When the robotic Maths tutor is wrong - can children identify mistakes generated by ChatGPT?

Manal Helal*, Patrick Holthaus, Luke Wood, Vignesh Velmurugan,
Gabriella Lakatos, Sílvia Moros, and Farshid Amirabdollahian
School of Physics, Engineering and Computer Science
University of Hertfordshire
College Lane, Hatfield AL10 9AB, UK
Email: {m.helal, p.holthaus, l.wood, v.velmurugan, g.lakatos, s.moros2, f.amirabdollahian2}@herts.ac.uk

***Abstract*—This study delves into integrating Large Language Models (LLMs), particularly ChatGPT-powered robots, as educational tools in primary school mathematics. Against the backdrop of Artificial Intelligence (AI) increasingly permeating educational settings, our investigation focuses on the response of young learners to errors made by these LLM-powered robots. Employing a user study approach, we conducted an experiment using the Pepper robot in a primary school classroom environment, where 77 primary school students from multiple grades (Year 3 to 5) took part in interacting with the robot. Our statistically significant findings highlight that most students, regardless of the year group, could discern between correct and incorrect responses generated by the robots, demonstrating a promising level of understanding and engagement with the AI-driven educational tool. Additionally, we observed that students' correctness in answering the Maths questions significantly influenced their ability to identify errors, underscoring the importance of prior knowledge in verifying LLM responses and detecting errors. Additionally, we examined potential confounding factors such as age and gender. Our findings underscore the importance of gradually integrating AI-powered educational tools under the guidance of domain experts following thorough verification processes. Moreover, our study calls for further research to establish best practices for implementing AI-driven pedagogical approaches in educational settings.**

***Index Terms*—Large Language Models, LLM Mathematical Correctness, Educational Robots, Cognition, Social Robotics**

I. Introduction

In the AI era, information is more accessible than in previous generations. However, accessing information might not lead to learning. Learning requires understanding how this information came together and contributed to cognition development. Learning from easily available information provided by AI applications can limit natural intelligence development and memory retention [1]. Education has a long-term impact on the quality of life of individuals and societies. This is why introducing new approaches without evaluating their long-term impact on the cognition development of students might have irrecoverable consequences that can be corrected only in the following generations. Integrating new approaches needs to be done while measuring impacts thoroughly and reevaluating when needed.

Generally, LLMs have been used widely in education [2], [3]. The consensus is that LLMs can be used to benefit students, but it is important to note that they are prone to a variety of errors [4]. Research looking at AI or LLMs in education has focused on the following aspects. One aspect concerns errors produced by a LLM that must be verified before acceptance [5]–[7]. Some other studies evaluated methods to detect AI generated vs Human content such as Writer, Copyleaks, GPTZero, and CrossPlag [8]. Some studies raised questions of integrity when AI authors students' submissions or academic manuscripts for scientific publications or academic and educational gains [9]. The downside of using AI for gaining such merits could be that the AI skills are being evaluated and not the long-term student's skills development [10]. Most previous work focuses on university students or textbooks, whether the LLM responses are identified as AI-generated or human-generated, and whether this enhances students' skills or not. Some publications report that the submissions are enhanced but do not focus on student skills [11]. ChatGPT, as a famous LLM example, showed outstanding performance enhancement in supporting students in economics, satisfactory in programming, and unsatisfactory in mathematics [12]. While other work has illustrated various shortcomings in LLM-assisted learning of computer science without prior knowledge [13]. The rush to AI-answers in various tasks at all ages and disciplines has created concerns over creativity loss and negative impacts on learning and natural intelligence development and mastery of skills [14]–[16].

The "AI for Kids" proposal emphasizes the importance of exposing children to AI literacy as part of digital literacy for intelligent societies [17]. The proposal presents the why, what and how to educate 3:8-year-old children about AI. Other studies discussed further challenges and opportunities in AI in early childhood education [18]. The World Economic Forum estimates that 65% of today's children will be working in jobs not yet created. This necessitates that AI education start early with children. The AI-powered education investments are estimated to grow to $20 billion by 2027 from $4 billion in 2022. Privacy and safety concerns over the children's excessive use of AI-enabled technologies in phone and smart home systems and others have grown consequently [19]. Concerns also over AI-powered education's various risks are expressed

at the AI+Education Summit. One of the risks discussed was bias in the training dataset of the AI models, which did not reflect cultural diversity and affected the hoped-for personalised learning. Another possible risk is that AI could harm students' critical thinking development, lack of explanation and lack of proper pedagogy, affecting learning how information is generated as opposed to making this information quickly available. The most relevant risk is the errors and incorrect information in the LLMs' responses are well packaged and sometimes difficult to detect [20].

Robots also have been frequently used in educational contexts [21], e.g. to support children with autism in their learning [22]. AI-powered robots' role in education and the learning impact has been evaluated in a systematic review [23]. The survey evaluated the different studies in which robots have been used as tutees, tutors, and tools to assist learners in various educational stages, from primary education to higher education and in different subjects such as languages, math and science, and STEM. The review provided various future recommendations, such as the need for long-term studies. AI provided robots with computer vision, voice recognition, and natural language understanding. Other work discussed different models in which LLM-powered Robots are used in education and other applications [24].

