAI Lab, Readrly Limited, UK

[3]University of Buckingham, UK

Citation: Esegbona-Isikeh OM, Adeyemi TS, Fagbohun O, et al. Assessing Zero-Shot and Zero-Shot Chain-of-Thought Reasoning Abilities in JAMB Mathematics and Physics Exams: Do LLMs 'Know' JAMB? *J Artif Intell Mach Learn & Data Sci 2025* 3(1), 2610-2616. DOI: doi.org/10.51219/JAIMLD/Ogheneruona-Maria-Esegbona-Isikeh/558

Received: 25 April, 2025; **Accepted:** 12 May, 2025; **Published:** 14 May, 2025

***Corresponding author:** Ogheneruona Maria Esegbona-Isikeh, Birmingham City University, England, UK

ABSTRACT

In this study, we investigate the zero-shot and zero-shot chain-of-thought reasoning capabilities of advanced language models GPT-4, Claude and Mistral on the Joint Admissions and Matriculation Board (JAMB) Mathematics and Physics examinations. The evaluation focuses on the models' inherent ability to solve standardised test questions that require logical reasoning and domain-specific knowledge, without any prior fine-tuning. Past JAMB exam questions were systematically presented to each model in both zero-shot and chain-of-thought prompting conditions. We analysed performance using metrics such as accuracy, reasoning quality, response time and computational efficiency to draw comparisons across models and evaluation methods.

The findings reveal the strengths and limitations of each model in tackling complex problem-solving tasks, with particular emphasis on the impact of chain-of-thought prompting on reasoning performance. These results provide valuable insights into the potential of large language models in educational contexts, particularly in the development of automated tutoring and assessment tools. The study also identifies areas where current models perform well and where further improvement is needed, offering guidance for future research and AI system design tailored to Mathematics and Physics education within the Nigerian curriculum.

Keywords: Zero-shot learning, Chain-of-thought reasoning, GPT-4, Language models, JAMB exams, Mathematics education, Physics education, AI in education, Comparative study, Computational efficiency

1. Introduction

The field of artificial intelligence has witnessed remarkable advancements in recent years, particularly with the development of large language models (LLMs) such as GPT-4, Claude and Mistral. These models have demonstrated unprecedented capabilities in natural language understanding and generation, achieving human-like proficiency in tasks ranging from language translation to creative writing. For instance, OpenAI's GPT-3 and its successor GPT-4 have shown that scaling up model parameters and training data leads to significant improvements in performance across a variety of benchmarks[1,2]. Similarly, models like Claude have been optimized for conversational tasks, further highlighting the versatility of LLMs in different contexts[3].

Despite these advancements, the application of LLMs in specialized domains such as education remains an area ripe for exploration. Solving standardized test problems, which often require domain-specific knowledge and logical reasoning, poses a unique challenge for these models. Prior research has shown that while LLMs can perform well on general language tasks, their effectiveness diminishes when confronted with complex problem-solving scenarios that necessitate step-by-step reasoning[4]. This gap underscores the need to evaluate LLMs in educational settings, particularly in subjects like Mathematics and Physics, where problem-solving skills are paramount.

The Joint Admissions and Matriculation Board (JAMB) exams are critical standardized tests in Nigeria that assess students' readiness for tertiary education. These exams cover a broad spectrum of topics in Mathematics and Physics, demanding not just rote memorization but also a deep understanding of concepts and the ability to apply them in problem-solving contexts. Given that over 1.9 million candidates registered for the JAMB exams in 2020 alone[5], the impact of enhancing educational tools and resources for this examination is substantial.

Evaluating LLMs on JAMB exam questions offers valuable insights into their potential applications in educational contexts within the Nigerian framework. Such an evaluation can help determine whether these models possess the inherent capability to understand and solve complex, domain-specific problems without prior fine-tuning. Moreover, it can shed light on how techniques like zero-shot learning and chain-of-thought prompting influence the reasoning processes of LLMs. Zero-shot learning enables models to make predictions about data they have not been explicitly trained on, leveraging their generalization capabilities[6]. When combined with chain-of-thought prompting-a technique that encourages models to generate intermediate reasoning steps-the potential for enhanced problem-solving emerges[4]. This approach aligns with cognitive theories of learning, which emphasize the importance of step-by-step reasoning in understanding complex problems[7].

