International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 1, pp. 332-352, January 2026
https://doi.org/10.26803/ijlter.25.1.17
Received Oct 25, 2025; Revised Nov 27, 2025; Accepted Dec 16, 2025

Can Machines Think? Assessing the Accuracy of GenAI Chatbots in a Physics University Entrance Exam

Samuel Jere*
University of Venda
Thohoyandou, South Africa

Abstract. The rapid advancement of artificial intelligence (AI) over the past few years has led to the development of Generative AI (GenAI) tools with enhanced capabilities, including multimodal functionality, reduced susceptibility to hallucinations, and real-time access to internet resources. Past studies have revealed that GenAI tools are used in daily life across various fields, including education, healthcare, engineering, and software development. School learners are increasingly relying on them for their academic activities. There is a paucity of empirical research on the accuracy of these tools' responses, particularly in the field of physics education. This mixed-method case study aims to evaluate the accuracy of responses from ChatGPT and Google Gemini chatbots in answering physics university entrance exams in South Africa. Technological Pedagogical Content Knowledge and Webb's Depth of Knowledge were used to construct the theoretical framework. The research instrument used in this study was the 2024 university entrance physics exam paper in South Africa. The question paper was loaded into each chatbot, then they were prompted to respond to the questions. Two expert examiners assessed the responses of the chatbots. The performance of each chatbot was compared to that of the learners who took the exams. The findings were that the chatbots outperformed the learners. This study's findings suggest that these chatbots can serve as teaching assistants to support learners in exam preparation and formative assessment tasks. Learners should employ critical thinking skills to assess the responses they receive from chatbots during interactions.

Keywords: Chatbots; ChatGPT; Gemini; Generative Artificial Intelligence; Webb's Depth of Knowledge

*Corresponding author: Samuel Jere; *samuel.jere@univen.ac.za*

1. Introduction

The question "Can machines think?" was asked several decades ago (Turing, 2009). It is only now that it appears like we are on the verge of getting the answer, as the producers of GenAI chatbots describe their models as being capable of thinking. For example, OpenAI describes its latest model, GPT-5, as the "… smartest, fastest, most useful model yet, with built-in thinking …" (OpenAI, 2025b). Similarly, Google's recent model, Gemini 2.5 Pro, is described as a thinking model with multimodal reasoning capabilities including solving complex problems such as coding, mathematics and scientific problems (Comanici et al., 2025).

There is evidence that chatbots are not only being produced for general purposes but are also being developed to assist learners (Chen et al., 2020). They can be used in assessment and evaluation, intelligent tutoring, the personalisation of learning, and assisting teachers in instructional activities (Chen et al., 2020; Jere et al., 2024; Zawacki-Richter et al., 2019). As these chatbots become increasingly widely available and are being continuously improved, there is a need for researchers to evaluate the claims that chatbots are capable of thinking and to determine the accuracy of their responses to educational problems (Kooli, 2023; Kuhail et al., 2023). As learners utilise these chatbots in their studies, it is essential to have empirical evidence of their capabilities to better assess their potential in education (Jere, 2025).

The purpose of this study was to assess the accuracy of the responses generated by the latest ChatGPT and Gemini models to answer the grade 12 physics university entrance examination in South Africa. This study aimed to answer the following research question: 'What kind of physics problems are chatbots capable of answering, and what is the nature of problems that chatbots find challenging?' The rest of this paper is structured as follows: the next section reviews the related literature to identify the research gap. The theoretical framework is then presented, followed by the methodology. Thereafter, the findings and discussion are presented, leading to the implications and limitations of this study and the conclusion.

2. Literature Review

In late 2022, ChatGPT released the first widely available chatbot, GPT-1 (Plevris et al., 2023). In a short period, the model underwent several revisions, progressing from version 1 to 3.5 to 4. Now that GPT-5 has been released, its capabilities have been further enhanced. During the same period, Google released its chatbot, first as Bard, and then it underwent several iterations. The current version is Gemini 2.5, with advanced capabilities. While ChatGPT and Gemini could initially only understand the information presented to them in text format, they have now advanced to a stage where they are regarded as multimodal (Ahmed et al., 2025; Rane et al., 2024). They can comprehend information presented in various formats, including text, image, audio, and video (Comanici et al., 2025; OpenAI, 2025a).

The architecture of GPT-5 is comprised of three models: a deeper reasoning model, which can either be GPT-5-Thinking or GPT-5-Thinking-Mini; a smart, faster model, which can either be GPT-5-Main or GPT-5-Main-Mini; and a real-time router that enables the architecture to switch between the reasoning model and the fast model (OpenAI, 2025a). The latest model of Google's AI chatbot, Gemini 2.5, is comprised of several versions, including the free version 2.5 Flash and the subscription-based version 2.5 Pro. Gemini 2.5 Pro is described as the most intelligent thinking model, capable of advanced reasoning and high coding abilities (Comanici et al., 2025).

There are claims that chatbots like GPT-5 and Gemini 2.5 Pro have advanced to the point where they can perform as well as human beings in examinations at various levels of education across multiple fields (Liu et al., 2025; Newton et al., 2025). The performance of these chatbots in public science examinations is important for several reasons. First, if chatbots can answer questions accurately, they can be used as teaching assistants to help learners with exam preparation, allowing them to focus on more challenging tasks at higher levels of Bloom's taxonomy of educational objectives (Jere & Mpeta, 2025).

The teacher will no longer be the sole source of knowledge, as knowledge becomes multifaceted. Secondly, chatbots can be useful for personalising learning (Chang et al., 2023). As learners have different abilities, having an AI chatbot that can personalise the learning content to each learner's level can help learners who are at risk of not performing well. Finally, the interaction between the learner and the chatbot can significantly enhance the comprehension of difficult concepts in physics.

There are conflicting findings regarding the accuracy of chatbots when answering examination questions in various fields. For example, while some studies demonstrate that chatbots can answer questions accurately, others reveal that they perform poorly in examinations (Al-Thani et al., 2025; Jere, 2025; Xuan-Quy et al., 2023). Some studies have demonstrated that AI chatbots can answer physics problems with great accuracy. For example, Chapagain et al. (2024) tested chatbots and found that ChatGPT 4 could correctly answer 90% of the grade 12 final exams in Nepal. However, other studies have shown that ChatGPT has not yet developed to answer questions accurately.

A study by Revalde et al. (2025) revealed that the chatbot achieved only 17% of the maximum possible score. Similarly, Demirci (2025) investigated the accuracy of ChatGPT, Gemini and Copilot when answering physics university entrance exams in Turkey. It was found that ChatGPT achieved 38.09% of the maximum possible score, while both Gemini and Copilot achieved 28.57%. Xuan-Quy et al. (2023) found that in Vietnam, chatbots were unable to accurately answer questions that required application, and learners outperformed the chatbots. These mixed results suggest that further research is needed to determine whether chatbots can accurately answer physics university entrance exams.

