**Implementing ChatGPT as Tutor, Tutee, and Tool in Physics and Chemistry**

Nawee Jaroonchokanan and Chitnarong Sirisathitkul*

School of Science, Walailak University, Nakhon Si Thammarat, 80160, Thailand

*Corresponding author's email: schitnar@mail.wu.ac.th

**Received**: Jun 06, 2024 **Revised**: Sep 16, 2024 **Just Accepted Online**: Sep 23, 2024 **Published**: Xxx



Please cite this article as:

N. Jaroonchokanan, C. Sirisathitkul, (2024) Implementing ChatGPT as Tutor, Tutee, and Tool in Physics and Chemistry. **Substantia**. *Just Accepted*. DOI: 10.36253/Substantia-2808



## Abstract

In the age of modern technology, generative artificial intelligience-powered chatbots offer a variety of uses for different purposes. Undoubtedly, ChatGPT is one of the most widely used chatbots in science education. In this paper, we review the implementations of chatbots, focusing particularly in teaching and learning physics and chemistry. Their roles in the context of science education are classified as tutee, tutor, and tool. We found the development of ChatGPT to be quite impressive. As a tutee, the latest version of ChatGPT is a fast learner, capable of passing standard tests and providing accurate scientific answers using approaches like Chain-of-Thought and Socratic-style dialogue. As a

tutor, it can help students learn through classroom teaching techniques such as scaffolding and enhance critical thinking by acting as a personal tutor that offers instantaneous feedback. As a tool, ChatGPT can assist in reviewing students' handwritten homework, drafting scientific writing, and generating code for science programming. Although ChatGPT offers many benefits, it can sometimes provide inaccurate information, necessitating human oversight in science education. Importantly, students should be taught to critically assess the responses provided by ChatGPT and understand its ethical use to ensure effective utilization.

**Keywords:** Artificial intelligence, ChatGPT, Physics, Chemistry, Science education

## 1. Introduction

Chatbots and other artificial intelligence (AI) tools are shaping the future across various domains, including education [1-3]. The successful integration of AI into education depends on collaboration among educators and policymakers. In science education, ChatGPT has the potential to enhance student learning and improve educational outcomes [4]. Educators can use ChatGPT to supplement traditional teaching methods, providing students with additional resources to support their learning and engage them further in the subject matter. However, ChatGPT can sometimes provide misleading information and be susceptible to negative and unethical outcomes [1-6]. Therefore, the challenge for educators is to leverage chatbots and other AI tools to maximize student learning efficiency and prepare them for their future professional lives [4]. The effectiveness of chatbots depends on several factors, including the academic discipline and specific domain [6,7].

This article reviews the implementation of ChatGPT in physics and chemistry education. To provide context, the development of chatbot technology is briefly summarized in the following section.

The subsequent sections (3 and 4) are structured around the three roles of chatbots: tutor, tutee, and tool. The roles of ChatGPT as a tutee and tutor in physics and chemistry are particularly intriguing as they reflect the chatbot's ability to comprehend concepts and deliver accurate information. Given the importance of reliability in education, ChatGPT must be capable of providing precise and trustworthy knowledge that can be effectively utilized in teaching physics and chemistry. Following this discussion, ChatGPT's performance is compared with other AI chatbots to offer broader insights. The article concludes by addressing ethical considerations and providing an outlook on the future of AI in education.

## 2. Brief History of Chatbot Development

Chatbots, or conversational agents, are programs designed to interact with humans via text- or voice-based interfaces. These systems are created and trained to comprehend user input, identify the purpose of the dialogue, and generate human-like responses. Chatbot development can be broadly categorized based on either pattern-matching algorithms or machine learning models, the latter of which is foundational to modern AI-driven chatbots. The history of chatbots dates back to 1966 with Eliza, one of the first conversational agents created by Joseph Weizenbaum [8]. Eliza simulated a psychotherapist by engaging in simple conversations based on pattern matching. In 1972, Parry was developed as a more advanced chatbot, employing a system of assumptions and simulated human reactions to replicate the thought patterns of someone with paranoid schizophrenia [9]. Both chatbots, while rudimentary by today's standards, laid the groundwork for future developments in AI dialogue systems.

Schobel et al. summarized the evolution of chatbot technology into five waves: the zero-hour wave, the explore wave, the kick-off wave, the hype wave, and the AI wave [10]. During the explore wave, the integration of Natural Language Processing (NLP) became a key focus. NLP enabled chatbots to not only understand but also analyze and interpret natural human language, vastly improving their conversational capabilities. Significant advancements in this period included Jabberwacky in 1988 and A.L.I.C.E in 1995, and the term AI was used firstly used for the chatbots with the former [9]. The kick-off wave was marked by the debut of IBM Watson in 2006, introducing AI-driven chatbot technology into the mainstream with real-world applications. This demonstrated the potential of chatbots to handle large-scale knowledge queries and complex problem-solving tasks. The hype wave saw the mass adoption of chatbots for a variety of consumer-facing roles. This wave introduced some of the most recognizable AI-driven assistants, including Apple's Siri (2011), Amazon Alexa (2014), Microsoft Cortana (2014), and Google Assistant (2016). These systems moved beyond simple Q&A and evolved into fully integrated personal assistants. They could understand complex voice commands and perform tasks ranging from setting reminders to controlling smart home devices. By this time, chatbots were widely implemented across sectors such as marketing, customer support, healthcare, education, and entertainment.

The launch of ChatGPT by OpenAI in late 2022 initiated the AI wave, which has sparked unprecedented global interest in generative AI technologies [10]. Unlike its predecessors, ChatGPT was built using a large language model (LLM), specifically GPT-3.5, and later GPT-4, allowing it to generate coherent and contextually relevant text responses across a wide array of topics. In response, major technology firms accelerated their AI chatbot development, leading to the introduction of Bard by Google and Bing Chat by Microsoft in 2023. These new generative AI chatbots offer capabilities such as answering complex questions, explaining scientific principles, summarizing texts, and even

generating academic essays or code, positioning them as indispensable tools for both casual users and professionals alike. As AI continues to evolve, the role of chatbots in education, research, and everyday life will likely expand, with generative models like ChatGPT serving as critical companions in learning, problem-solving, and content creation.

## 3. ChatGPT in Physics Education

### 3.1 Tutee

Many researchers and educators investigate and train ChatGPT to test its knowledge of physics and train it to perform better. In the following paragraphs, we will briefly describe several studies considering ChatGPT to be a tutee.