For these reasons, we evaluated primary school children's abilities in detecting errors generated by robots empowered by LLMs, such as ChatGPT, when answering common questions from the maths curriculum. When errors are not detected, children could be misled/wrongly educated. We also investigated the confounding factors affecting children's ability to detect LLM errors, particularly students' correctness in answering the questions, age and gender. The presented experiment is one of the very few experiments using LLM-powered robots in the education of primary school students that is student-focused. Another primary school experiment is teachers-focused [25]. In particular, we investigated if LLM-powered robots could be used as maths tutors in primary schools and how the children would respond in case of mistakes.

The following section, section II, will discuss some of the related work that considers robots in education as well as the benefits and drawbacks of using LLMs. Section III then describes the methods we applied in setting up the experiment to answer the questions above, how we collected the dataset, and analysed it. The results from this experiment are explained in section IV, followed by discussions in section V, which puts our findings into the wider context of using AI in education. We further provide some limitations and future work of our work in section VI and draw a conclusion in section VII.

II. Related work

In this section, we will introduce related work. The first subsection introduces the use of AI-powered Robots in education, and the following subsection focuses on the LLMs error types.

A. Robots and AI in educational settings

Since the 2010s, the rise of AI in education attracted substantial attention. AI technologies, including machine learning, natural language processing, and computer vision, made significant inroads into education. AI-powered applications began offering personalised learning experiences [26], intelligent tutoring systems [27], and automated grading [28]. The advent of advanced LLMs, such as OpenAI's Generative Pre-trained Transformer (GPT) series, marked a shift towards natural language understanding and generation. Educators started exploring using LLMs to create educational content, generate quizzes and essay scoring, and provide instant feedback. Virtual assistants and AI-powered chatbots have been integrated into educational platforms for instant student support. These systems can answer queries, offer guidance, and facilitate interactive learning experiences. As AI and LLMs become more prevalent in education, concerns about data privacy, bias in algorithms, and the ethical use of technology have surfaced. Striking a balance between automation and maintaining a human touch in education remains an ongoing challenge.

Educators and learners find it easy to use LLMs to retrieve answers faster than reading books and traditional references, and they use publicly available portals without fine-tuning. Cognitive abilities are linked to educational approaches affected by learners' genetic variations and affect long-term health, wealth, morbidity and mortality [1]. Many studies identified successful approaches in education to improve society's welfare in the AI era. Brain-based learning activities such as practical activities, problem-solving exercises, multisensory engagement, and discussions promote active learning and profoundly impact students' engagement, knowledge retention, and critical thinking abilities. Time to reflect on learned knowledge connects to prior knowledge and creates meta-cognitive skills.

Artificial Intelligence in Education (AIED) has enabled personalised education experiences to be tailored to different students' genetic and educational experiences. Models such as museum brain-based instruction, virtual reality, asynchronous discussion groups, learning analytics, recommendation systems, adaptive learning, adaptive mastering tests, dynamic assessment, intelligent tutoring systems, real-time feedback, and targeted interventions are identified in the literature to support cognitive development better. LLMs such as GPTs and other models managed to process and generate massive amounts of text after training on the massive corpus from internet-wide content and other sources. The training dataset bias caused ethical concerns about the validity of the responses. However, fine-tuning these models into specific curriculum content can enhance the responses personalised to different educational contexts [29].

B. Large language models and accuracy

LLMs such as Google Fine-tuned Language Net (FLAN) or Pathways Language Model (PaLM), Facebook Open Pre-trained Transformers (OPT), Microsoft BARD and OpenAI Generative Pre-trained Transformers (GPT) have managed to

process and generate large amounts of text after training on the very large corpus from the internet-wide content and other sources [30]. To reach this unprecedented higher level of accuracy, these models have billions of parameters and are trained on many petaflops-days of computation. Being trained on unverified datasets on diverse domains, LLMs are known to produce various types of errors, with the most notable being hallucination and factual errors. Other errors include repetition, grammatical errors, semantic incoherence, out-of-domain responses, biased responses, undergeneration, overgeneration and uncertainty. Hallucination refers to the generation of text that is not grounded in reality or factual information. This could involve the creation of entirely fictional events, entities, or statements that do not exist in the real world. Factual errors, or factuality, refer to inaccuracies in the information the LLM presents. These errors occur when the model generates text that contradicts established facts, empirical evidence, or common knowledge. Factual errors can range from minor inaccuracies to significant distortions of reality. The work in [5] emphasises the importance of verifying the LLM erroneous responses that are labelled as hallucinations. The authors identified physics, chemistry, computer science and mathematics LLM errors taken from university-level textbooks and exams. The verification process can include private hosting to eliminate breach of privacy of use history, fine-tuning to eliminate irrelevant responses, and changing response vocabulary to eliminate negative emotions building up on learners and blocking their cognitive development.

Errors in LLM retrieved information in various domains can cause safety issues or even life-threatening issues in Healthcare, loss or harm in economics and legal domains, and misinformation and deteriorated mastery and negative cognition development in education domains. The work in [6] identified factuality problems in the various domains rather than the explicit hallucinations. This extensive review identified the factuality of information quantitatively using various metrics, causes of non-factual errors at the various levels of training, inference or retrieval and approaches to enhance the factuality of LLMs such as fine-tuning to specific knowledge bases, multi-agents, decoding approaches, interactive retrieval among others. While fine-tuning these models into specific curriculum content can enhance the responses personalised to different educational contexts [29], choice of training dataset and other biases can still negatively impact the validity of the responses and might introduce other mistakes [7].