This study aimed to evaluate the accuracy of the most recent models of ChatGPT and Gemini when answering physics university entrance examination questions, and to investigate the types of questions the chatbots could accurately answer and those where they encountered difficulties. This study contributes to closing the gap left by previous research by providing a deeper analysis of the accuracy of GenAI chatbots when used in physics education. If these chatbots can be found to offer accurate responses to examination questions, then teachers and learners can incorporate them into their activities while preparing for examinations and other informal formative assessments. The next section discusses the theoretical framework that guided this study.

3. Theoretical Framework

The theoretical framework was crafted from two different but complementary theoretical perspectives to develop a comprehensive understanding of the integration of GenAI chatbots into physics education. This study draws on the scholarship of the Technological Pedagogical Content Knowledge (TPACK) framework (Mishra & Koehler, 2006) and the Depth of Knowledge (DoK) framework (Webb, 2002). The TPACK model extends the understanding that AI tools can be used as mediating agents. The model suggests integrating technology, proposing that teachers need to develop a comprehensive understanding of how knowledge of content, pedagogy, and technology can be blended to maximise learning. This study will contribute to the teachers' TPACK knowledge of the capabilities and role of AI tools in physics education.

Apart from lenses that offer insight into how learning can be enhanced by AI chatbots, a theoretical framework was required to analyse the questions presented to the chatbots. Webb's (2002) Depth of Knowledge framework was used for this. Questions in physics assessment instruments can be ranked according to their level of difficulty. Test items range from those that require recall to more complex questions involving combining the learner's knowledge in new ways (Woitkowski, 2020). The exam questions assess the extent to which the curriculum objectives and goals have been attained (Marzano & Kendall, 2006). Webb (2002) classifies objectives and test items into four levels, ranging from those that require simple recall to those that require more complex thinking and reasoning. These four categories are Level 1, recall and reproduction; Level 2, skills and concepts; Level 3, strategic thinking; and Level 4, extended thinking (Webb, 2002).

Test items were classified into Level 1 if they required recalling facts, definitions or scientific procedures. Level 2 test items were those that required the learner to engage in cognitive processes that extend beyond recalling, necessitating combining several steps to arrive at a solution (Webb, 2002). This means that questions that require learners to collect, organise, compare, and interpret data (Webb, 2002) are considered Level 2 questions. The defining feature of Level 3 questions, classified as strategic thinking, is that they require complex and abstract thinking at a higher level of reasoning than the previous two levels (Webb, 2002).

Questions that have several possible solutions and require learners to consider the best approach are generally regarded as Level 3 questions. These questions are more cognitively demanding. Asking learners to explain how they arrived at the solution or to justify their answer is regarded as a Level 3 question. Questions that require learners to apply concepts they have acquired to solve novel problems, develop logical arguments, design investigations and scientific models, and formulate conclusions from experimental data are considered to be Level 3 questions (Webb, 2002). Level 4 questions require significantly more complex thinking and should be addressed over an extended period, typically through learner projects (Cvenic et al., 2022). Therefore, Level 4 questions are beyond the scope of physics university entrance examinations.

4. Methodology

This study employed a case study research design to allow for an in-depth exploration of complex issues according to a naturalistic approach (Crowe, et. al, 2011). This design was suitable for the study as it enables the researcher to answer the "how", "what", and "why" questions to explore, describe and explain the case (Crowe, et al.). al, 2011). The case study examined the accuracy of the responses generated by Gemini and ChatGPT chatbots to answer the November 2024 university entrance physics exam in South Africa.

In South Africa, physics is offered as part of the physical sciences from grade 10 to grade 12. The major topics covered in physics include mechanics (vectors and scalars, kinematics, Newton's laws, work energy and power, momentum and impulse, and vertical projectile motion), waves, sound, light (transverse and longitudinal waves, wave properties, sound waves and the Doppler effect, and wave-particle duality and the photoelectric effect) and electricity and magnetism (electrostatics and current electricity, electric circuits and electromagnetic effects). In the final year of secondary education, learners sit a physics paper (Paper 1) and a chemistry paper (Paper 2). These are high-stakes exams used as university entrance exams, allowing learners to pursue science-related careers such as engineering and medical professions. To achieve the purpose of this study, the November 2024 exam paper was purposively sampled as it was the most recent examination.

The physics papers were downloaded from the Department of Basic Education website on two different devices. These papers were then uploaded into Gemini 2.5 Pro (Deep Research) and ChatGPT (GPT-5) on two different devices on August 12, 2025. For the responses from ChatGPT, the platform used was Android, and the interface was the ChatGPT mobile app with the App version being ChatGPT/1.2025.315, running on OS version 16. The default configuration settings were used without modification, and the decoding parameters, replicas and aggregation methods were not visible to the user. The responses from Google Gemini were obtained from the Google Gemini web platform through the standard conversational interface. The model used was the Gemini Advanced Deep Thinking model, using the default configuration settings.

The chatbots were prompted to answer all questions with no further instructions. The responses from the chatbots were copied into two answer scripts. These were labelled Script 1 and Script 2, and were printed. The marking guidelines document and question papers that were used that year were downloaded from the same website. Four copies of the marking guidelines and question papers were made. Two experienced examiners from the Department of Basic Education were requested to participate in the study. The author held a meeting with the two experts to discuss the two tasks related to assessing the scripts. The first task involved classifying the questions using Webb's (2002) DoK framework.

The panel discussed Webb's (2002b) DoK framework and the classification of the questions. This was followed by a discussion on the question paper and the marking guidelines. The marking guidelines were strictly adhered to during the marking process. As the marking guidelines require that partially correct answers are awarded as indicated in the guide, the examiners followed this instruction. The examiners were then allowed to follow the marking guidelines to assess the chatbots' answer scripts and to classify the questions according to Webb's Depth of Knowledge Framework. It was agreed that the results would be submitted after one week.

The reliability of the coding process for classifying the exam questions into Webb's DoK framework was determined using SPSS version 29 to determine Cohen's kappa (k) (Cohen, 1960) after receiving the results from the examiners. The results indicated almost perfect agreement ($k = 0.914$, $p < .001$) (Landis & Koch, 1977) for the classification of the exam questions into Webb's DoK framework. The intraclass correlation coefficient (ICC) was used in the interrater reliability analysis of the exam marking process (Koo & Li, 2016). The ICC was determined using a two-way model and absolute agreement in SPSS version 29, with the results suggesting excellent agreement between the examiners for the questions answered by ChatGPT (ICC (2,1) = 0.93, CI [0.88, 0.96], $p < .001$) (Koo & Li, 2016).

Excellent agreement was also observed between the two examiners in their marking of the answers generated by Gemini (ICC (2,1) = 0.94, CI [0.90, 0.97], $p < .001$). A meeting was held between the researcher and the two examiners to discuss the classification of the exam questions and the marking process. Any differences between the examiners in both the marking process and the classification of questions were resolved, and the results are presented in the findings and discussion section that follows.