One such investigation by Wang scrutinizes ChatGPT's proficiency in solving physics problems [11]. Initially tasked with resolving the motion of a body on a frictionless incline, ChatGPT adeptly interprets the query and correctly identifies physical variables such as angle, gravitational acceleration, force, and mass. However, the acceleration sign is incorrect, which could cause coordinate system confusion. The author then requests that ChatGPT simulate a model of the situation. The results are extremely incredible. ChatGPT can generate a simulation to resolve this problem correctly. The author challenges ChatGPT with a more difficult question about the Stern-Gerlach experiment. The results showed that ChatGPT was unable to answer the question accurately. Consequently, while ChatGPT appears capable of solving simple physics problems, such conceptual topics remain challenging to tackle.

Interestingly, Kortemeyer delves into whether ChatGPT could pass an introductory physics course [12]. Employing a multifaceted assessment approach that includes multiple-choice questions, homework assignments, clicker questions, programming exercises, and exams, ChatGPT exhibits varying degrees of competence. The results found that ChatGPT scored 18 out of 30 points on the Force Concept Inventory (FCI) [13]. For homework, this allowed it to make 5 attempts on topics including trajectory motion, friction, thermodynamics, capacitance, and special relativity, covering a total of 76 homework problems. It was found that ChatGPT frequently made numerical errors and was unable to correct these mistakes even after they were pointed out. Despite this, ChatGPT solved 55% of the homework problems using an average of 1.88 attempts per problem. For clicker questions, ChatGPT correctly answered 10 out of 12 questions, a score better than most students in the actual course. Additionally, ChatGPT was also assigned to write Python code related to an anharmonic oscillator for programming exercises. The author noted that ChatGPT performed much better than many students in the course. ChatGPT scored 14 out of 30 points on the midterm and final exams. The results showed that five of the incorrect answers were due to numerical calculation errors. Although ChatGPT could answer correctly after reverse token verification, full credit was not awarded. Considering the grading policies (20% homework, 5% clicker questions, 5% programming exercises, and 70% exams), the author determined that ChatGPT would receive a course grade of 1.5. However, if it had been more accurate in numerical operations, the course grade could have been 2, reflecting a 60% performance. Hence, ChatGPT can pass an introductory physics course.

Another perspective to evaluate the performance of ChatGPT is to use it to write short essays. Yeadon et al. examined ChatGPT's ability to write essays on physics concepts and historical and philosophical themes, such as "Is physics based on facts that follow from observations?" and "How did natural philosophers' understanding of electricity change during the 18th or 19th centuries?" [14]. The

exam comprised five short-form essay questions, each limited to 300 words, with a maximum score of 100 per question. These essays were graded by five different markers, and ChatGPT achieved an average score of 71±2%, which is high enough to qualify for a First Class grade, the highest distinction available at UK universities. The plagiarism rates, checked by both Grammarly and Turnitin, were found to be 2±1% and 7±2%, respectively. The authors pointed out that ChatGPT poses a significant threat to the integrity of short-form essays as an assessment method in physics courses.

In term of educational learning objectives, López-Simó and Rezende explored ChatGPT's capability to solve five types of physics questions related to Bloom's taxonomy: dictionary definitions, simple calculations, multistep calculations, reasoning problems, and Fermi problems [15]. Using GPT-3 for its broader accessibility, each question was asked ten times in separate windows. ChatGPT performed well on dictionary definitions (e.g., Newton's second law) and simple calculations, answering correctly 7 out of 10 times. However, it failed to solve multistep calculations correctly even once and showed a preference for certain options in reasoning problems. For Fermi problems, which require interpretation and informed reasoning, ChatGPT provided answers closer to the expected order of magnitude but with inconsistency. The authors concluded that ChatGPT is still unreliable as a self-help tool for introductory physics. However, the authors suggested leveraging its limitations to engage students in critical discussions, enhancing their understanding of complex physics problems.

The subsequent version of ChatGPT, ChatGPT-4, demonstrates a remarkable ability to tackle advanced physics problems. This capability was thoroughly investigated across various physics domains. Dazhen et al. investigated the performance of ChatGPT-4 in solving physics conceptual understanding and reasoning problems, encompassing mechanics and electromagnetism [16]. The research utilized two primary problem sets in physics education, namely the FCI and the Conceptual Survey of Electricity and Magnetism (CSEM) [17]. The authors selected sixteen multiple-choice

questions to assess ChatGPT's abilities. Impressively, ChatGPT answered all questions correctly, whereas the average score for the FCI test among 415 university students in 2018 was 56.3%. For CSEM, the average score among 9,905 students who completed an electromagnetic course was 44.6%.

The author conducted another test that utilized primitive physics problems (PPPs). This evaluates various knowledge representations such as abstraction, assignment, image, physics, methodology, and mathematics, as proposed by Xing et al. [18]. However, ChatGPT struggled with image representation, which was subsequently excluded from the evaluation. In this assessment, the authors compared ChatGPT's performance with that of 388 middle school students across four questions following PPPs. Three researchers independently scored the physics reasoning problems based on the aforementioned representations. The results revealed that ChatGPT achieved a significantly higher score of 87.5% overall, compared to the middle school students' average score of 23.32%. A similar study conducted by West compared the performance of ChatGPT-3.5 and ChatGPT-4 on a modified version of the FCI, consisting of 23 usable questions [19]. The results showed that ChatGPT-3.5 answered fifteen questions correctly (65%), performing just below or slightly above the average student. Impressively, ChatGPT-4 answered 22 out of the 23 questions correctly. The authors noted that ChatGPT-4's ability to engage in metaphor and utilize multiple representations distinguishes it from novices and aligns more closely with expert-level understanding.

In addition, Kumar and Kats conducted a comparative analysis of ChatGPT's performance in solving 13 electromagnetism problems of differing complexities from an introductory course at the Department of Electrical and Computer Engineering, University of Wisconsin-Madison [20]. The study included different versions of ChatGPT, namely ChatGPT-3.5, ChatGPT-4, and ChatGPT-4 with a Code Interpreter (4/CI). ChatGPT-4 introduces a range of plugins, such as WolframAlpha and Code Interpreter, enabling it to execute calculations and generate Python code for plotting graphs. The study

found that ChatGPT-4 with Code Interpreter (4/CI) outperformed its predecessors in resolving all 13 electromagnetic problems. ChatGPT-4/CI consistently demonstrated high accuracy in solving these problems, including tasks like integrating charge density in Cartesian coordinates and calculating electric fields in dielectrics. However, it should be noted that ChatGPT-4/CI occasionally identified its own errors when prompted for explanations. Furthermore, it exhibited a stochastic nature when solving vector calculus problems, occasionally yielding different answers, both correct and incorrect.