III. Methods

Given the current state-of-the-art investigations of LLM-powered robots in education described in section II, we explore the ability of primary school students to trust their knowledge when faced with errors introduced by LLM responses. We focused on a fundamental knowledge area, specifically the UK Maths Year 3 curriculum, where maths reasoning plays a crucial role. Lack of mathematical reasoning can lead to various issues in different domains, such as providing inaccurate financial advice, misinterpretation of medical statistics affecting treatment decisions and/or dosage needs, errors in scheduling and planning tasks, grading errors by teachers in various disciplines, misinterpretation of statistical reasoning from datasets by data scientists, algorithms errors by programmers, and drawing incorrect conclusions in everyday life tasks even as simple as incorrect measurements in cooking recipes. Mathematical fallacies often result in misinformation, particularly for individuals experiencing Maths anxiety. Building confidence in mathematical skills and identifying errors becomes essential to address these challenges [31], [32].

We specifically investigate whether ChatGPT-powered robots could be used as maths tutors in primary schools and how children would respond in case of mistakes. We conducted a user study bringing a Pepper robot into a classroom at a primary school. The LLM responses were retrieved from the OpenAI `gpt-3.5-turbo` model using its published APIs. LLM and the Pepper Robot SDK text-to-speech were connected using custom-built Python code. The experiment was designed to investigate the following four hypotheses:

- **H1** Primary school students can detect whether ChatGPT-generated answers presented by a Pepper robot are correct or incorrect.
- **H2** The accuracy of spotting such mistakes is influenced by the student's correctness in answering the question.
- **H3** The accuracy of spotting such mistakes depends on the student's school year.
- **H4** The accuracy of spotting such mistakes is influenced by the student's gender.

The first two hypotheses indicate the ability to identify errors and how this is affected by students' correctness as a measure of confidence. To study other confounding characteristics in student responses, we analysed the effect of age on the results and the effect of gender because these were available in the collected dataset.

A. Participants

We recruited 77 primary school students from the Hatfield Community Free School, a local school in our area, from years 3 to 5 in the UK elementary school system, which corresponds roughly to between 7 and 11 years of age. This study was approved by the University of Hertfordshire ethics board under approval number (SPECS/SF/UH/05395). Informed consent was obtained from the parents or legal guardians of the children.

B. Generation of the LLM answers

The LLM was provided with the 29 questions as prompts, and it generated answers and explanations that were recorded and compared to the correct responses from the curriculum [33]. The analysis revealed that LLM's responses for 13 out of the 29 questions were inaccurate, with an approximate accuracy rate of 55%. Among the 13 incorrect answers, the LLM's answers and explanations were found to be wrong in 8 questions, while the remaining 4 had correct answers but incorrect explanations, and 1 question had partial correct answers from the several possible answers.

C. Procedure

The students were divided into groups according to their school year and participated in an activity where the Pepper robot was involved in asking them Maths questions, with the lead author as a facilitator, supported by the school teachers. The activity lasted for a total of 1 hour and 20 minutes, and we ensured that no more than 20 students were present at the time to avoid the activity from being too crowded.

The students were presented with 20 questions out of a pool of 29 questions chosen from the UK Maths curriculum for grade three [33]. The students in each session were presented with ten questions with LLM-generated correct answers and ten questions with LLM-generated incorrect answers in random order for each session. The students also had to complete a questionnaire asking their views about the robot's answers, following the steps below.

1) The respective question from the questionnaire was projected onto a screen and simultaneously displayed on Pepper's tablet. Additionally, Pepper read the question aloud.
2) Students were asked to write their answers on the provided questionnaire.
3) Pepper vocalised the generated answer from the LLM. This response was also displayed on Pepper's tablet and projected onto the screen.
4) Subsequently, students commented on the perceived correctness of Pepper's answer using a 5-point Likert scale, as illustrated in Figure 1.
5) In cases where Pepper's answer or explanation was incorrect, Pepper apologised and read the hard-coded correct answer text. Also, the lead author discussed with the students to clarify the accurate answer or explanation to avoid misinformation or lack of confidence in their correct answers.

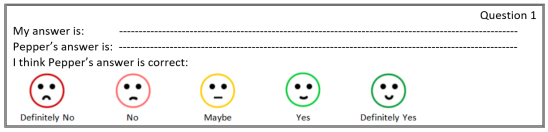

Fig. 1. Questionnaire items presented to students after each Maths question

IV. Results

The questionnaires completed by the students were collated for statistical analysis. We augmented the questionnaire responses with student scores about the correctness of their answers, age at the grade level, and their gender. Hypothesis testing was conducted to address the research questions, specifically comparing observed frequencies to expected frequencies. Given the nature of the comparisons, Chi-square statistical hypothesis testing was employed with a confidence level of 95%. The graphs illustrate student percentages on the y-axis, while the numbers displayed on the bars represent the corresponding student counts. Initially, the first graph presents LLM correctness as the sole bars along the x-axis. However, this metric is depicted atop grey shades in subsequent graphs, with the bars indicating the influence of confounding factors.