5. Findings and Discussion

The physics examination used as the research instrument had ten questions and fifty-seven sub-questions. These questions and sub-questions were classified using Webb's DoK framework, and the findings are presented in Table 1. The physics exam questions were presented to Google's Gemini 2.5 Pro Deep Thinking chatbot model and ChatGPT-5, and the findings on the chatbots' performance are presented in Table 1.

Table Error! No text of specified style in document..**1: ChatGPT-5 and Gemini 2.5 Pro Performance**

Depth of Knowledge Level	Questions	ChatGPT Marks	Gemini Marks	Total Marks
Level 1: Recall and Reproduction	1.1; 1.6; 1.8; 1.10; 2.1; 4.1; 5.1; 6.1; 7.3.1; 8.1; 9.1; 9.2; 9.5; 10.1; 10.3.1	25 (93%)	26 (96%)	27 (100%)
Level 2: Skills and Concepts	1.2; 1.3; 1.4; 1.5; 1.7; 1.9; 2.2; 2.4; 3.1.2; 3.1.3; 3.2; 4.2; 4.3.1; 4.3.2; 4.4; 5.2; 5.4; 6.2; 6.3.1; 7.1; 7.2; 8.2.1; 8.2.2; 9.3; 9.4; 9.6; 10.2.1; 10.2.2; 10.2.4; 10.3.2	50 (68%)	57 (78%)	73 (100%)
Level: 3 Strategic Thinking	2.3.1; 2.3.2; 3.1.1; 5.3; 6.3.2; 7.3.2; 7.3.3; 8.2.3; 8.2.4; 8.3; 9.7; 10.2.3;	31 (62%)	37 (74%)	50 (100%)
Total		106 (71%)	120 (80%)	150 (100%)

The results in Table 1 indicate that Gemini was able to accurately answer four out of 10 examination questions, while the sampled learners and ChatGPT could not accurately answer all questions for any of the 10 exam questions. The chatbots outperformed the learners in all questions, with the following exceptions. In question 2, learners outperformed ChatGPT; in question 8, learners outperformed Gemini; and in question 6, both chatbots were outperformed by the learners. Overall, Gemini was more accurate with a performance of 80% followed by ChatGPT, which performed at 71%. The learners managed an average performance of 49%. Key definitions of terms are provided in Table 2 to facilitate the understanding of the values in Table 1. From this analysis, it is apparent that the chatbots outperformed the learners in terms of generating accurate responses to the exam questions (Figure 1).

Table Error! No text of specified style in document..**2: Definitions of the Key Terms for Interpreting Chatbot Performance**

Term	Definition	Denominator for percentage values
Item	A single question in the exam paper made up of sub-questions, e.g. Question 2.	Total marks for the entire question.
Sub-Item	Part or component of a question, e.g. Question 2.1	Total marks for the sub-question.
"Totally Correct" Question	When the chatbot answers all sub-questions correctly, scoring all marks for that item.	Total marks for the question
Correct Answer per Sub-question	Total or partial marks obtained per sub-question.	Total marks for the sub-question.

Recent studies support the findings that chatbots are now increasingly competent at answering physics questions at the university entrance level. For example, Tschisgale et al. (2025) assessed the problem-solving abilities of LLMs in

answering Olympiad-type physics questions and found that the LLMs outperformed human subjects. Where LLMs are found to underperform, prompt engineering can be used to obtain responses that approach expert-level reasoning from the chatbots (Polverini & Gregorcic, 2024).

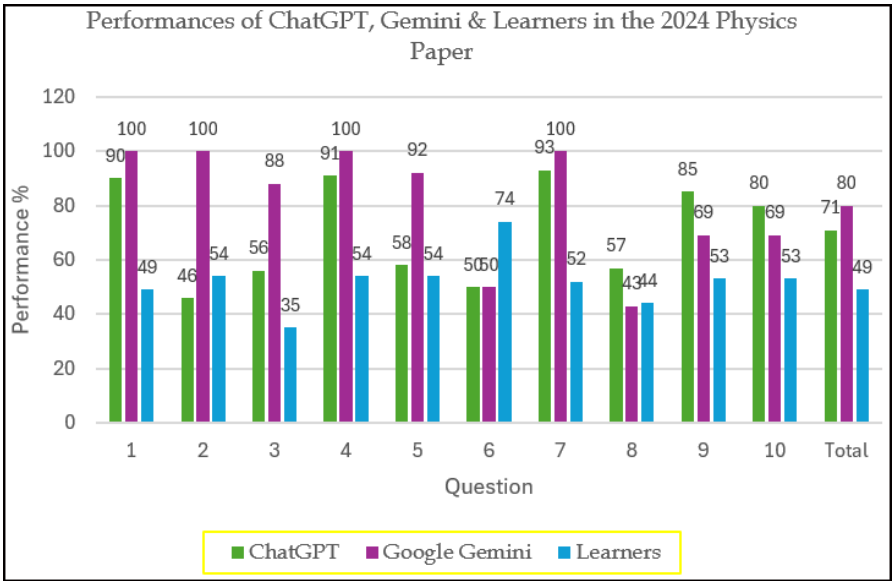


Figure 1: **Performance of ChatGPT-5, Gemini 2.5 and learners per question**

The questions sampled to illustrate how chatbots performed in this exam are now presented in terms of Webb's DoK framework, ranging from Level 1 (recall and reproductions) to Level 2 (skills and concepts), and finally to Level 3 (strategic thinking). No questions were classified under Level 4 (extended thinking) by the examiners. The following sub-section presents the findings for Level 1 questions.

5.1 Level 1 - Recall and Reproduction

Question 1 had ten multiple choice items. Four of these items were classified as Level 1, recall and reproduction questions. These items were 1.1, 1.6, 1.8 and 1.10 (Figure 2). Both ChatGPT and Gemini were able to accurately answer all recall and reproduction multiple choice questions.

1.1 Several forces are acting on a moving object. Which ONE of the following statements is CORRECT when these forces are in equilibrium?

A. The velocity of the object is increasing.
B. The object is moving at a constant velocity.
C. The kinetic energy of the object is decreasing.
D. The object has a non-zero acceleration. (2)

1.6 The absorption spectrum of an element surrounding a moving star is observed on Earth and found to be redshifted. Which ONE of the following combinations is CORRECT for the movement of the star and the frequency of the observed light on Earth?

	MOVEMENT OF STAR	FREQUENCY OF OBSERVED LIGHT ON EARTH
A	Away from Earth	Decreased
B	Towards Earth	Decreased
C	Away from Earth	Increased
D	Towards Earth	Increased

(2)

1.8 The kilowatt-hour (kWh) is a unit of …

A. power.
B. electric current.
C. electrical energy.
D. potential difference. (2)

1.10 Which of the following statements is/are TRUE for the photoelectric effect? The photoelectric effect demonstrates that: (i) Light has a wave nature, (ii) Light has a particle nature, (iii) Light energy is quantised

A. (i) only
B. (ii) only
C. (i) and (iii) only
D. (ii) and (iii) only (2)

Figure 2: **Recall and Reproduction Multiple Choice Questions**

The first question, 1.1, required the candidates to recall that when the forces acting on an object are at equilibrium, the net force acting on it is zero as the forces are balanced. Therefore, the object moves at constant velocity. Most candidates had a good understanding of these concepts, as evidenced by the average performance of 72% for this question (Figure 3).