In contrast to studies with a small number of questions, Yeadon and Halliday assessed ChatGPT's performance on Durham University physics exams [21]. They analyzed 42 exam papers from 10 different physics courses, spanning 2018 to 2022, with 593 questions, including traditional and COVID-era adaptive formats. During the COVID period (2021-2022), open-book exams were allowed. ChatGPT-4 outperformed ChatGPT-3.5, achieving average scores of 49.4% compared to 38.6%. Pre-COVID, ChatGPT-4 scored 50.8% while ChatGPT-3.5 scored 41.6%, and post-COVID scores dropped to 47.5% and 33.6%, respectively. ChatGPT-4's performance showed minimal variation between the two periods. This demonstrates ChatGPT-4's improved performance.

The ability to interpret graphs is also crucial in physics. ChatGPT-4 was enhanced to process image data, a feature examined by Polverini and Gregorcic [22]. The authors tested this capability using the Test of Understanding Graphs in Kinematics (TUG-K), a multiple-choice assessment widely used to evaluate students' comprehension of one-dimensional motion graphs. Each of the 26 survey items was uploaded as a "png" screenshot, and the test was submitted to ChatGPT 60 times across 1,560 separate chats. The average score was 10.85 points (41.7%), a performance comparable to that of high school students. The authors noted ChatGPT's tendency to answer certain items correctly or incorrectly consistently. Additionally, the authors cautioned against relying on ChatGPT for tutoring

students with typical learning difficulties, highlighting the need for careful consideration in its educational application.

Furthermore, Polverini and Gregorcic demonstrated techniques for improving ChatGPT's conceptual physics task performance [23]. They highlighted the Chain-of-Thought (CoT) strategy, which involves prompting the AI to think step-by-step. For example, asking, "If two bodies with different masses have the same kinetic energy, which one has the largest momentum?" without CoT led to incorrect answers four out of eight times. However, framing the question to include reasoning improved accuracy, for example, "If two bodies with different mass have the same kinetic energy, which one has the largest momentum? Provide your reasoning first and only then provide the answer," with correct answers seven out of eight times. This illustrates the potential of CoT prompts in enhancing AI performance. The authors recommended dialogue-based approaches, such as asking, "Isn't there another way to do it?" to train ChatGPT and help students use AI chatbots effectively in physics education.

### 3.2 Tutor

As ChatGPT and other chatbots are anticipated to become indispensable tools for academic training and assessment in future education, the role of ChatGPT as a physics tutor is discussed. Traditionally, students do not receive instantaneous instructor feedback, but modern technology like ChatGPT can significantly enhance student learning. In this section, we summarize methods for using ChatGPT in the classroom.

Liang et al. investigated ChatGPT's potential in physics education [24]. They found that ChatGPT can provide scaffolding by generating step-by-step guidance. For instance, ChatGPT can explain projectile motion by breaking down complex movements into components and offering detailed

explanations. It can also generate questions and hints to assess students' understanding and summarize variables from questions in a table, making problem-solving easier. However, it sometimes makes errors when judging vector directions and calculations. ChatGPT can permute variables to create new computation problems, helping instructors prepare homework materials with commands like "Permute the physics variables and give me another problem: [input problem]." Additionally, the authors suggest using commands to enhance physics study and motivation, such as "Please tell me some stories about physicist [name]" and "Physics concepts are applied in various real-world scenarios. Please provide examples of how [concept name] is applied." Despite these benefits, the study's limitation is that it has not yet been applied in real classrooms and lacks empirical support.

Another study claimed that ChatGPT can encourage critical thinking. This experiment was conducted by using high school students as samples. Bitzenbauer uses ChatGPT to teach quantum physics to 53 secondary school students [25]. The author first instructs students to ask ChatGPT questions such as "What is a photon?" and "What kind of particle is a photon?". After obtaining a response from ChatGPT, students can double-check the answer by looking for relevant answers to compare errors provided by ChatGPT, which they have already studied in class. This allows students to think about and become aware of the ChatGPT answer. This refers to critical thinking ideas such as generating conclusions based on facts, making decisions, or forming perceptions about something. Students were assigned to discuss ChatGPT-generated statements in pairs. In this phase, students could look through textbooks and scientific papers to check and revise their ChatGPT answers before debating them. This helps to cultivate the habit of verifying and sharing discoveries. In the last stage, the author utilizes ChatGPT to generate conceptual questions to assess students. This will allow students to review the topics they discussed with their friends. Finally, the authors compare a questionnaire from before and after the course regarding perspectives on learning with ChatGPT.

Overall, the average student's rate of agreement after the class is greater than before the class. The question "we should all learn to incorporate ChatGPT in our lives" has a substantial favorable impact on student perceptions of ChatGPT. However, "We can use ChatGPT even if we do not understand how it works" lowers after the class, indicating that students are more aware of ChatGPT results.

Alneyadi and Wardat utilized ChatGPT to teach the concrete idea of electromagnetism [26]. The authors employ ChatGPT to teach electromagnetics to eleventh-grade students, using 58 students for experimental and 64 students for control. In their experiment, the authors placed the students into two groups: those who used ChatGPT and those who did not. This experiment lasts four weeks. The data in this study is separated into two categories: quantitative and qualitative data. The electromagnetism exam is a part of quantitative data that contains a pretest and a post-test. It consists of 25 multiple-choice questions based on Bloom's taxonomy, including recall, comprehend, apply, and analyze. Following data collection, the authors apply the t-test to determine whether there is a significant difference in average pretest and posttest scores for both the experimental and control groups. The results revealed that the experimental group's mean post-test score was higher than that of the control group.