In this paper, we aim to assess the zero-shot and zero-shot chain-of-thought reasoning abilities of selected LLMs on JAMB Mathematics and Physics exams. By conducting a comparative analysis of GPT-4, Claude and Mistral, we seek to understand:

- How well these models can handle domain-specific, standardized test questions without prior fine-tuning.
- The impact of chain-of-thought prompting on their reasoning processes.
- The implications for deploying LLMs in educational tools and assessments.

Our contributions are threefold:

- **Comparative analysis:** We provide a comprehensive evaluation of four advanced LLMs on JAMB exam questions, highlighting their strengths and weaknesses in a standardized testing context.
- **Effectiveness of chain-of-thought prompting:** We assess how chain-of-thought prompting influences the problem-solving abilities of LLMs, contributing to the understanding of how reasoning processes can be enhanced in these models.
- **Implications for educational settings:** We discuss the potential applications and limitations of using LLMs in educational environments within Nigeria, offering insights for future developments in AI-assisted learning tools.

By addressing these points, we hope to contribute to the growing body of research on the intersection of artificial intelligence and education. Understanding how LLMs perform in educational assessments not only informs the development of more effective AI models but also paves the way for innovative educational technologies that can support students in their learning journeys.

2. Related Work

2.1. Large language models in education

The integration of large language models (LLMs) into educational settings has been a growing area of interest. LLMs like GPT-3[1] and GPT-4[2] have demonstrated impressive capabilities in generating coherent and contextually relevant text, which has significant implications for education. These models have been explored for applications such as automated tutoring, content generation and personalized learning support.

For instance, Hu, et al, investigated the use of LLMs for generating educational content[8], finding that these models could create practice questions and explanations that align with curricular standards. Similarly, examined how GPT-4 could support students in problem-solving by providing hints and feedback in real-time, enhancing the learning experience[9].

However, challenges persist in ensuring the reliability and accuracy of LLM-generated content. Issues such as factual errors, potential biases and the alignment of generated material with educational objectives are critical concerns[10]. Researchers emphasize the need for careful validation and oversight when integrating LLMs into educational tools to prevent the dissemination of misleading information.

2.2. Zero-shot learning and chain-of-thought reasoning

Zero-shot learning enables models to perform tasks without explicit task-specific training by leveraging their ability to generalize from existing knowledge[6]. In the context of LLMs, this means that models can respond to prompts about unfamiliar tasks using their broad language understanding.

Chain-of-thought prompting is a technique that encourages models to generate intermediate reasoning steps before arriving at a final answer[4]. By articulating the reasoning process, models can improve their performance on complex tasks that require logical progression and multi-step problem-solving.

Kojima, et al, demonstrated that zero-shot chain-of-thought prompting significantly enhances the reasoning abilities of LLMs on arithmetic and commonsense reasoning tasks[11]. This approach allows models to break down problems into manageable steps, leading to more accurate and explainable outcomes.

Moreover, Fung, Wong and Tan, explored the use of chain-of-thought reasoning in mathematical problem-solving[12], finding that it not only improved accuracy but also provided insights into the model's reasoning process.

Such transparency is valuable in educational settings, where understanding the steps leading to a solution is as important as the solution itself.

2.3. Evaluations on standardized tests

Evaluating LLMs on standardized tests serves as a benchmark for their reasoning and problem-solving capabilities.

Prior studies have assessed models on exams like the SAT, GRE and LSAT to gauge their performance in academic contexts.

For example, OpenAI's GPT-3 was evaluated on a range of standardized tests, revealing that while the model excelled in language comprehension and vocabulary, it struggled with quantitative reasoning sections[13]. Clark et, al, assessed LLMs on science questions from standardized tests, finding that models performed well on questions requiring factual recall but faced challenges with those necessitating complex reasoning[14].

In mathematics, the MATH dataset introduced by Hendrycks, et al, provided a collection of problems from high school competitions to test the mathematical reasoning of LLMs[15]. Results indicated that models benefited from explicit reasoning steps, aligning with the benefits observed from chain-of-thought prompting.