In question 1.6, recalling that a redshift means that the light on Earth would have a lower frequency and longer wavelength, implying that the star is moving away from the Earth, was sufficient for candidates to answer the question. However, only 65% of the candidates were able to understand this (Figure 3). Similarly, question 1.8 was a recall question. The candidates were expected to recall that the kilowatt-hour is the unit of energy, as it is a unit obtained by multiplying the unit of power by the unit of time. Surprisingly, only 39% of candidates could answer the question correctly.

Question 10.1 required the recall of concepts regarding the photoelectric effect. In accordance with Einstein's theory, light is made of particles called photons. During the photoelectric effect, each electron on a metal surface absorbs a single photon, and if the energy of a photon, hf, is greater than the work function, Φ, then the electron is ejected. Recalling this information would allow the learners to realise that (ii) and (iii) are correct. However, the photoelectric effect does not demonstrate the wave nature of light, making (i) false. Only 33% of the learners were able to understand these concepts (Figure 3). GenAI tools can be used to assist learners in comprehending recall and reproduction questions related to these concepts as this study has shown that they are capable of generating accurate and correct responses.

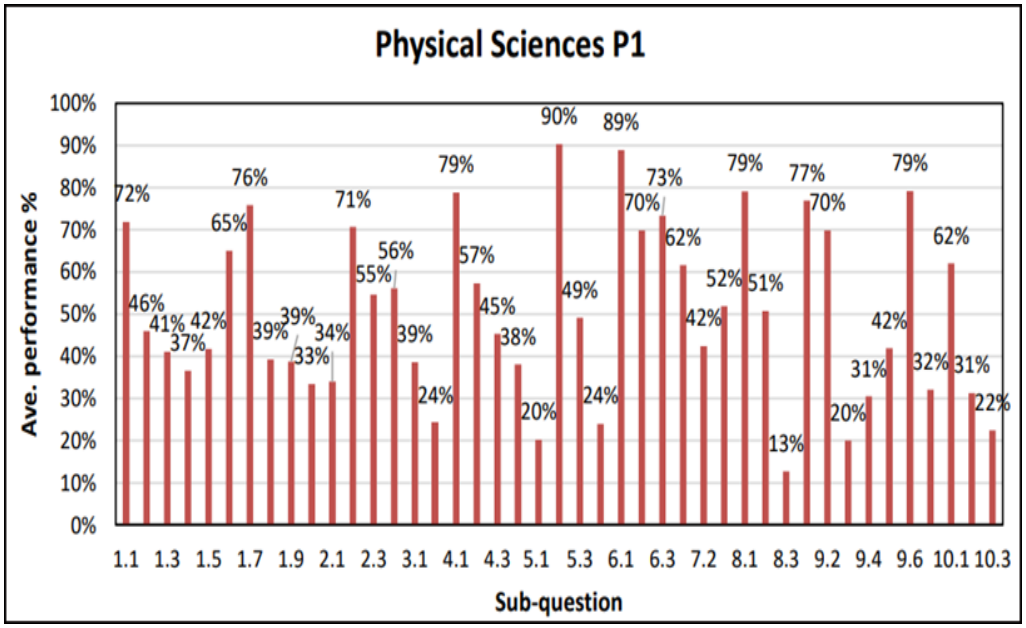


Figure 3: **Average performance of candidates per sub-question (DBE, 2024)**

The chatbots were able to provide comprehensive responses to questions that required the recall and reproduction of information, answering questions 2 to 10. The responses of the two chatbots, together with the expected answers from the marking guidelines, are presented in Table 2. A significant proportion of learners struggled to answer these questions accurately. For example, Figure 3 shows that the average performance of learners was only 34% for accurately answering question 2.1, and only 20% for accurately answering question 5.1.

Table Error! No text of specified style in document..**3: Responses of the chatbots to questions requiring recall and reproduction**

Question	Marking Guideline Answer (DBE, 2024)	ChatGPT Answer	Gemini Answer
2.1 Defining *static friction*	Force that opposes the tendency of motion of a stationary object relative/parallel to a surface.	Frictional force that acts between two surfaces when there is no relative motion; it adjusts up to a maximum value of $f_{s,\,max} = \mu_s N$.	Force that opposes the tendency of motion between two surfaces in contact when they are at rest relative to each other.
4.1 Stating the *principle of conservation of mechanical energy*	The total mechanical energy/sum of gravitational potential energy and kinetic energy, in an isolated system, remains constant or is conserved.	In an isolated system with only conservative forces acting, the total mechanical energy (kinetic + potential) remains constant.	The total mechanical energy (sum of the kinetic energy and potential energy) of an isolated system remains constant.
5.1 Defining *work done by a force*	The work done on an object by a constant force F is $F.\Delta x.\cos\theta$, where F is the magnitude of the force, Δx the magnitude of the displacement and θ the angle between force and displacement.	Scalar product of force and displacement: $W = F.\Delta x.\cos\theta$. It is the energy transferred by the force when it moves the object through displacement Δx.	It is quantitatively defined as the product of the magnitude of the force, the magnitude of the displacement of the object, and the cosine of the angle between the direction of the force

			and the direction of the displacement ($W = F.\Delta x.\cos\theta$). Work is a scalar quantity.
7.3.1 Stating *Coulomb's law*	The magnitude of the electrostatic force exerted by one point charge to another point charge is directly proportional to the product of the magnitudes of their charges, and inversely proportional to the square of the distance between them.	The magnitude of the electrostatic force between two point charges is directly proportional to the product of the magnitudes of the charges, and inversely proportional to the square of the distance between them.	Coulomb's Law states that the electrostatic force between two point charges is directly proportional to the product of the magnitudes of the charges, and inversely proportional to the square of the distance between their centres.
8.1 Defining *power*	The rate at which work is done/dissipated OR energy transferred OR Work done per unit time.	Power is the rate of transfer of energy; electrical power $P = VI$ or $P = I^2R$ or $P = V^2/R$.	Power is defined as the rate at which work is done or energy is transferred or converted.
10.1 Defining *photoelectric effect*	The process whereby electrons are ejected from a metal surface when light (of suitable frequency) is shining on that surface.	The emission of electrons from a metal surface when electromagnetic radiation (light) of a frequency greater than the threshold frequency; electrons are emitted if hf exceeds the work function and the maximum kinetic energy is $K_{max} = hf - \Phi$.	The photoelectric effect is the phenomenon in which electrons are ejected from the surface of a metal when light of a sufficiently high frequency (above a certain threshold frequency) shines on it.