Open-ended survey questions by interviewing record audio were used to gather qualitative data. Authors conducted interviews covering six themes related to students' perceptions of ChatGPT: (i) ChatGPT as a Learning Tool: Students noted its usefulness in providing instant answers and explanations, particularly aiding in overcoming language barriers. (ii) Impact on Performance: While students appreciated its assistance with visual aids and homework problem-solving, some expressed concern over the lack of grading feedback. (iii) Comfort with Using ChatGPT: Overall, students reported feeling comfortable, though a few voiced occasional hesitations or frustrations with the technology. (iv) Differences in ChatGPT Use: Male students tended to use ChatGPT for quick answer

confirmations, whereas females utilized it for deeper understanding, investing more time in its use. (v) Suggestions for Improvement: Students suggested enhancements such as accent and dialect recognition, more visual aids and animations, and faster response times to questions. (vi) Recommendation to Other Students: Students indicated they would recommend ChatGPT to peers, underscoring its perceived value in aiding learning.

In higher education, ChatGPT is still useful for assessing students' understanding of many physics ideas. Dahlkemper et al. conducted an experiment to assess first-year students' understanding of physics principles and their attitudes regarding AI [27]. This contains students from two university physics courses totaling 95 people. The authors initially collected general data on students' perceptions of using ChatGPT; the findings revealed that the majority of students (84%) had heard of ChatGPT, but only half had used the chatbot. Furthermore, 74% of respondents say they would never employ ChatGPT in physics. Following that, the formal experiment was performed. The author tasked students with evaluating the performance of four physics responses across three topics: rolling motion, waves, and fluid dynamics, generated by ChatGPT. Notably, one response in each topic was provided by experts. The familiarity level varied among the topics: rolling motion, being the easiest, was previously assigned as homework; waves were studied a few weeks prior, while fluid dynamics, being the least familiar, required application of knowledge from various fields and was considered the most challenging.

Initially, students self-assessed their performance on a scale of 0 to 6 without attempting the problems. It was found that a decrease in average score was observed across the less familiar topics. The next step involved assessing the answers labeled as both ChatGPT and expert responses, followed by students evaluating both the scientific accuracy and linguistic quality. A two-way repeated-measures analysis of variance was conducted to compare the average scores between two groups: scientific

accuracy and linguistic quality. The analysis revealed significant differences in assessing ChatGPT responses across the topics. In the case of rolling motion and waves, students rated the expert answer highest for both scientific accuracy and linguistic quality. In contrast, for fluid dynamics, the expert solution was rated significantly higher in scientific accuracy, while linguistic quality did not differ significantly. This study sheds light on student perceptions of ChatGPT responses and the factors influencing their assessments.

Ding et al. also conducted a study involving 40 college-level students to explore their perceptions of using ChatGPT [28]. The students were tasked with regaining lost exam credits by interacting with ChatGPT on topics related to light and radioactivity. Over 1.5 weeks, students posed questions to ChatGPT, reviewed its responses, and assessed their validity, providing reasons for their decisions. If they disagreed with ChatGPT's answers, they had to argue their points and repeat the process. Following this activity, the students were surveyed about their experience. During the task, out of 362 questions asked, ChatGPT answered 85% correctly. However, upon further questioning, it revised its responses, correcting itself from incorrect to correct answers in 7 instances but also changing from correct to incorrect in 34 instances. Additionally, the authors employed K-means clustering to categorize students into three groups based on their level of trust in ChatGPT's responses: the trust group, partial trust group, and distrust group. ANOVA and MANOVA analyses were used to examine the statistical significance among these groups and assess differences in perception. The findings revealed that nearly half of the students trusted ChatGPT's answers regardless of their accuracy, perceiving it as a knowledgeable machine. These students found ChatGPT easy to use and expressed a greater likelihood of using it in the future than partial trust and distrust groups.

### 3.3 Tool

The literature discussed presents diverse viewpoints regarding the effectiveness of ChatGPT as both a physics tutee and tutor. In this section, we introduce additional literature advocating ChatGPT as a valuable tool for educators.

Gregorcic and Pendrill conducted an experiment wherein basic physics questions were posed to ChatGPT [29]. For instance, the question “A teddy bear is thrown into the air. What is its acceleration at the highest point?" yielded inconsistent and sometimes incorrect responses from ChatGPT. Despite providing some correct information, its answers often contradicted themselves, indicating a lack of coherence in understanding concepts like net force. Despite its linguistic prowess, ChatGPT struggled to recognize and rectify its own contradictions. However, the authors suggest potential applications in education, particularly in teacher training as a tool, where it could aid in recognizing and interpreting problematic argumentation. Subsequently, Gregorcic et al. reported that ChatGPT-4 demonstrated improved performance through repeated questioning, accurately answering all queries [30]. When tasked with describing graphs, ChatGPT4 exhibited proficiency, albeit with occasional unclear expressions regarding graph slopes. Notably, this version of ChatGPT showed the ability to detect inconsistencies in its responses and could provide correct answers with prompting. The authors concluded that ChatGPT-4 holds promise as a training tool for teaching physics through Socratic dialogue, based on both their direct experiences and insights from their pilot study.

In addition, one of the intensive tasks for physics teachers is grading students’ work. When it comes to physics problems that require derivations, this can become a significant workload, demanding both time and effort. Fortunately, technology nowadays can help address these challenges. Kortemeyer examines an AI-assisted workflow to grade handwritten physics derivations using MathPix and GPT-4 [31]. The process begins with scanning handwritten papers into PDF files using the smartphone app Scanner Pro, which are then transcribed into LaTeX using MathPix. The output from MathPix is

subsequently refined using GPT-4. The author evaluates the effectiveness of ChatGPT in grading electricity problems compared to human graders. The scores range from 0 (worst) to 4 (best) based on a rubric assessing the correctness of approach, symbolic derivations, numerical results, and straightforwardness. ChatGPT has demonstrated considerable potential for grading student work. While AI-assigned grades show a strong correlation to manually assigned grades ($R^2$ = 0.84), they are currently unreliable enough for summative assessments with limits and errors when conducting symbolic and numerical computations. However, it is reliable enough to help human graders by sorting or grouping answers and offering preliminary grades.