Despite these evaluations, there is a notable gap in assessing LLMs on standardized tests from non-Western educational systems. The majority of research has focused on exams prevalent in the United States and Europe, overlooking assessments like the JAMB exams in Nigeria. This gap is significant given the linguistic and cultural differences that can affect model performance[16].

2.4. LLMs and african educational assessments

Research on applying LLMs to African educational contexts, particularly in standardized testing, is limited. The diversity of languages and educational curricula across African nations presents unique challenges for NLP applications[17].

Adelani DI, highlighted the scarcity of NLP resources for African languages and the need for models that can understand and process local educational content[18]. They emphasized the importance of developing AI systems that are inclusive and representative of the linguistic diversity in Africa.

Evaluating LLMs on the JAMB exams addresses this gap by providing insights into how these models perform on assessments critical to Nigerian students. Understanding their capabilities and limitations can inform the development of educational tools tailored to the Nigerian context, supporting students in subjects like Mathematics and Physics where problem-solving skills are crucial.

3. Methodology

3.1. Models evaluated

In this study, we evaluated three advanced large language models (LLMs):

- **GPT-4:** Developed by OpenAI, GPT-4 is known for its strong reasoning abilities and has demonstrated significant improvements over its predecessors in various natural language processing tasks[2].
- **Claude:** An AI assistant developed by Anthropic, designed to excel in conversational tasks while maintaining coherent and contextually relevant responses[3].
- **Mistral:** An open-source model recognized for its computational efficiency and effectiveness in tasks requiring language comprehension1[9].

3.2. Dataset

We curated a representative sample of past Joint Admissions and Matriculation Board (JAMB) Mathematics and Physics examination questions. The dataset encompassed a range of topics, including algebra, calculus, mechanics and electromagnetism. To ensure diversity and relevance, we selected problems from multiple years and included various difficulty levels. The JAMB exams are standardized tests administered in Nigeria to assess students' readiness for tertiary education. These exams are rigorous and require not only subject knowledge but also critical thinking and problem-solving skills[5].

3.3. Evaluation settings

We assessed the models under two evaluation settings:

- **Zero-shot:** Models received questions without any additional context or prior examples, testing their inherent ability to generate correct answers based solely on their pre-trained knowledge[1].
- **Zero-Shot chain-of-thought (CoT):** Models were prompted to provide step-by-step reasoning before arriving at the final answer. This approach aims to enhance the models' problem-solving capabilities by encouraging them to articulate intermediate reasoning steps[4].

3.4. Metrics

We evaluated the models using the following metrics:

- **Accuracy:** The percentage of questions for which the model provided the correct final answer.
- **Reasoning quality:** Assessed based on the coherence, logical progression and correctness of the reasoning steps provided by the model.
- **Response time:** The time taken by the model to generate an answer, measured in seconds.
- **Computational efficiency:** Evaluated by monitoring resource utilization during inference, including CPU/GPU usage and memory consumption.

3.5. Procedure

- **Data preprocessing:** We formatted each question to ensure compatibility with the input requirements of each model. Questions involving diagrams or visual components were excluded due to input limitations. All textual content was carefully proofread for clarity and consistency.
- **Question presentation:** For the zero-shot setting, each model was presented with questions in plain text. For the zero-shot CoT setting, we prefixed each question with a prompt encouraging step-by-step reasoning (e.g., "Solve step by step:").
- **Model interaction:** We utilized the API or interface provided by each model to submit the questions and receive responses. Consistent settings were used across models to ensure a fair comparison.
- **Response collection:** The models' answers and, where applicable, their reasoning steps were recorded. We captured both the final answer and any intermediate reasoning provided.
- **Analysis:** Responses were evaluated against the official answer keys provided by JAMB. For reasoning quality, we developed a rubric to assess the logical coherence and correctness of each step. Two independent reviewers scored the reasoning to ensure objectivity.

- **Ethical considerations:** We adhered to ethical guidelines for AI research, ensuring that data privacy and intellectual property rights were respected[20].