A. Robot Error

The most important research question is the ability of the students to identify when the LLM (Robot's first answer) is wrong, partially correct, or correct and rate it accordingly, formalised in **H1**. From Figure 1, we identify a low rating as the first two left-hand side faces. This means the student thinks the LLM answer is wrong. We considered that the middle face is chosen when the student is undecided about the LLM correctness, and the last two faces are chosen when the student thinks that the LLM answer is correct. Testing against the null hypothesis that there is no difference in the proportions of different students' ratings and whether LLM is in error or not, we calculated (χ^2 = 504.7, df=4, N = 1540, p <.001), and hence we reject the null hypothesis. The results suggest that 76.5% of students identified correct answers, 72% identified incorrect answers, and partial correctness caused some confusion. The confusion might be caused as they realise from previous questions that LLM errors are possible. The visualisation of the proportions is illustrated in Fig. 2.

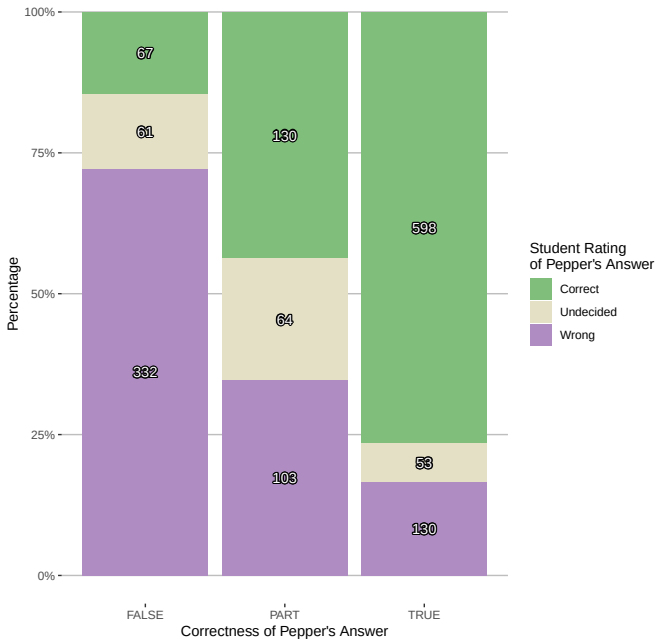

Fig. 2. Student rating of Pepper's answer by the correctness of the answer.

B. Student Correctness

When students correctly answer the presented Maths questions, we expected their confidence in identifying the LLM errors to be higher (**H2**). We tested against the null hypothesis that students' ability to identify LLM mistakes is not influenced by their ability to solve the question correctly. We calculated (χ^2 = 374.95, df=8, N = 1540, p <.001), hence rejecting the null hypothesis. The results suggest that 81% of students identified incorrect LLM answers when they were

correct as opposed to the earlier 72% regardless of whether they were correct. The visualisation of the proportions is illustrated in Figure 3.

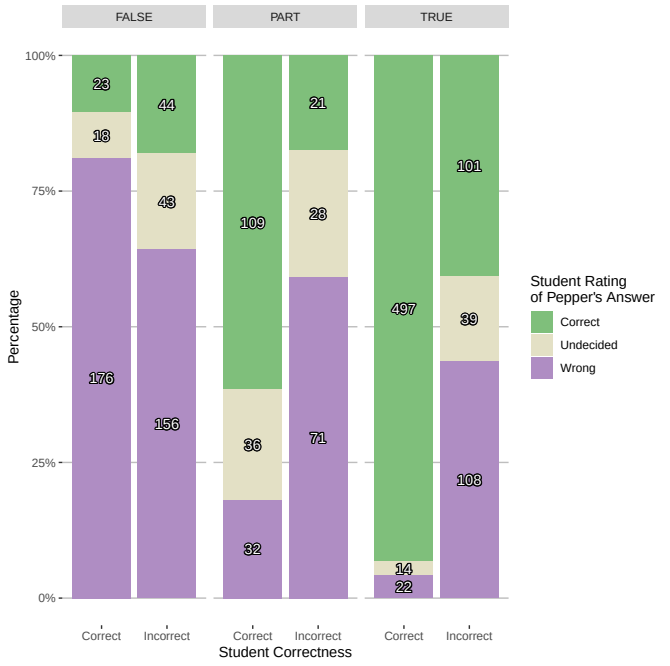

Fig. 3. Student rating of Pepper's answer by the correctness of the answer, split by student's correctness.