## 4. ChatGPT in Chemistry Education

### 4.1 Tutee

Much like those in physics, chemistry educators and researchers have investigated ChatGPT's understanding and accuracy in responding to various topics. Clark et al. discovered that ChatGPT's accuracy significantly varied across topics [32]. It performed well in pH calculations for strong acids and bases but struggled with more complex problems like titrations and aqueous salts. Unlike students, the chatbot avoided heuristic errors but made uncommon mathematical mistakes. Leon and Vidhani noted that while ChatGPT can provide correct answers within a given context, it often has difficulty verifying the computational or analytical accuracy of those answers [33]. Fergus et al. reported that

ChatGPT's responses to chemistry assessments were generally well-written but varied in quality [34]. The chatbot had difficulties with application and interpretation questions, especially those involving non-text information. This observation was consistent with Clark, who found that ChatGPT could identify concepts in closed-response questions with significant chemical symbolism but performed below the class average in problem-solving [35]. For open-response questions, ChatGPT demonstrated strong language processing abilities, performing better on questions that could be solved with generalizable information rather than specific skills taught in lectures. However, incorrect responses and flawed explanations often seemed logically sound and persuasive to students.

Nascimento and Pimentel highlighted ChatGPT's low accuracy in converting SMILES representations into compound names and vice versa, with errors such as missing or adding methyl groups, including nonexistent atoms, confusing regular cyclic and aromatic compounds, or misunderstanding isomers [36]. The chatbot also had difficulty with the most current and robust string representation. Daher et al. used the theoretical framework encompassing transfer, depth, predict/explain, problem-solving, and translation to evaluate ChatGPT's conceptual understanding in the material science domain [37]. They found significant difficulties in conceptual knowledge across various categories, particularly in representations and depth, hindering effective knowledge transfer.

### 4.2 Tutor

While chatbots should not be solely relied upon for providing answers or explanations to students due to their shortcomings noted in sections 2.1 and 3.1, they have significant potential as teaching assistants to complement educators' efforts. When financial limitations prevent institutions from hiring multiple teachers, chatbots provide a cost-effective solution to meet the individual needs of students. According to Alasadi and Baiz [38], AI-assisted teaching benefits students by providing

additional support and personalized attention. Chatbots can function as virtual co-teachers, assisting in evaluating student progress, offering tailored feedback, and delivering targeted interventions. This allows human educators to focus on fostering meaningful interactions and deepening students' understanding of the subject. Guo and Lee observed significant improvements in students' confidence to ask insightful questions, analyze information, and understand complex concepts, thanks to ChatGPT's ability to present diverse perspectives and challenge existing thought processes [39]. Students also reported using ChatGPT more frequently to enhance their critical thinking skills and expressed a willingness to recommend it to others. Exintaris et al. described a classroom activity combining metacognitive scaffolding, problem-solving practice, and critiquing ChatGPT-generated solutions [40]. This approach showed that students engaged with metacognition as a key part of their problem-solving toolkit and appreciated the collaborative nature of the exercise. They also identified errors and flaws in the provided incorrect solutions, although to varying degrees.

### 4.3 Tool

Chatbots have gained popularity as tools for academic writing due to their abilities to generate ideas, draft content, edit, and proofread. West et al. analyzed the strengths and weaknesses of laboratory reports generated by ChatGPT [41]. Rojas et al. noted that although students found ChatGPT helpful for scientific writing, they were reluctant to use it to generate entire texts [42]. Clark et al. discovered that students could distinguish between essays on sustainability written by ChatGPT and those written by humans [43]. While students' essays contained more scientific reasons and chemistry concepts, they were impressed by ChatGPT's ability to discuss sustainability solutions, policies, and

practices. Desaire et al. pointed out that manuscripts generated by ChatGPT are likely to be detected by chemistry journals [44].

Beyond writing assistance, chemistry researchers recognize the potential of chatbots in laboratory research and design. Araujo and Saude showed that ChatGPT could conceptualize problems and laboratory activities accessible to chemistry students, although the accuracy and safety of these activities require human oversight [45]. Scoggin and Smith examined ChatGPT's ability to help students generate experimental designs based on general chemistry textbook questions, noting that success depended on the clarity of the questions [46]. Chatbots also show promise in designing chemical reactions. Zheng et al. used ChemPrompt Engineering to train ChatGPT to text-mine peer-reviewed articles on Metal-Organic Frameworks (MOFs) [47]. ChatGPT could then answer questions about synthesis procedures, identify critical factors in MOF crystallization, and predict experimental outcomes. Mahjour et al. demonstrated that ChatGPT could formulate reaction arrays for common pharmaceutical reactions, with these results usable as inputs for management software like Phactor, enabling automated execution and analysis of assays [48]. Kong et al. employed ChatGPT to create interactive learning environments and simulate real-world engineering thinking processes in distillation column design for undergraduate mass-transfer courses [49]. Hasrod et al. used human prompts and an error message feedback loop with ChatGPT to generate working code for a graphic user interface (GUI) to predict sulfate levels in acid mine drainage [50]. This template allows students to create their GUIs for codes or models developed during their studies, demonstrating the potential for augmenting analytical data to infer or approximate non-directly analyzable parameters.

## 5. Comparison of the Performance of Various Chatbots in Science, Engineering, and Medical Education

Chatbots offer valuable educational information across various fields, serving as tutee, tutor, and tool—key roles in advancing educational technology. Apart from reviewing ChatGPT, we also compare its performance with other chatbots. To offer diverse perspectives, we include literature from various fields such as science, engineering, and medical education, as there are few comparative studies on chatbots in physics and chemistry.

In a study by Dos Santos [51], ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard were compared to investigate their effectiveness. The author posed questions related to motion and energy concepts to these chatbots across three sessions: analyzing the acceleration of a teddy bear at its highest point (similar to references [29,30]), understanding the speed of a ball at half height, and discussing a roller coaster loop. ChatGPT-4 emerged as particularly notable for its accurate application of physics concepts. Nikolic et al. [52] investigated the performance of ChatGPT, Copilot, Gemini, SciSpace, and Wolfram across assessment tasks in 10 engineering subjects, including quizzes, numerical problems, oral, visual, programming, and writing tasks. The results showed that chatbots are generally unlikely to pass these assessments, with visual, project-based written, and research-based written assessments being secure. In addition, the authors note that ChatGPT-4 is particularly reliable for most engineering applications, suggesting the benefits of the paid version.