4. Experiments

4.1. Zero-shot evaluation

In the zero-shot setting, we directly posed questions to the models without any additional context or instruction to elaborate on their reasoning **(Figure 1)**. For example:

Question:
Calculate the derivative of $y=3x^2+2x-5$

Expected Answer:
$y| = 6x+2$

This setting tests each model's ability to recall and apply mathematical rules solely from their pre-trained knowledge, without step-by-step guidance.

4.2. Zero-shot chain-of-thought evaluation

In the zero-shot chain-of-thought (CoT) evaluation, models were prompted to solve problems by providing step-by-step reasoning **(Figure 2)**. For example:

Prompt:
Solve step by step: Calculate the derivative of y= 3x2 + 2x -5
Expected Reasoning:
Step-1: The derivative of 3x2 is 6x.
Step-2: The derivative of 2x is 2.
Step-3: The derivative of the constant -5 is 0.
Step-4: Therefore, y' = 6x + 2

Prompting models to show their working enables a more detailed assessment of how structured reasoning impacts their mathematical accuracy and logic[11]. Table 2 illustrates the differing reasoning styles of GPT-4, Claude and Mistral.

4.3. Data preprocessing

To ensure the integrity of the evaluation:

- **Exclusion of diagram-based questions:** Questions requiring visual interpretation were omitted due to the models' inability to process images in this context.
- **Clarity and consistency:** All questions were standardized in terms of notation and language to prevent any ambiguity that could affect the models' understanding.
- **Validation:** A subject matter expert reviewed the dataset to confirm the accuracy of the questions and expected answers.

4.4. Experimental setup

- **Hardware:** The experiments were conducted on a system equipped with NVIDIA Tesla V100 GPUs and 256 GB RAM to accommodate the computational requirements of the models.
- **Software:** We used the latest versions of the models' APIs as of October 2023. All models were accessed in their default configurations without any fine-tuning or additional training.
- **Randomization:** The order of questions was randomized for each model to mitigate any potential ordering effects.

4.5. Statistical analysis

We performed statistical analyses to determine the significance of differences in performance between models and evaluation settings:

- **Accuracy comparison:** Chi-squared tests were used to compare accuracy rates between models.
- **Response time analysis:** ANOVA tests were conducted to compare mean response times across models.
- **Inter-rater reliability:** Cohen's kappa coefficient was calculated to assess the agreement between reviewers on reasoning quality scores[21].

5. Results

5.1. Accuracy

The evaluation of the models revealed significant differences in their ability to correctly answer JAMB Mathematics and Physics exam questions under both zero-shot and zero-shot chain-of-thought (CoT) settings. **(Table 1)** summarizes the accuracy percentages achieved by each model.

Table 1: Accuracy of Models in Zero-Shot and Zero-Shot Chain-of-Thought (CoT) Settings.

Model	Zero-Shot (%)	Zero-Shot COT (%)
GPT-4	67	82
Claude	61	77
Mistral	43	58

In the zero-shot setting, GPT-4 achieved the highest accuracy at 67%, indicating its superior ability to handle domain-specific questions without prior fine-tuning. Claude followed with 61% and Mistral with 43%. When employing chain-of-thought prompting, all models exhibited improved performance. GPT-4's accuracy increased to 82%, highlighting a 15% improvement, consistent with findings by Wei et al. (2022) that chain-of-thought reasoning enhances model performance on complex tasks.

Statistical analysis using chi-squared tests confirmed that the accuracy improvements with chain-of-thought prompting were significant for all models ($p < 0.01$). These results suggest that guiding models to articulate their reasoning processes can effectively enhance their problem-solving abilities.

5.2. Reasoning quality

The reasoning quality was assessed based on coherence, logical progression and correctness of the steps provided. GPT-4 consistently exhibited clear and logical reasoning that closely mirrored human problem-solving approaches. Its step-by-step explanations were detailed and accurate, aligning with observations in prior studies on GPT-4's reasoning capabilities[2].

Claude delivered coherent reasoning but occasionally oversimplified explanations, potentially omitting critical intermediate steps. Mistral's reasoning was often fragmented, with frequent errors in calculations and logical inconsistencies. Inter-rater reliability for reasoning quality assessment was high (Cohen's kappa = 0.85), indicating strong agreement between reviewers **(Table 2)**.