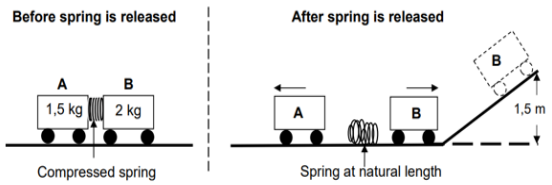
It is clear that the chatbots can help learners to comprehend questions at Level 1 using Webb's DoK framework, dealing with recall and reproduction. Apart from providing accurate responses to such questions, chatbots can be used by learners to gain a deeper understanding of any aspects of these concepts that they may not initially comprehend. The chatbots can be prompted by the learners to provide clarity on any aspects that may be unclear to them.

The findings show that GenAI chatbots do not find definitions and recall and reproduction questions challenging, and can accurately answer these questions, supporting the earlier literature. Chapagain et al. (2024) found that the chatbots were able to provide accurate responses to questions requiring the definition of scientific terms in physics. Additionally, López-Simó and Rezende (2024) found that ChatGPT was able to provide accurate and correct responses to questions requiring basic definitions and simple calculations in physics, as confirmed in this study. This study examines the integration of GenAI tools in physics education and suggests that AI tools are valuable for learners when practising recall and reproduction questions, including simple calculations. The next sub-section analyses how ChatGPT and Gemini responded to questions at Level 2, skills and concepts, using Webb's DoK framework.

5.2 Level 2 - Skills and Concepts

Sub-questions, such as 1.2, 2.2, 3.1.3, 4.4, 7.1, and 9.6, assessed skills and concepts (Table 1). Sample questions illustrating the chatbots' responses to questions assessing skills and concepts are shown in Figures 4 and 7.

Two trolleys, A and B, of masses of 1,5 kg and 2 kg respectively, are held in a stationary position on a straight, horizontal, frictionless track, with a compressed spring between them. The trolleys are released, and the spring takes t seconds to return to its natural length. The spring then falls to the ground. Trolley A moves to the left, while trolley B moves to the right and then up a frictionless inclined plane, rising to a maximum vertical height of 1,5 m, as shown in the diagram below.



Ignore the rotational effects of the wheels.

4.2 Calculate the speed of trolley B at the bottom of the inclined plane. (4)

4.3 For the t seconds that the spring takes to return to its natural length:

4.3.1 Calculate the change in momentum of trolley B (3)

4.3.2 Write down the change in momentum of trolley A (1)

4.4 Calculate the speed of trolley A after t seconds. (2)

Figure 4: **Questions on mechanics, assessing skills, and concepts (DBE, 2024)**

The response of the Google Gemini Deep Thinking model to question 4 (Figure 4) is shown in Figure 5. The examiners awarded Gemini full credit for question 4.2 for stating the conservation equation, substituting the correct values into the equation, and then providing the answer with the correct units. The average performance of the learners in 4.2 was 57%, with most candidates failing to write down the formula for the law of conservation of mechanical energy. Alternatively, they could have used the work-energy theorem or equations of motion. The question required an understanding of the principle of conservation of mechanical energy, then using an equation to solve the problem.

In question 4.3.1, the learners were required to understand that the initial velocity before the explosion was zero, and that the final velocity after the explosion was the answer of 4.2. Substituting this value into the equation for the change in momentum would yield the required result. The change in the momentum of trolley A would be the same as that of B but in the opposite direction. This was what learners were required to understand in order to answer 4.3.2 correctly. The last question, 4.4, was again an application of the concept of a change in momentum. While both ChatGPT (Figure 6) and Gemini (Figure 5) accurately answered all questions, demonstrating that the chatbots comprehended the information provided, the same could not be said about the learners regarding impulse and momentum. Only about half of the learners were able to attempt this question competently.

4.2 Calculate the speed of trolley B at the bottom of the inclined plane. (4)

Substitute into the conservation equation:

$\frac{1}{2}m_B v_B^2 + 0 = 0 + m_B g h_{\text{top}}$

Solve for v_B (speed of trolley B at the bottom):

$\frac{1}{2}m_B v_B^2 = m_B g h_{\text{top}}$

Notice that the mass m_B cancels out, which is typical for problems involving only gravitational potential energy and kinetic energy.

$v_B^2 = 2gh_{\text{top}}$

$v_B^2 = 2(9.8\ \text{m} \cdot \text{s}^{-2})(1.5\ \text{m})$

$v_B^2 = 29.4\ \text{m}^2 \cdot \text{s}^{-2}$

$v_B = \sqrt{29.4}$

$v_B = 5.4221...\ \text{m} \cdot \text{s}^{-1}$

Rounding to two decimal places, $v_B = \mathbf{5.42}\ \text{m} \cdot \text{s}^{-1}$.

4.3.1 For the t seconds that the spring takes to return to its natural length:

Calculate the change in momentum of trolley B (3)

Apply the change in momentum formula:

$\Delta p_B = m_B v_{fB} - m_B v_{iB}$

$\Delta p_B = (2\ \text{kg})(5.42\ \text{m} \cdot \text{s}^{-1}) - (2\ \text{kg})(0\ \text{m} \cdot \text{s}^{-1})$

$\Delta p_B = 10.84\ \text{kg} \cdot \text{m} \cdot \text{s}^{-1}$

- The change in momentum of trolley B is $\mathbf{10.84}\ \text{kg} \cdot \text{m} \cdot \text{s}^{-1}$ **to the right.**

4.3.2 Write down the change in momentum of trolley A (1)

- The change in momentum of trolley A is $\mathbf{10.84}\ \text{kg} \cdot \text{m} \cdot \text{s}^{-1}$ **to the left.**

4.4 Calculate the speed of trolley A after t seconds. (2)

Apply the change in momentum formula:

$\Delta p_A = m_A v_{fA} - m_A v_{iA}$

$-10.84 = (1.5)v_{fA} - (1.5)(0)$

$-10.84 = 1.5v_{fA}$

$v_{fA} = \frac{-10.84}{1.5}$

$v_{fA} = -7.2266...\ \text{m} \cdot \text{s}^{-1}$

Speed: Speed is the magnitude of velocity.

Speed of trolley A = $\mathbf{7.23}\ \text{m} \cdot \text{s}^{-1}$.

Figure 5: **Responses of Gemini to questions on mechanics on concepts, and skills**

The response of ChatGPT to question 4 is shown in Figure 6. Unlike most learners, ChatGPT received full credit. The chatbot was able to use the principle of the conservation of mechanical energy to calculate the speed of the trolley at the bottom of the incline, and then to calculate the change in momentum of trolley B. The chatbot deduced that the change in momentum of A would be the same as the change in momentum of B but the opposite in direction. The chatbot was also able to reason comprehensively to calculate the speed of trolley A after *t* seconds.