In chemistry, Watts et al. compared writing-to-learn assignments produced by ChatGPT-3.5, ChatGPT-4, and Bard [53]. They found that while the responses varied, the chatbots seldom discussed electron movement, a critical component of mechanistic reasoning. As a result, the chatbots did not engage in reasoning to the same extent as students. Regarding structural notations, Hallal et al.

compared ChatGPT and Bard's understanding of condensed structures, InChi, and SMILES, and their ability to answer organic chemistry-related questions [54]. They assessed the chatbots' abilities to convert IUPAC names, InChi, and SMILES notations into condensed forms and vice versa, identify functional groups, generate molecular formulas, and predict resonance patterns. Leite [55] also compared the performance of ChatGPT, Gemini, and Copilot in defining five basic chemistry concepts (atom, electron, mole, molecule, and chemical substance) against IUPAC definitions. The author found that Copilot was the only chatbot to cite sources for its generated text, while Gemini provided references. Additionally, the author used GPTZero, Plagium, Smodim, and AI Content Detector to determine if the texts were AI-generated. The results showed that GPTZero, Plagium, and Smodim detected AI-generated text 33.3% of the time, while AI Content Detector detected it only 6.7% of the time. Overall, the author noted that Gemini delivered the most satisfactory responses, followed by Copilot and ChatGPT. Nascimento Jr et al. [56] compare ChatGPT 3.5, ChatGPT 4.0, Google Bard, Bing Chat, Adobe Firefly, Leonardo.AI, and DALL-E, focusing on both textual and imagery content. These AIs are classified as Free or Paid (ChatGPT 4.0 and DALL-E). For textual content, the chatbots perform well in chemical bonding, aligning with scientific consensus (Ct3). In imagery, only the paid ChatGPT 4.0 effectively identifies chemical content, Lewis structures, and arrow orientations, generating mostly accurate responses with minor errors.

In biophysical phenomena, liquid–liquid phase separation (LLPS) was presented to ChatGPT-4 and Gemini to evaluate their explanations and understanding [57]. The authors analyzed accuracy, response time, response length, and cosine similarity index (CSI) of the responses. Gemini consistently provided more accurate answers than ChatGPT, though neither model answered all questions correctly. The CSI was 0.62, indicating moderate similarity between the models.

In medicine, Rossettini et al. [58] investigated the performance of ChatGPT-4, Microsoft Copilot, and Google Gemini in passing the Italian entrance exam for healthcare science degrees (CINECA test). The results show that ChatGPT-4 and Microsoft Copilot outperformed Google Gemini. The authors note that differences in neural network architecture impact accuracy, with ChatGPT-4 and Microsoft Copilot using GPT (Generative Pre-trained Transformer) architecture, while Google Gemini employs LaMDA (Language Model for Dialogue Application) and later PaLM 2 (Pathways Language Model) combined with web search. Meyer et al. [59] compared responses from ChatGPT, Gemini, and Le Chat in interpreting complete blood counts in an online health forum. The results showed inaccuracies in the chatbots' interpretations, particularly with complex patient questions. This research highlights the need for patient caution when using chatbots for self-diagnosis. Saeedi and Aghajanzadeh [60] used ChatGPT and Perplexity to assess dysphonia's perceptual level by analyzing voice self-assessments and acoustic data. Chatbots were asked to classify the severity of voice disorders (vocally healthy, mild, moderate, or severe dysphonia). The authors found that while chatbots occasionally made correct assessments, their reliability was inconsistent for clinical use. Kaba et al. [61] tested ChatGPT-3.5, ChatGPT-4, Gemini, and Perplexity with MRI safety-related questions, finding that ChatGPT-4 outperformed the others with an accuracy of 93.3%, followed by Gemini, Perplexity, and ChatGPT-3.5. The authors suggest that these chatbots could potentially assist healthcare professionals in the future.

Throughout our review of comparison studies on chatbots used as tutees, tutors, and tools across various fields, many studies suggest that the paid version of ChatGPT-4 outperforms other chatbots. However, we believe that more analysis is still needed to compare chatbots, as paid versions should provide greater accuracy in all chatbots, not just ChatGPT. It is noted that the paid versions of ChatGPT, Gemini, and Perplexity, which are popular chatbots, are priced at 20 USD per month. This

advantage raises concerns about educational inequality, particularly for those with financial constraints. Different chatbots offer various benefits and perspectives. For instance, ChatGPT relies on its dataset to create responses based on trends discovered during training, while Gemini retrieves and analyzes data from Google Search in real-time for more realistic queries. However, both of them have drawbacks related to transparency, unlike Perplexity AI, which allows for the citation of its information sources [62]. In terms of research tools and education, we believe Perplexity is more suitable for providing scientific citations. However, its overall performance requires further exploration, as this is just the beginning of the chatbot era, where training biases and user interactions impact accuracy. Furthermore, we believe that a comparison of paid chatbots is needed, as much of the literature compares free versions of chatbots with paid versions of ChatGPT, which could introduce bias in assessing the accuracy of chatbot performance.



## 6. Ethical Considerations and Outlook

This literature review offers insights into the emerging roles of ChatGPT in physics and chemistry education. ChatGPT demonstrates impressive capabilities as a tutee, effectively providing scientific descriptions, solving conceptual problems, and passing standard tests. However, obtaining effective answers often requires an understanding of ChatGPT's functioning and specific approaches.

Utilizing strategies like Chain-of-Thought and Socratic-style dialogue can significantly enhance its efficacy and utility [63,64]. Advanced versions of ChatGPT and other chatbots exhibit improved abilities in creating simulations and reading graphs. Nonetheless, chatbots are inherently limited in their in-depth knowledge and advanced analysis, occasionally resulting in misleading information. It follows that teaching should not rely solely on current chatbot technology. The literature reveals varying levels of effectiveness among current chatbots as tutors, suggesting that they can be helpful as teaching assistants. ChatGPT-4 is particularly notable for its accurate application of scientific concepts, demonstrating a deep understanding of the learning process by exhibiting excellent facilitation skills, delivering content knowledge, and encouraging student engagement. The literature also highlights substantial advancements in chatbots' roles as facilitators for academic writing and assessment, showcasing their versatility as tools for science education. Interestingly, chemistry educators and researchers have explored the potential of ChatGPT in laboratory and research design. Such chatbot assistance in idea development will be extremely useful across various other science disciplines.