As observed, GPT-4 provides a clearer and more structured step-by-step explanation, which aligns better with human reasoning processes. Claude and Mistral offer correct final answers but demonstrate less comprehensive intermediate reasoning.

Table 2: Sample Reasoning Quality Comparison in the Zero-Shot Chain-of-Thought Setting.

Model	Response Excerpt
GPT-4	"1. The derivative of $3x^2$ is 62. 2. The derivative of 2x is 2. 3. The derivative of -5 is 0. 4. Summing up, y' prime = 6x +2:
Claude	"First, find the derivative of each term. The derivative is y' = 6x + 2 "
Mistral	"Derivative is y' = 3(2x) + 2 So y' = 6x + 2 "

5.3. Response time

The average response time for each model was measured to assess efficiency. **(Table 3)** presents the findings.

Table 3: Average Response Time of Models.

Model	Average Response Time (s)
GPT-4	4.8
Claude	5.5
Mistral	3.6

Mistral demonstrated the fastest response time at an average of 3.6 seconds, attributable to its optimization for computational efficiency (Mistral AI, 2023). GPT-4 had moderate response times, while Claude had the longest average response time at 5.5 seconds, possibly due to its design for in-depth conversational engagement. An ANOVA test indicated that the differences in response times were statistically significant ($p < 0.05$). However, all models responded within a time frame suitable for real-time applications in educational settings.

5.4. Computational efficiency

Computational efficiency was evaluated based on resource utilization during inference, considering factors such as CPU/GPU usage and memory consumption.

- **Mistral:** Demonstrated the highest computational efficiency, utilizing fewer computational resources due to its smaller model size and optimized architecture. This efficiency makes it suitable for deployment in resource-constrained environments[19].
- **GPT-4 and claude:** Exhibited higher resource consumption, which is consistent with their larger parameter sizes and more complex architectures. While they offer superior performance in accuracy and reasoning quality, their computational demands may pose challenges for scalability and accessibility, particularly in regions with limited computational infrastructure.

The trade-off between performance and computational efficiency aligns with observations in prior research, where larger models often require more resources but deliver enhanced capabilities[22].

6. Discussion

6.1. Effect of chain-of-thought prompting

Our findings indicate that chain-of-thought prompting significantly enhances the problem-solving performance of large language models (LLMs) on standardized test questions. Across all models evaluated, there was an average accuracy increase of 15% when chain-of-thought prompting was employed. This improvement aligns with prior research demonstrating that guiding LLMs to articulate intermediate reasoning steps can lead to better outcomes on complex tasks[4].

The enhancement is likely due to the reduction of ambiguity and the provision of a structured framework for the models to follow. By encouraging step-by-step reasoning, the models can better navigate the problem space and avoid heuristic shortcuts that might lead to incorrect answers[11]. This approach mirrors human cognitive strategies, where explicit reasoning aids in understanding and solving complex problems[7].

6.2. Model performance analysis

Among the models evaluated, GPT-4 exhibited the highest accuracy and reasoning quality. Its superior performance can be attributed to its extensive training data and advanced architecture, enabling it to capture nuanced patterns and perform sophisticated reasoning[2]. Claude showed competent reasoning abilities but was slightly less accurate than GPT-4, possibly due to differences in training objectives or data.

Mistral, while notable for its computational efficiency, was limited by its smaller model size, which impacted its ability to handle complex reasoning tasks. This limitation is consistent with established scaling laws, where larger models tend to perform better on a variety of tasks due to their increased capacity to learn and represent information.

6.3. Limitations

Several limitations of our study should be acknowledged. First, the exclusion of diagram-based questions may have skewed the results, as visual information is integral to many Mathematics and Physics problems. Future research should incorporate multimodal models capable of processing both text and images to provide a more comprehensive evaluation[23]. Second, cultural and linguistic nuances specific to the Nigerian educational context may not have been fully captured by the models, leading to occasional misinterpretations of questions. This issue highlights the importance of training models on localized data to improve their relevance and accuracy in specific contexts[16].