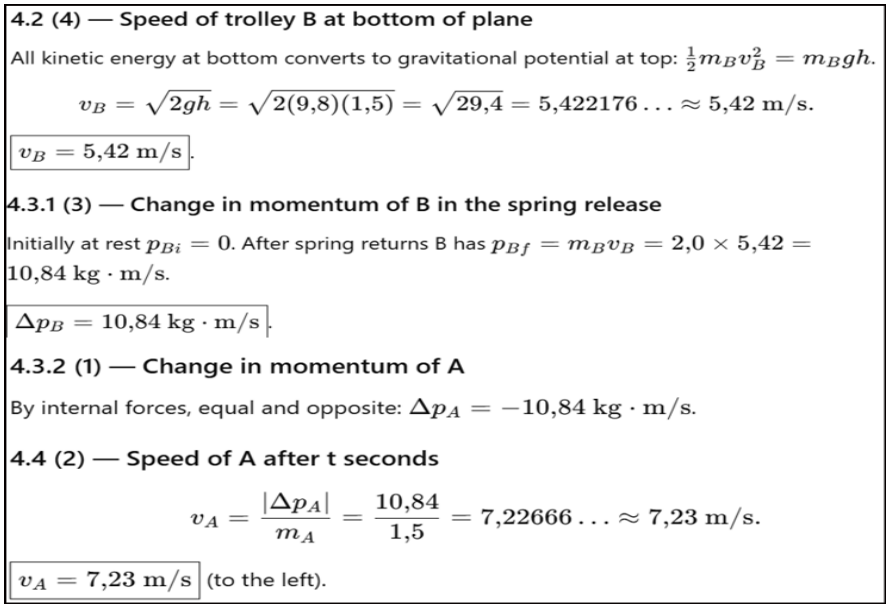


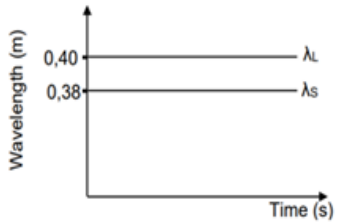
Figure 6: **Responses of ChatGPT to questions on mechanics, assessing concepts, and skills**

The results presented in this section indicate that GenAI models appear to outperform learners. This finding aligns with other scholars who have recently observed a significant improvement in the performance of LLMs, with similar results reported by Tong et al. (2024).

One of the questions where both chatbots performed poorly was related to the Doppler Effect (Figure 8). Learners outperformed the chatbots in this question with an average performance of 74% (Figure 1). Sub-questions 6.2 and 6.3.1 were classified under Level 2, concepts and skills. It was important for the candidates to realise that from the given information in the graph, the wavelength recorded by the listener was larger than that of the source. They should recall that wavefronts spread out and the observed wavelength increases (frequency decreases) if the source is moving away from the observer.

ChatGPT was able to provide an accurate response to this question, while Gemini failed to provide a correct answer. Learners performed at 70% for this question. Sub-question 6.3.1 required candidates to use the wave equation, $v = f.\lambda$. While ChatGPT obtained the appropriate value of the wavelength from the graph, Gemini used the wrong value of the wavelength for the sound from the source. Gemini's response was inaccurate while that of ChatGPT was correct, raising some doubt about the comprehension of Gemini of information in graphs and its claim of multimodal abilities.

A stationary listener, standing on the roadside, records the wavelength of the sound emitted by the siren of a police car travelling at a constant velocity. In the wavelength-time graph below, NOT drawn to scale, λL is the wavelength of the sound recorded by the listener and λS is the wavelength of the sound emitted by the siren. Take the speed of sound in air to be 343 $m \cdot s^{-1}$.



6.1 Name the phenomenon that explains why the wavelengths shown in the graph differ. (1)
6.2 Is the car moving TOWARDS or AWAY FROM the listener? Give a reason for the answer. (2)
6.3 Calculate the: 6.3.1 Frequency of the sound emitted by the siren (3) 6.3.2 Magnitude of the velocity of the car. (6)

Figure 7: **Questions 6.2 and 6.3.1 on the Doppler Effect, assessing concepts and skills**

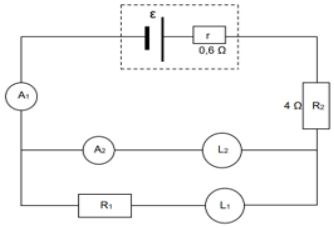
The analysis of the responses of GenAI chatbots to Level 2 questions of Webb's DoK framework, which assesses the comprehension of concepts and skills, suggests that the chatbots are capable of applying fundamental physics principles, particularly in mechanics. They were able to correctly formulate and substitute relevant conservation and kinematic equations, showing conceptual understanding and procedural accuracy that exceeded the learners' average performance. This study confirms and supports recent findings in physics education, which indicate that large language models now possess greater capabilities when it comes to accurately responding to physics problems (Chapagain et al., 2024; Tong et al., 2024).

Although the chatbots surpass the learners' question-answering performance at Level 2, they still have some limitations that require improvement. For the questions on the Doppler Effect, the failure to comprehend graphical information implies that Gemini still requires improvement to generate more accurate responses. The next sub-section analyses the responses of the GenAI chatbots to Level 3 questions and compares their performance to that of learners.

5.3 Strategic Thinking

There were a considerable number of cognitively demanding Level 3 questions, requiring learners to reason beyond recall and reproduction or familiarity with concepts and skills. For example, question 8, sub-questions 8.2.3, 8.2.4 and 8.3 (Figure 8), assessed the candidates' strategic thinking skills using Webb's DoK framework.

L_1 and L_2 are two light bulbs that have the following ratings: L_1: 36 W; 20 V, and L_2: 48 W; 32 V. The two bulbs are connected as shown in the circuit diagram below. The battery has an internal resistance of 0,6 Ω while the conducting wires and the ammeters have negligible resistance. R_1 and R_2 are resistors.



8.2 If both light bulbs operate as RATED, calculate the:

8.2.1 Reading on ammeter A_2 (3)

8.2.2 Reading on ammeter A_1 (3)

8.2.3 Resistance of resistor R_1 (4)

8.2.4 Emf of the battery (4)

8.3 Bulb L_1 burns out after a while. Assume that the resistance of bulb L_2 remains constant. Will bulb L_2 continue to glow after bulb L_1 burns out? Choose from YES or NO. Support your answer with a suitable calculation. (5)

Figure 8: **Questions on electric circuits, assessing strategic thinking (DBE, 2024)**

Although all questions on electric circuits at Level 3 required calculations, not all calculation questions in physics are Level 3. For example, simple calculations that require substituting given values into a formula (often provided on data sheets) are less cognitively demanding and can be regarded as Level 1 or 2 questions, depending on their level of complexity. Multistep calculations (López-Simó & Reze, 2024) which often require several equations, with some values provided in the problem and others not, are more cognitively demanding and can be classified as Level 2 or 3, depending on the degree of complexity required for the reasoning. This was the case with sub-questions 8.2.3, 8.2.4 and 8.3 (Figure 8).

Sub-question 8.2.3 required the candidates to calculate the resistance of resistor R_1 in the given circuit (Figure 8). The sub-question is a Level 3 question for several reasons. First, it is not a single-step question that requires substituting given values into a formula. Second, there are several different approaches that a candidate can follow to obtain the solution. The problem requires multiple steps and extensive reasoning to arrive at the result. For example, candidates could have determined V_{R1} first. This step required them to understand that the voltage across the resistors in parallel is the same but the voltage across resistors in series is divided among said resistors.