Given the massive impact of generative AI, the focus has shifted to supporting educators in integrating the technology effectively and raising student awareness of its ethical use. Ethical considerations regarding the use of AI, as highlighted by several authors [2,65-68], are summarized as follows. Over-reliance on technology, bypassing critical thinking, writing, or problem-solving processes, risks diminishing human creativity, ingenuity, and intellectual development in both teaching and learning. Misuse of AI can lead to academic dishonesty, plagiarism, and reduced engagement in the learning process. Transparency and accountability in AI decision-making, as well as copyright issues, remain areas of concern. Educators and students must verify AI-generated outputs through trusted academic sources to ensure accuracy, as AI systems can produce incorrect or misleading information. Generative AI, trained on vast datasets, may inadvertently amplify biases related to gender, race, or

socioeconomic status, and without fully understanding the context, AI can produce inappropriate outputs for specific age groups or academic levels. Additionally, concerns about security and privacy arise, as personal data and interactions with AI systems may be stored and analyzed without explicit consent. The introduction of AI in education could also exacerbate the digital divide, disadvantaging students without access to high-quality technology and resources. Furthermore, integrating AI in education may reduce human-to-human interaction, which is crucial for developing social and emotional skills. Overcoming these challenges is a moral obligation for users of the technology. Jobin et al. identified eleven key ethical principles in AI implementation [69], while Petricini emphasized the importance of truthfulness, temperance, prudence, and courage—four specific Aristotelian virtues—to guide ethical AI practices [70]. While the emergence of chatbots and other AI tools should indeed be cherished as opportunities for progress, their implementation must remain anchored in human intelligence and moral virtue.



**Acknowledgments**

During the preparation of this work the authors used ChatGPT (powered by OpenAI's language model, GPT-3.5) in order to improve the readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

**Reference**

[1] M. Ivanova, G. Grosseck, C. Holotescu, Unveiling insights: A bibliometric analysis of artificial intelligence in teaching. *Informatics* **2024**, 11, 10. DOI 10.3390/informatics11010010

[2] M. R. Zheltukhina, O. V. Sergeeva, A. R. Masalimova, R. L. Budkevich, N. N. Kosarenko, G. V. Nesterov, A bibliometric analysis of publications on ChatGPT in education: Research patterns and topics. *Online J. Commun. Media Technol.* **2024**, 14(1), e202405. DOI 10.30935/ojcmt/14103

[3] E. Dimitriadou, A. Lanitis, A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learn. Environ.* **2024** DOI 10.1186/s40561-023-00231-3

[4] F. Jia, D. Sun, C.-K. Looi, Artificial intelligence in science education (2013–2023): Research trends in ten years. *J. Sci. Educ. Technol.* **2024**, 33, 94-117. DOI 10.1007/s10956-023-10077-6

[5] P. Lo Nostro, Artificial intelligence vs. Natural stupidity. *Substantia* **2023**, 7, 5-6, DOI 10.36253/Substantia-2250

[6] L. S. Balhorn, J. M. Weber, S. Buijsman, J. R. Hildebrandt, M. Ziefle, A. M. Schweidtmann, Empirical assessment of ChatGPT's answering capabilities in natural science and engineering. *Sci. Rep.* **2024**, 14, 4998. DOI 10.1038/s41598-024-54936-7

[7] C. K. Lo, What is the impact of ChatGPT on education? A rapid review of the literature. *Educ. Sci.* **2023,** 13, 410. DOI 10.3390/educsci13040410

[8] J. Weizenbaum, Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, 9, 36-45.

[9] E. Adamopoulou, L. Moussiades, Chatbots: History, technology, and applications. *Machine Learn. Appl.* **2020**, 2, 100006. DOI 10.1016/j.mlwa.2020.100006

[10] S. Schobel, A. Schmitt, D. Benner, M. Saqr, A. Janson, J. M. Leimeister, Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. *Inform. Sys. Front.* **2024**, online first. DOI 10.1007/s10796-023-10375-9

[11] J. Wang, ChatGPT: A test drive. *Am. J. Phys.* **2023**, 91, 255-256. DOI 10.1119/5.0145897

[12] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* **2023**, 19, 010132. DOI 10.1103/PhysRevPhysEducRes.19.020163

[13] D. Hestenes, I. Halloun, Interpreting the force concept inventory: A response to Huffman and Heller. *Phys. Teach.* **1995**, 33. 502-502. DOI 10.1119/1.2344278

[14] W. Yeadon, O. Inyang, A. Mizouri, A. Peach, C. Testrow, The death of the short-form physics essay in the coming AI revolution. *Phys. Educ.* **2023**, 58, 035027. DOI 10.1088/1361-6552/acc5cf

[15] V. López, M. Rezende, Challenging ChatGPT with different types of physics education questions. *Phys. Teach.* **2024**, 62. 290-294. DOI 10.1119/5.0160160

[16] D. Tong, Y. Tao, K. Zhang, X. Dong, Y. Hu, S. Pan, Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pac. Educ. Rev.* **2024**, DOI 10.1007/s12564-023-09913-6

[17] D. Maloney, T. O'Kuma, C. Hieggelke, A. Heuvelen, Surveying students' conceptual knowledge of electricity & magnetism. *Am. J. Phys.* **2001**, 69, S12-S23. DOI 10.1119/1.1371296.

[18] H. Xing, W. Gong, P. Wang, Y. Zhai, Y. Zhao, The measuring instrument of primitive physics problem for junior high school students: Compilation and research. *J Baltic Sci Educ.* **2022**, 21, 305-324. DOI 10.33225/jbse/22.21.305

[19] C. West, AI and the FCI: Can ChatGPT project an understanding of introductory physics? *Arxiv* **2023**, DOI 10.48550/arXiv.2303.01067

[20] T. Kumar, M. A. Kats; ChatGPT-4 with Code Interpreter can be used to solve introductory college-level vector calculus and electromagnetism problems. *Am. J. Phys.* **2023**, 91, 955-956. DOI 10.1119/5.0182627

[21] W. Yeadon, D. P. Halliday, Exploring Durham University physics exams with large language models. *Arxiv* **2023**, DOI 10.48550/arXiv.2306.15609

[22] G. Polverini, B. Gregorcic, Performance of ChatGPT on the test of understanding graphs in kinematics. *Phys. Rev. Phys. Educ. Res.* **2024**, 20, 010109. DOI 10.1103/PhysRevPhysEducRes.20.010109

[23] G. Polverini, B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education. *Eur. J. Phys.* **2024**, 45, 025701. DOI 10.1088/1361-6404/ad1420

[24] Y. Liang, D. Zou, H. Xie, L. Wang, Exploring the potential of using ChatGPT in physics education. *Smart Learn. Environ.* **2024**, 10, 52. DOI 10.1186/s40561-023-00273-7

[25] P. Bitzenbauer, ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemp. Educ. Technol.* **2023**, 15, ep430. DOI 10.30935/cedtech/13176

[26] S. Alneyadi, Y. Wardat, ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemp. Educ. Technol.* **2023**, 15. 448-1309. DOI 10.30935/cedtech/13417

[27] M. Dahlkemper, S. Lahme, P. Klein, How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic

quality of ChatGPT. *Phys. Rev. Phys. Educ. Res.* **2023**, 19, 010142. DOI 10.1103/PhysRevPhysEducRes.19.010142.