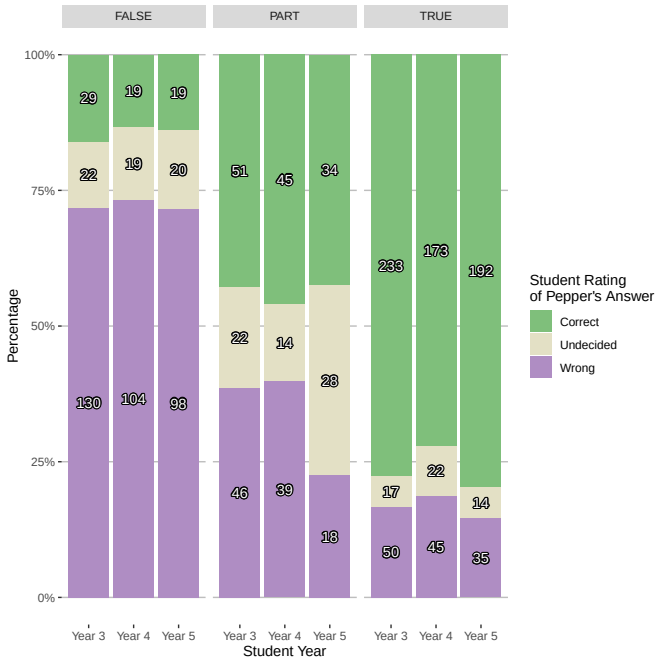

Fig. 4. Student rating of Pepper's answer by the correctness of the answer, split by the student's year.

C. Student Age/Year

Age or school year reflects the amount of training they had on the curriculum tested (cf. **H3**). We had 30 students in Year 3, 24 in Year 4, and 23 in Year 5. We tested against the null hypothesis that there is no difference in the accuracy of spotting LLM mistakes depending on the student's school year. We calculated (χ^2= 19.68, df=16, N = 1540, p >.05), and hence we fail to reject the null hypothesis. Although students are trained more on this content, it seems that the experiment setting, with other confounding factors, did not show a higher ability to detect LLM mistakes affected by the student's school grade. The illustration in Figure 4 further supports this conclusion.

D. Student Gender

We wanted to explore if gender significantly affected students' spotting LLM mistakes (**H4**). The study included 37 female and 40 male participants. We tested against the null hypothesis that there is no difference in the accuracy of spotting LLM mistakes influenced by the student's gender. We calculated (χ^2 = 37.88, df=8, N = 1540, p <.001). We can conclude that the observed difference in LLM error spotting varied significantly by gender. In our observations, female students tend to be correct more often than male students. The illustration in Figure 5 shows the proportions of correct vs incorrect answers for both genders.

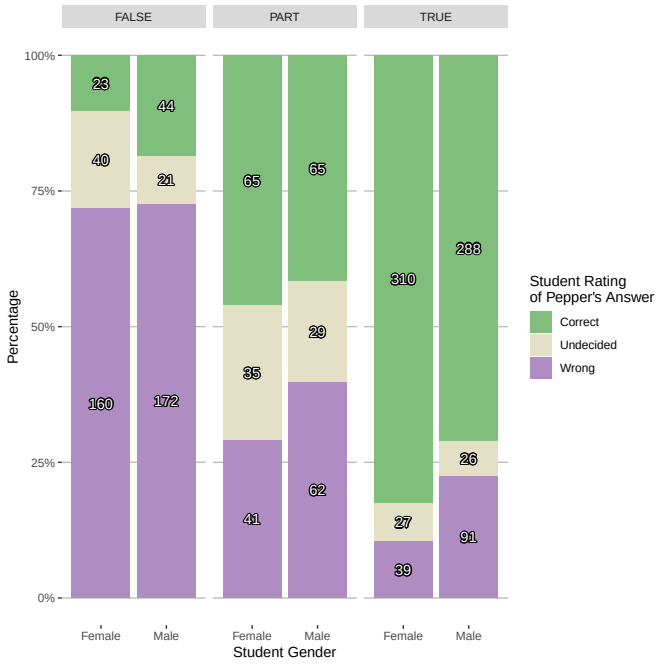

Fig. 5. Student rating of Pepper's answer by the correctness of the answer, split by the student's gender.

V. Discussion

In the first research question, students' ability to identify LLM errors at such a young age is observed with statistical significance. Our results reflect children's concentration on the different questions, although the difficulty of the questions, topics, and LLM-correctness varied among the questions, which were randomly presented to the children. It will be interesting to check if older learners can show such skill in identifying errors or if they will rely on technology, which could lead to accepting errors without verification due to a lack of time to check further.

Regarding the second research question, the results supported that students' ability to solve the question first increased their ability to identify LLM errors. This reflects confidence in their knowledge that did not change when Pepper, as an LLM-powered Maths Tutor, made some correct answers and some incorrect ones. This confidence in children is known to decrease as people get older [34]. This confirms the recommendation of [13] that prior knowledge in the domain LLM is used for is required to be able to verify and properly use the generated response.

The results of the third research question did not support our hypothesis. Although it is commonly agreed that maths practice through the years is expected to increase students' ability to identify correct answers, other confounding factors in the experimental settings could have distracted the older students. Noting that higher year groups responded to a Year 3 question set, it might also be that the students in higher year groups follow a different curriculum. Because of this, their performance was comparable with their younger peers, not better. A future extension of this study could consider exposing each year group to the specific questions extracted from their year curriculum, thus addressing this uncertainty.

The findings of the fourth research question align with other studies such as [35]. However, some studies find that males are better than females in Maths. In this experiment, our findings support the counter-argument. However, we have no baseline data regarding these students' mathematical knowledge, their interest in technology and their equal attentiveness to the task on the day.