Third, computational efficiency varied among the models, impacting their scalability in real-world applications. Models like GPT-4, despite their high performance, require substantial computational resources, which may not be accessible in all educational settings, particularly in resource-constrained environments[24].

6.4. Implications for AI in education

The enhanced performance observed with chain-of-thought prompting suggests that LLMs have significant potential as educational tools. They can assist students by providing detailed explanations and step-by-step solutions, facilitating personalized learning experiences[9]. However, the inconsistencies and inaccuracies identified necessitate cautious implementation.

Human oversight remains essential to ensure the accuracy and appropriateness of the content generated by LLMs. Educators should be involved in the deployment process to validate information and guide the integration of these models into curricula[25]. Additionally, ethical considerations such as fairness, accessibility and bias mitigation must be addressed to prevent exacerbating existing educational inequalities[10].

6.5. Error analysis

We conducted a detailed qualitative analysis of model

errors across both the zero-shot and zero-shot chain-of-thought (CoT) prompting settings to better understand the nature and distribution of reasoning failures.

6.5.1. Zero-shot setting: In the zero-shot condition, the most prevalent errors stemmed from the absence of explicit reasoning scaffolds. Common issues included:

- **Misinterpretation of problem phrasing:** Models often struggled with question wording that relied on implicit assumptions or culturally specific syntax. For instance, Mistral misread "find the least number that satisfies..." as a maximum-finding task, leading to incorrect conclusions.
- **Omission of critical formulae:** GPT-4 and Claude sometimes failed to invoke standard formulae in Mechanics or Algebra (e.g., ignoring Newton's Second Law or the quadratic formula), instead offering approximations or surface-level logic.
- **Incorrect application of units:** All models occasionally mismatched units, such as treating km/h and m/s interchangeably, especially in Physics problems involving motion or force.

These errors highlight the limitations of relying solely on internalised pretraining when solving domain-specific tasks that require formal precision.

6.5.2. Zero-shot chain-of-thought (CoT) setting: In the CoT prompting condition, the inclusion of structured reasoning led to improved performance but also introduced new categories of errors:

- **Hallucinated intermediate steps:** GPT-4 occasionally inserted plausible but incorrect steps (e.g., fabricating intermediary values for resistance or acceleration), reflecting a tendency to fill in gaps with confident but ungrounded assertions.
- **Arithmetic miscalculations:** Claude and Mistral, in particular, frequently performed flawed arithmetic despite logically sound initial reasoning. For example, a multi-step ratio problem with correct logic broke down due to a multiplication slip in the final step.
- **Truncated logic chains:** Mistral often failed to complete multi-step reasoning sequences. In one instance, it began solving a kinematics question with the right approach but stopped prematurely before deriving the final velocity.

These findings suggest that while CoT prompting generally strengthens logical structure and interpretability, it also increases the cognitive load on the model, raising the likelihood of internal inconsistency or overconfident fabrication.

6.6. Socio-educational impact of LLMs in high-stakes exam preparation

The deployment of large language models (LLMs) in tools designed to support high-stakes exam preparation offers a promising avenue for educational transformation. In Nigeria, where approximately 1.9 million students sit the JAMB examination annually, access to high-quality academic support is deeply uneven across geographic and socioeconomic lines.

LLMs, when embedded in scalable and accessible platforms-including mobile apps or USSD-based tools-have the potential to democratise learning by providing instant, personalised instruction. These systems can deliver detailed explanations, step-by-step problem-solving walkthroughs and real-time feedback, addressing gaps in regions where qualified teachers are scarce, overburdened or entirely unavailable.

This potential, however, is not without caveats. Algorithmic fairness must be carefully addressed to prevent the amplification of existing educational biases. LLMs can sometimes hallucinate explanations or provide inaccurate reasoning paths, posing a risk in high-stakes learning environments. Moreover, the digital divide presents a fundamental barrier: students lacking access to internet connectivity, smartphones or laptops may be excluded unless deliberate design efforts are made.