Candidates performed poorly in this question, obtaining around 50% (Figure 3). ChatGPT provided a comprehensive response with accurate reasoning. It determined the voltage across R_1, reasoned that the current through R_1 must be the same as through L_1, and then proceeded to use Ohm's law to find R_1. The response from Gemini was fraught with errors. It failed to determine the current through R_2 and the total current in the circuit. Although it obtained the accurate voltage across R_1, its application of Ohm's law yielded an incorrect response due to the errors in the preceding discussion.

Determining the EMF (sub-question 8.2.4) of the battery was also a Level 3 question. There were several ways to determine the EMF. One way was to find the voltage drop across R_2 using Ohm's Law. Candidates had to understand that the current through R_2 was the total current flowing in the circuit. They could then apply Kirchhoff's loop rule to determine the EMF. ChatGPT provided an accurate response, while Gemini did not. Sub-question 8.3 required candidates to determine whether bulb L_2 would continue glowing if bulb L_1 were to burn out. ChatGPT accurately reasoned that when L_1 burns out, all current would flow through L_2. It correctly calculated the current that would flow through L_2 (which was greater than 1.50A) by first determining the resistance of L_2, concluding that L_2 would continue to glow brighter. The calculations from Gemini had some errors, although it also concluded that L_2 would continue to glow.

The sub-questions on electric circuits at Level 3 in Webb's DoK framework required a multi-step approach, specifically conceptual reasoning that integrated multiple concepts and strategic thinking that involved selecting appropriate strategies. They needed to demonstrate a deeper understanding of fundamental principles such as Ohm's law and Kirchhoff's rule. Consequently, learners struggle with attempting such questions.

This study suggests that recent GenAI models have advanced to a level where they can exhibit comprehensive thinking, enabling them to solve complex physics problems that require multi-step reasoning, as demonstrated by ChatGPT. This suggests that the chatbots are approaching expert reasoning in physics. The less consistent accuracy of the responses from Gemini implies that GenAI systems do not yet possess the level of sophisticated reasoning required for problems that necessitate strategic thinking.

In terms of the taxonomy of errors displayed by GenAI chatbots, this study identified some apparent limitations when reading graphs, particularly in the case of Google Gemini. The study did not find evidence of failure by GenAI chatbots when choosing appropriate formulae for solving physics problems or using the correct units. The GenAI chatbots displayed proficiency in algebraic manipulations. The implications and limitations of this study are presented in the next section.

6. Implications and Limitations of the Study

The findings in this study indicate that generative AI tools, such as Gemini and ChatGPT, continue to evolve and become capable of providing accurate responses to physics questions on university entrance exams, opening up new possibilities. It builds upon earlier work in physics and chemistry education by researchers such as Demirci (2025) and Daher, Diab, and Rayan (2023), who found that GenAI chatbots were unable to accurately answer physics and chemistry questions on university entrance exams. The implications of the findings are that the accuracy of GenAI chatbots continues to improve, extending to assessment and curriculum design, educational policy, professional teacher development, teaching and learning, and future research.

This study is significant for practice in South Africa and beyond, particularly in the field of physics education. Regarding assessment and curriculum design, educators should reconsider the formative assessment tasks they require their learners to complete. If generative AI tools can provide accurate responses to standardised test items, then setting these types of exercises for homework may serve no purpose as they will use chatbots to answer the questions. It is suggested that when educators want to assess recall and reproduction, assessments should be in the form of short class tests.

Homework should focus on tasks that assess strategic thinking, such as analysis, synthesis, and interpretation, rather than concepts and skills, and the recall of information. Learners can utilize AI but they should be critical of AI-generated responses. They should also be aware of the limitations of AI (Tang et al., 2024). Assessment activities should focus on problem-solving, creativity, and collaboration between AI tools and learners (Khlaif et al., 2025).

There is a need for a review of educational policy regarding AI integration in schools to ensure that all learners have access to AI tools, that there is an incorporating of AI tools into formal curricula, and to ensure data privacy and the ethical use of AI (Zhao et al., 2024). In-service and pre-service teacher training should be designed to focus on integrating AI into pedagogical practices, ensuring the appropriate use of AI in lesson planning, teaching, learning, and assessment, while being aware of the current limitations (Seufert et al., 2021). Thus, it is suggested that GenAI tools, such as ChatGPT and Gemini, should be used as pedagogical assistants in classroom practice.

This study has some limitations. A major limitation is that the findings only apply to the period in which the study was conducted. GenAI tools are continuing to evolve at a rapid pace and are becoming increasingly advanced. As new models of AI tools emerge, the types of questions that this study reveals as problematic for AI tools to answer accurately may change. Future research should focus on investigating the accuracy of GenAI tools in STEM fields, including mathematics, biology, and chemistry, as well as the impact of integrating AI into academia on academic achievement and the learners' attitudes towards AI integration. This is to ensure the effective integration of GenAI.

7. Conclusion

This study has demonstrated that GenAI tools, such as ChatGPT and Gemini, continue to evolve. They are becoming more adept at providing accurate responses to physics questions that require higher cognitive thinking, specifically at level 3 according to Webb's DoK framework. This means that physics education at secondary school level can be transformed by integrating these tools. It is suggested that the integration of GenAI in high school curricula should be accompanied by the building of teacher capacity through professional development. When integrated appropriately into science education, GenAI can enhance inquiry-based learning, promote higher-order thinking, and improve academic performance.

8. Conflict of Interest

The author declares no conflict of interest.

9. Acknowledgements

The author would like to acknowledge the use of SPSS version 29 for statistical analysis and Grammarly for editing and writing this paper. Grammarly was used to help improve the language and grammar in the paper. The paper remains an accurate representation of the authors' work and intellectual contributions. The author expresses gratitude to the DBE examiners who participated in the study.

10. References

Ahmed, J., Nadeem, G., Majeed, M. K., Ghaffar, R., Baig, A. K. K., Shah, S. R., Razzaq, R. A., & Irfan, T. (2025). The Rise of Multimodal AI: A Quick Review Of GPT-4v and Gemini. *Spectrum of Engineering Sciences*, *3*(6), 778-786. https://thesesjournal.com/index.php/1/article/view/506/452

Al-Thani, S. N., Anjum, S., Bhutta, Z. A., Bashir, S., Majeed, M. A., Khan, A. S., & Bashir, K. (2025). Comparative performance of ChatGPT, Gemini, and final-year emergency medicine clerkship students in answering multiple-choice questions: implications for the use of AI in medical education. *International Journal of Emergency Medicine*, *18*, 146. https://doi.org/10.1186/s12245-025-00949-6

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability*, *15*(17), 12921. https://doi.org/10.3390/su151712921

Chapagain, P., Malakar, N., & Rimal, D. (2024). Can AI solve physics problems? Evaluating efficacy of AI models in solving higher secondary physics exam problems: A comparative study. *Journal of Nepal Physical Society*, *10*(1), 58-64. https://doi.org/10.3126/jnphyssoc.v10i1.72836