VI. Limitations and Future Work

The observations and results are valid only for the recruited cohort and could be representative of similar age groups. Various confounding factors might have introduced noise in our dataset, that we did not account for in this experiment. A main confounding factor is socioeconomic student data, which usually affects learning skills. The school did not release information about students' socioeconomic backgrounds, so these factors might have impacted our observations.

A potential complicating factor when using a robot with primary school students is the possibility that they perceive it more as a toy than a learning tool. This perception could impact their ability to focus on detecting the error, especially at a young age. However, we cannot confidently assess this influence. Additionally, we question whether this factor could explain why older students, who presumably have more advanced skills and longer exposure to training, may also be distracted by the novel setting of math tutoring with a robot. The students' mutual interactions and further interactions with teachers might be influencing the results to be different than working independently with AI. It is a common interest of the UK government to evaluate all factors affecting Maths education, and study the various aspects of enhancing the UK international ratings [36].

A. Long-term impacts and Recommendations

Our experiment presents a snapshot of the current situation with other potential confounding factors. Long-term impact on cognition development from psychological, societal, deep effective learning and creativity and problem-solving skills are not well evaluated in the literature. Previous work discusses the use of LLMs in psychotherapy and its effects on behavioural health [14]. The authors discuss the consequences of AI failures that could lead to self-harm or other societal damage if no experts are involved to validate the outcomes. Also in defence and self-driving car applications, AI failure causes direct loss of life. Education does not pose fewer risks than other AI applications. Complete automation, labelled autonomous AI, is the highest level of trusting AI, as opposed to the lowest level in assistive AI, in which humans use AI but manage the process fully. Collaborative AI is a middle stage in which AI is heavily used and guided by humans in the loop. The authors recommend gradual integration of AI through these stages while measuring the outcome to ensure safe AI integration is achieved. The same recommendation can be reached about all AI integration in all application domains to avoid irrecoverable failures. Education has a long-term impact on the quality of life of individuals and societies. Measuring the safety of AI integration in education might require measuring the outcomes over a year or more to identify best practices. Other previous work iterated that the impact of using AI in teaching and learning on cognition development has yet to be evaluated [15]. The assessment for learning (AFL) link with AI-powered education has been systematically reviewed, exposing some challenges that still need to be addressed in the literature [16]. These challenges include fewer human interactions and their emotional impact on the learner, limited understanding due to lack of the required depth and breadth of the learning objectives, lack of creativity due to easy information finding, lack of contextual understanding due to fragmented pieces of information, lack of time and effort consideration in student assessments in disciplines where practice to mastery is required. This last challenge led to the proposal of using multi-modal systems in assessments to evaluate students' active involvement in their submissions rather than passively copying information. This is all besides the known issue of bias in training data of these models, the nontransparent learning algorithms owned by the companies producing these models, the responses generated depending on the provided prompts or data, and the privacy challenges. The following summarise our recommendations:

1) **AI Gradual integration:** Teachers should verify AI-generated knowledge, starting from assistive AI to collaborative AI with careful evaluation. However, we discourage reaching autonomous AI in Education (learners alone).
2) **Prior knowledge:** AI should be used after learner exposure to the educational content using brain activities over a reasonable time.
3) **Early Age Exposure:** AI-assisted education needs to be introduced to all ages to increase their confidence in identifying errors, accompanied by brain activities to aid in information retention, connecting to prior knowledge and mastery development.
4) **Teacher/Student - Focused:** AI should empower teachers more than learners. The presented experiment was student-focused, while many previous work was teacher-focused only. Measuring effects in hybrid teacher and student-focused experiments will be more realistic as AI is available to both.

B. Future work Directions

In future work, instead of tailoring the code connecting the LLM to Robots for the experiment details, more advanced LLM-powered Robots can use emerging APIs for a standardised code [37]. Aspects of selecting the level of challenge according to the curriculum relevant to each year group, and also considerations for assessing other STEM topics using the same approach remain the focus of our future work. We can reduce LLM inaccuracies by applying prompt engineering guidance [38]. We can also use fine-tuned LLM models trained on mathematics datasets to reduce the inaccuracy rate [39] or Retrieval Augmented Generation (RAG) techniques across different modalities [40]. We can also increase the number of participants from other schools to increase the representation and the significance of the study. We can collect more information about participants' scores in previous Math exams and other topics, and possibly more confounding factors.

VII. Conclusions

We support other studies' recommendations that LLM-powered education, whether using robots or indirectly in the teaching content preparation, assessments, or student support and feedback, should all be introduced gradually and after careful verification by domain experts. We recommend introducing LLM-powered learning support at all ages, though, to increase people's confidence in their knowledge and increase learners' abilities to identify mistakes and continue depending on brain activities that enhance mastery levels from valid references. Education relies on teachers' skills and this will not change in the AI era, however, teachers can be significantly empowered when using AI technologies responsibly. We recommend repeating similar experiments with young children, as we observed students' confidence in their knowledge increase, and their trust in AI becomes more balanced as they learn at a young age that AI can make mistakes. This study contributes to the growing body of research on AI in education, shedding light on the potential benefits and challenges associated with integrating LLM-powered technologies into classroom settings. While our findings suggest positive outcomes in terms of student engagement, error detection and some confounding factors, further research is warranted to explore optimal strategies for integrating LLM-powered educational tools into curriculum design and pedagogical practices. Ultimately, our study underscores the need for responsible and thoughtful implementation of AI technologies in education to enhance learning outcomes and support cognitive development in young learners.