To realise the benefits of LLM-powered exam preparation tools while mitigating risks, we recommend:

- Designing low-bandwidth, offline-capable solutions that reduce dependence on continuous internet access.
- Building hybrid human-AI systems, where teachers co-pilot or moderate AI-generated feedback to maintain instructional integrity.
- Localising content, ensuring alignment with Nigerian curriculum standards, linguistic norms and cultural framing.

When thoughtfully deployed with an emphasis on inclusion, transparency and pedagogical alignment, LLMs can enhance learning resilience, reduce reliance on costly private tutoring and foster more equitable outcomes for students across Nigeria and the broader African educational landscape.

7. Conclusion

This study assessed the zero-shot and zero-shot chain-of-thought reasoning abilities of GPT-4, Claude and Mistral on JAMB Mathematics and Physics exams. The findings reveal that chain-of-thought prompting significantly improves model performance, with GPT-4 demonstrating the highest proficiency. These results underscore the potential of LLMs to tackle complex, domain-specific tasks without prior fine-tuning.

The implications of our research are twofold. First, LLMs hold promise for augmenting learning experiences by providing students with tailored assistance in understanding complex subjects. However, they require further refinement to ensure reliability and accuracy. Second, deploying LLMs in educational contexts must consider policy and ethical aspects, including fairness, accessibility and bias mitigation, to ensure equitable benefits across diverse student populations[20].

Note: These experiments were conducted in September 2024. Since then, the evaluated models may have undergone further updates and improvements, which could affect their current performance.

8. References

1. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020;33: 1877-1901.
2. Open AI. GPT-4 Technical Report. OpenAI Documentation, 2023.
3. Anthropic. Claude: An AI Assistant for Conversational Tasks, 2023.
4. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought Prompting Elicits Reasoning in Large Language Models, 2022.

5. JAMB. JAMB Registration Statistics. Joint Admissions and Matriculation Board, 2020.

6. Xian Y, Schiele B, Akata Z. Zero-Shot Learning-The Good, the Bad and the Ugly. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4582-4591.

7. Sweller J, Ayres P, Kalyuga S. Cognitive Load Theory. Springer, 2011.

8. Hu B, Zheng L, Zhu J, et al. Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design. IEEE Transactions on Learning Technologies, 2024.

9. Chohan I, Khan I. Enhancing Mathematics Education with ChatGPT-4 Personalized Problem-Solving and Consistent Learning. In 2024 2nd International Conference on Computing and Data Analytics (ICCDA), 2024: 1-6.

10. Bender EM, Gebru T, McMillan-Major A, et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency, 2021: 610-623.

11. Kojima T, Gu SS, Reid M, et al. Large Language Models are Zero-Shot Reasoners, 2022.

12. Fung SCE, Wong MF, Tan CW. Chain-of-Thoughts Prompting with Language Models for Accurate Math Problem-Solving. 2023 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2023: 1-5.

13. OpenAI. GPT-3: Language Models are Few-Shot Learners. OpenAI Blog, 2020.

14. Clark P, Tafjord O, Richardson K, et al. From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project, 2020.

15. Hendrycks D, Burns C, Basart S, et al. Measuring Mathematical Problem Solving with the MATH Dataset. Advances in Neural Information Processing Systems, 2021;34: 24293-24307.

16. Joshi P, Santy S, Budhiraja A, et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6282-6293.

17. Nekoto W, Marivate V, Matsila T, et al. Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. Findings of the Association for Computational Linguistics: EMNLP, 2020: 2144-2160.

18. Adelani DI. Natural language processing for African languages, 2022.

19. Mistral AI. Mistral: An Open-Source Language Model, 2023.

20. Floridi L, Cowls J. A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 2022;1.

21. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1960;20: 37-46.

22. Kaplan J, McCandlish S, Henighan T, et al. Scaling Laws for Neural Language Models, 2020.

23. Lu J, Batra D, Parikh D, Lee S. ViLBERT: Pretraining Task-Agnostic Visio linguistic Representations for Vision-and-Language Tasks. Advances in Neural Information Processing Systems, 2019;32: 13-23.

24. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3645-3650.

25. Holstein K, McLaren BM, Aleven V. Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. Proceedings of the AAAI Conference on Artificial Intelligence, 2019;33: 5310-5317.