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, *8*, 75264-75278. https://doi.org/10.1109/access.2020.2988510

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., & Rosen, E. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*. arxiv.org/pdf/2507.06261

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. BMC Medical Research Methodology, 11(1). https://doi.org/10.1186/1471-2288-11-100

Demirci, N. (2025). How Successful Are Artificial Intelligence Chatbots on Higher Education Entrance Physics Exams in Turkey. *TOJET: The Turkish Online Journal of Educational Technology*, *24*(2). researchgate.net/profile/Neset-Demirci/publication/392059590_How_Successful_are_Artificial_Intelligence_Chatbots_on_Higher_Education_Entrance_Physics_Exams_in_Turkey/links/68319a696b5a287c304450a3/How-Successful-are-Artificial-Intelligence-Chatbots-on-Higher-Education-Entrance-Physics-Exams-in-Turkey.pdf

Department of Basic Education. (2024). Previous exam papers (Gr 10, 11 & 12). Pretoria. https://www.education.gov.za/Portals/0/CD/2024%20November%20past%20papers/Physical%20Sciences%20P1%20Nov%202024%20Eng.pdf?ver=2025-03-04-112701-620

Jere, S. (2025). Evaluating artificial intelligence large language models' performances in a South African high school chemistry exam. *EURASIA Journal of Mathematics,*

Science and Technology Education, *21*(2), em2582. https://doi.org/10.29333/ejmste/15932

Jere, S., & Mpeta, M. (2025). Integrating generative artificial intelligence chatbots into chemistry teaching: Impact of affective factors on engagement and conceptual understanding. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(10), em2713. https://doi.org/10.29333/ejmste/17077

Jere, S., Bessong, R., Mpeta, M., & Litshani, N. F. (2024). Exploring Pre-Service Teachers' Perceptions of ChatGPT Integration into Physical Sciences Teaching: A Case Study at a Rural South African University. *International Journal of Learning, Teaching and Educational Research*, *23*(11), 464-486. https://doi.org/10.26803/ijlter.23.11.24

Khlaif, Z. N., Alkouk, W. A., Salama, N., & Abu Eideh, B. (2025). Redesigning assessments for AI-enhanced learning: A framework for educators in the generative AI era. *Education sciences*, *15*(2), 174. https://doi.org/10.3390/educsci15020174

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, *15*(7), 5614. https://doi.org/10.3390/su15075614

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973-1018. https://doi.org/10.1007/s10639-022-11177-3

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. https://doi.org/10.2307/2529310

Liu, M., Okuhara, T., Dai, Z., Zhao, M., Yin, W., Okada, H., Furukawa, E., & Kiuchi, T. (2025). Large language models (GPT-5, Grok-4, Claude Opus 4.1, Gemini 2.5 Pro) achieved textbook-level accuracy on the Japanese medical licensing examination by 2025: A comparative study. *medRxiv*, 2025.2009. 2010.25335398. medrxiv.org/content/10.1101/2025.09.10.25335398v1.full.pdf

López-Simó, V., & Rezende, M. F. (2024). Challenging ChatGPT with different types of physics education questions. *The Physics Teacher*, *62*(4), 290-294. https://doi.org/10.1119/5.0160160

Marzano, R. J., & Kendall, J. S. (2006). *The new taxonomy of educational objectives*. Corwin Press. ifeet.org/files/The-New-taxonomy-of-Educational-Objectives.pdf

Matejak Cvenic, K., Planinic, M., Susac, A., Ivanjek, L., Jelicic, K., & Hopf, M. (2022). Development and validation of the Conceptual Survey on Wave Optics. *Physical Review Physics Education Research*, *18*(1), 010103. https://doi.org/10.1103/physrevphyseducres.18.010103

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017-1054. https://doi.org/10.1177/016146810610800610

Newton, P. M., Summers, C. J., Zaheer, U., Xiromeriti, M., Stokes, J. R., Bhangu, J. S., Roome, E. G., Roberts-Phillips, A., Mazaheri-Asadi, D., & Jones, C. D. (2025). Can ChatGPT-4o really pass medical science exams? A pragmatic analysis using novel questions. *Medical Science Educator*, *35*(2), 721-729. https://doi.org/10.1007/s40670-025-02293-z

OpenAI. (2025a). *GPT-5 System Card*. OpenAI. Retrieved 11 August 2025 from https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf

OpenAI. (2025b). *Introducing GPT-5*. OpenAI. Retrieved 11 August from https://openai.com/

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI*, *4*(4), 949-969. https://doi.org/10.3390/ai4040048

Polverini, G., & Gregorcic, B. (2024). How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics*, *45*(2), 025701. https://doi.org/10.1088/1361-6404/ad1420

Rane, N., Choudhary, S., & Rane, J. (2024). Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *Journal of Applied Artificial Intelligence*, *5*(1), 69-93. https://doi.org/10.48185/jaai.v5i1.1052

Seufert, S., Guggemos, J., & Sailer, M. (2021). Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior*, *115*, 106552. https://doi.org/10.1016/j.chb.2020.106552

Tang, K.-S., Cooper, G., Rappa, N., Cooper, M., Sims, C., & Nonis, K. (2024). A dialogic approach to transform teaching, learning & assessment with generative AI in secondary education: A proof of concept. *Pedagogies: An International Journal*, *19*(3), 493-503. https://doi.org/10.1080/1554480x.2024.2379774

Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan, S., & Liu, Q. (2024). Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pacific Education Review*, *25*(5), 1379-1389. https://doi.org/10.1007/s12564-023-09913-6

Tschisgale, P., Maus, H., Kieser, F., Kroehs, B., Petersen, S., & Wulff, P. (2025). Evaluating GPT-and reasoning-based large language models on Physics Olympiad problems: Surpassing human performance and implications for educational assessment. *Physical Review Physics Education Research*, *21*(2), 020115. https://doi.org/10.1103/6fmx-bsnl

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer* (pp. 23-65). Springer. https://doi.org/10.1007/978-1-4020-6710-5_3

Webb, N. L. (2002). Depth-of-Knowledge Levels for Four Content Areas. *Language Arts*. ossucurr.pbworks.com/w/file/fetch/49691156/Norm web dok by subject area.pdf

Woitkowski, D. (2020). Tracing physics content knowledge gains using content complexity levels. *International journal of science education*, *42*(10), 1585-1608. https://doi.org/10.1080/09500693.2020.1772520

Xuan-Quy, D., Ngoc-Bich, L., Xuan-Dung, P., Bac-Bien, N., & The-Duy, V. (2023). Evaluation of ChatGPT and Microsoft Bing AI chat performances on physics exams of Vietnamese national high school graduation examination. *arXiv preprint arXiv:2306.04538*. arxiv.org/pdf/2306.04538

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhao, J., Chapman, E., & Sabet, P. G. (2024). Generative AI and educational assessments: A systematic review. *Education Research and Perspectives*, *51*, 124-155. https://doi.org/10.70953/erpv51.2412006