Acknowledgments

We would like to gratefully acknowledge the participating students, their teachers, and the school administration hosting us during the 2023 UK Robotics Week.

References

[1] M. Malanchini, K. Rimfeld, A. G. Allegrini, S. J. Ritchie, and R. Plomin, "Cognitive ability and education: How behavioural genetic research has advanced our knowledge and understanding of their association," *Neuroscience & Biobehavioral Reviews*, vol. 111, pp. 229–245, Apr. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0149763419306748

[2] T. Alqahtani, H. A. Badreldin, M. Alrashed, A. I. Alshaya, S. S. Alghamdi, K. Bin Saleh, S. A. Alowais, O. A. Alshaya, I. Rahman, M. S. Al Yami, and A. M. Albekairy, "The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research," *Research in Social and Administrative Pharmacy*, vol. 19, no. 8, pp. 1236–1242, Aug. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1551741123002802

[3] M. Javaid, A. Haleem, R. P. Singh, S. Khan, and I. H. Khan, "Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 2, p. 100115, Jun. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2772485923000327

[4] M. Helal, P. Holthaus, G. Lakatos, and F. Amirabdollahian, "Chat failures and troubles: Reasons and solutions," in *CUI 2023 Workshop on Working with Trouble and Failures in conversation between humans and robots - WTF*, 2023.

[5] A. Extance, "ChatGPT has entered the classroom: how LLMs could transform education," *Nature*, vol. 623, no. 7987, pp. 474–477, Nov. 2023. [Online]. Available: https://www.nature.com/articles/d41586-023-03507-3

[6] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, Y. Wang, L. Yang, J. Wang, X. Xie, Z. Zhang, and Y. Zhang, "Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity," Dec. 2023, arXiv:2310.07521 [cs]. [Online]. Available: http://arxiv.org/abs/2310.07521

[7] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models," Dec. 2023, arXiv:2303.10420 [cs]. [Online]. Available: http://arxiv.org/abs/2303.10420

[8] A. M. Elkhatat, K. Elsaid, and S. Almeer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," *International Journal for Educational Integrity*, vol. 19, no. 1, p. 17, Sep. 2023. [Online]. Available: https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00140-5

[9] T. Foltynek, S. Bjelobaba, I. Glendinning, Z. R. Khan, R. Santos, P. Pavletic, and J. Kravjar, "ENAI Recommendations on the ethical use of Artificial Intelligence in Education," *International Journal for Educational Integrity*, vol. 19, no. 1, pp. 1–4, Dec. 2023, number: 1 Publisher: BioMed Central. [Online]. Available: https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00133-4

[10] M. Hosseini, D. B. Resnik, and K. Holmes, "The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts," *Research Ethics*, vol. 19, no. 4, pp. 449–465, Oct. 2023. [Online]. Available: http://journals.sagepub.com/doi/10.1177/17470161231180449

[11] M. Bernabei, S. Colabianchi, A. Falegnami, and F. Costantino, "Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100172, 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666920X23000516

[12] C. K. Lo, "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature," *Education Sciences*, vol. 13, no. 4, p. 410, Apr. 2023. [Online]. Available: https://www.mdpi.com/2227-7102/13/4/410

[13] A. Shoufan, "Can Students without Prior Knowledge Use ChatGPT to Answer Test Questions? An Empirical Study," *ACM Transactions on Computing Education*, vol. 23, no. 4, pp. 1–29, Dec. 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3628162

[14] E. C. Stade, S. W. Stirman, L. H. Ungar, C. L. Boland, H. A. Schwartz, D. B. Yaden, J. Sedoc, R. DeRubeis, R. Willer, and J. C. Eichstaedt, "Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation," PsyArXiv, preprint, Apr. 2023. [Online]. Available: https://osf.io/cuzvr

[15] I. Tuomi, *The impact of artificial intelligence on learning, teaching, and education: policies for the future*, M. Cabrera, R. Vuorikari, and Y. Punie, Eds. Luxembourg: Publications Office of the European Union, 2018, oCLC: 1076558603.

[16] B. Memarian and T. Doleck, "A review of assessment for learning with artificial intelligence," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, p. 100040, Jan. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2949882123000403

[17] W. Yang, "Artificial Intelligence education for young children: Why, what, and how in curriculum design and implementation," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100061, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666920X22000169

[18] J. Su, D. T. K. Ng, and S. K. W. Chu, "Artificial Intelligence (AI) Literacy in Early Childhood Education: The Challenges and Opportunities," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100124, 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666920X23000036

[19] N. Perucica, "Our children are growing up with AI: what you need to know," Jan. 2022. [Online]. Available: https://www.weforum.org/agenda/2022/01/artificial-intelligence-children-technology/

[20] C. Chen, "AI Will Transform Teaching and Learning. Let's Get it Right." Mar. 2023. [Online]. Available: https://hai.stanford.edu/news/ai-will-transform-teaching-and-learning-lets-get-it-right

[21] K. Wang, G.-Y. Sang, L.-Z. Huang, S.-H. Li, and J.-W. Guo, "The Effectiveness of Educational Robots in Improving Learning Outcomes: A Meta-Analysis," *Sustainability*, vol. 15, no. 5, p. 4637, Mar. 2023. [Online]. Available: https://www.mdpi.com/2071-1050/15/5/4637

[22] B. Robins, K. Dautenhahn, and J. Nadel, "Kaspar, the social robot and ways it may help children with autism–an overview," *Enfance*, no. 1, pp. 91–102, 2018.

[23] S.-T. Chu, G.-J. Hwang, and Y.-F. Tu, "Artificial intelligence-based robots in education: A systematic review of selected SSCI publications," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100091, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666920X22000467

[24] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, "Large Language Models for Robotics: A Survey," Nov. 2023, arXiv:2311.07226 [cs]. [Online]. Available: http://arxiv.org/abs/2311.07226

[25] F. Sacco, G. Rossini, F. Manzi, C. Di Dio, L. Aquilino, A. Cangelosi, L. Raggioli, D. Massaro, and A. Marchetti, "An Antropomorphic Robot with ChatGPT for Learning Activities: The Teachers' Perspective," in *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. Milano, Italy: IEEE, Oct. 2023, pp. 1166–1170. [Online]. Available: https://ieeexplore.ieee.org/document/10405596/

[26] H. Ismail, N. Hussein, S. Harous, and A. Khalil, "Survey of Personalized Learning Software Systems: A Taxonomy of Environments, Learning Content, and User Models," *Education Sciences*, vol. 13, no. 7, p. 741, Jul. 2023. [Online]. Available: https://www.mdpi.com/2227-7102/13/7/741

[27] A. Alkhatlan and J. Kalita, "Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments," Dec. 2018, arXiv:1812.09628 [cs]. [Online]. Available: http://arxiv.org/abs/1812.09628

[28] M. Messer, N. C. C. Brown, M. Kölling, and M. Shi, "Machine Learning-Based Automated Grading and Feedback Tools for Programming: A Meta-Analysis," in *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. Turku Finland: ACM, Jun. 2023, pp. 491–497. [Online]. Available: https://dl.acm.org/doi/10.1145/3587102.3588822

[29] V. E. S. Seaba, "Revolutionizing Education: Exploring the Potential of AI-Enabled Brain-Based Learning for Enhanced Cognitive Development," *OALib*, vol. 10, no. 10, pp. 1–20, 2023. [Online]. Available: http://www.oalib.com/paper/pdf/6806078

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[31] C. Foster, S. Woodhead, C. Barton, and A. Clark-Wilson, "School students' confidence when answering diagnostic questions online," *Educational Studies in Mathematics*, vol. 109, no. 3, pp. 491–521, Mar. 2022. [Online]. Available: https://link.springer.com/10.1007/s10649-021-10084-7

[32] J. J. Rolison, K. Morsanyi, and E. Peters, "Understanding Health Risk Comprehension: The Role of Math Anxiety, Subjective Numeracy, and Objective Numeracy," *Medical Decision Making*, vol. 40, no. 2, pp. 222–234, Feb. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0272989X20904725

[33] C. Yemm, *Maths Problem Solving Year 3*. Brilliant Publications, 2005.

[34] R. W. Robins and K. H. Trzesniewski, "Self-Esteem Development Across the Lifespan," *Current Directions in Psychological Science*, vol. 14, no. 3, pp. 158–162, Jun. 2005. [Online]. Available: http://journals.sagepub.com/doi/10.1111/j.0963-7214.2005.00353.x

[35] J. S. Hyde, E. Fennema, and S. J. Lamon, "Gender differences in mathematics performance: A meta-analysis." *Psychological Bulletin*, vol. 107, no. 2, pp. 139–155, 1990. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.107.2.139

[36] H. Blausten, C. Gyngell, H. Aichmayr, and N. Spengler, "Supporting Mathematics Teaching for Mastery in England," in *Empowering Teachers to Build a Better World: How Six Nations Support Teachers for 21st Century Education*, ser. SpringerBriefs in Education, F. M. Reimers, Ed. Singapore: Springer Singapore, 2020. [Online]. Available: http://link.springer.com/10.1007/978-981-15-2137-9

[37] S. Paulo, "Build LLM-powered robots in your garage with MachinaScript For Robots: GitHub - babycommando/machinascript-for-robots:." [Online]. Available: https://github.com/babycommando/machinascript-for-robots/tree/main

[38] X. Liu, J. Wang, J. Sun, X. Yuan, G. Dong, P. Di, W. Wang, and D. Wang, "Prompting Frameworks for Large Language Models: A Survey," Nov. 2023, arXiv:2311.12785 [cs]. [Online]. Available: http://arxiv.org/abs/2311.12785

[39] Y. Zhang, Y. Luo, Y. Yuan, and A. C.-C. Yao, "Autonomous Data Selection with Language Models for Mathematical Texts," Apr. 2024, arXiv:2402.07625 [cs]. [Online]. Available: http://arxiv.org/abs/2402.07625

[40] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, and B. Cui, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," Mar. 2024, arXiv:2402.19473 [cs]. [Online]. Available: http://arxiv.org/abs/2402.19473