[28] L. Ding, T. Li, S. Jiang, A. Gapud, Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *Int. J. Educ. Technol. Higher Educ.* **2023**, 20, 63. DOI 10.1186/s41239-023-00434-1

[29] B. Gregorcic, A.-M. Pendrill, ChatGPT and the frustrated Socrates. *Phys. Educ.* **2023**, 58. 035021. DOI 10.1088/1361-6552/acc299

[30] B. Gregorcic, G. Polverini, A. Sarlah, ChatGPT as a tool for honing teachers' Socratic dialogue skills. *Phys. Educ.* **2024**, 59. 045005. DOI 10.1088/1361-6552/ad3d21

[31] G. Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Phys. Rev. Phys. Educ. Res.* **2023**, 19, 020163. DOI 10.1103/PhysRevPhysEducRes.19.010132

[32] T. M. Clark, E. Anderson, N. M. Dickson-Karn, C. Soltanirad, N. Tafini, Comparing the performance of college chemistry students with ChatGPT for calculations involving acids and bases. *J. Chem. Educ.* **2023**, 100, 3934-3944. DOI 10.1021/acs.jchemed.3c00500

[33] A. J. Leon, D. Vidhani, ChatGPT needs a chemistry tutor too. *J. Chem. Educ.* **2023**, 100, 3859-3865. DOI 10.1021/acs.jchemed.3c00288

[34] S. Fergus, M. Botha, M. Ostovar, Evaluating academic answers generated using ChatGPT. *J. Chem. Educ.* **2023**, 100,1672-1675. DOI 10.1021/acs.jchemed.3c00087

[35] T. M. Clark, Investigating the use of an artificial intelligence chatbot with general chemistry exam question. *J. Chem. Educ.* **2023**, 100, 1905-1916. DOI 10.1021/acs.jchemed.3c00027

[36] C. M. C. Nascimento, A. S. Pimentel, Do large language models understand chemistry? A conversation with ChatGPT. *J. Chem. Inform. Model.* **2023**, 63, 1649-1655. DOI 10.1021/acs.jcim.3c00285

[37] W. Daher, H. Diab, A. Rayan, Artificial intelligence generative tools and conceptual knowledge in problem solving in chemistry. *Information* **2023**, 14, 409. DOI 10.3390/info14070409

[38] E. A. Alasadi, C. R. Baiz, Generative AI in education and research: Opportunities, concerns, and solutions. *J. Chem. Educ.* **2023**, 100, 2965-2971. DOI 10.1021/acs.jchemed.3c00323

[39] Y. Guo, D. Lee, Leveraging ChatGPT for enhancing critical thinking skills. *J. Chem. Educ.* **2023**, 100, 4876-4883. DOI 10.1021/acs.jchemed.3c00505

[40] B. Exintaris, N. Karunaratne, E. Yuriev, Metacognition and critical thinking: Using ChatGPT-generated responses as prompts for critique in a problem-solving workshop (SMARTCHEMPer). *J. Chem. Educ.* **2023**, 100, 2972-2980. DOI 10.1021/acs.jchemed.3c00481

[41] J. K. West, J. L. Franz, S. M. Hein, H. R. Leverentz-Culp, J. F. Mauser, E. F. Ruff, J. M. Zemke, An analysis of ai-generated laboratory reports across the chemistry curriculum and student perceptions of ChatGPT. *J. Chem. Educ.* **2023**, 100, 4351-4359. DOI 10.1021/acs.jchemed.3c00581

[42] A. J. Rojas, An investigation into ChatGPT's application for a scientific writing assignment. *J. Chem. Educ.* **2024**, DOI 10.1021/acs.jchemed.4c00034

[43] M. J. Clark, M. Reynders, T. A. Holme, Students' experience of a ChatGPT enabled final examina non majors chemistry course. *J. Chem. Educ.* **2024**, DOI 10.1021/acs.jchemed.4c00161

[44] H. Desaire, A. E. Chua, M.-G. Kim, D. Hua, Accurately detecting AI text when ChatGPT is told to write like a chemist. *Cell Rep. Phys. Sci.* **2023**, 4, 101672. DOI 10.1016/j.xcrp.2023.10167

[45] J. L. Araujo, I. Saude, Can ChatGPT enhance chemistry laboratory teaching? using prompt engineering to enable AI in generating laboratory activities. *J. Chem. Educ.* **2024**, DOI 10.1021/acs.jchemed.3c00745

[46] J. Scoggin, K. C. Smith, Enabling general chemistry students to take part in experimental design activities. *Chem. Educ. Res. Pract.* **2023**, 24, 1229. DOI 10.1039/d3rp00088e

[47] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, O. M. Yaghi, ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **2023**, 145, 18048-18062. DOI 10.1021/jacs.3c05819

[48] B. Mahjour, J. Hoffstadt, T. Cernak, Designing chemical reaction arrays using Phactor and ChatGPT. *Org. Process Res. Dev.* **2023**, 27, 1510-1516. DOI 10.1021/acs.oprd.3c00186

[49] Z. Y. Kong, V. S. K. Adi, J. Sunarso, J. G. Segovia-Hernandez, Complementary role of large language models in educating undergraduate design of distillation column: Methodology development. *Digit. Chem. Eng.* **2023**, 9, 100126. DOI 10.1016/j.dche.2023.100126

[50] T. Hasrod, Y. B. Nuapia, H. Tutu, ChatGPT helped me build a chemistry App, and here's how you can make one also. *J. Chem. Educ.* **2024**, 101, 653-660. DOI 10.1021/acs.jchemed.3c01170

[51] R. Dos Santos, Enhancing physics learning with ChatGPT, Bing Chat, and Bard as agents-to-think-with: A comparative case study. *arXiv* **2023**, DOI 10.48550/arXiv.2306.00724

[52] S. Nikolic, C. Sandison, R. Haque, S. Daniel, S. Grundy, M. Belkina, S. Lyden, G. M. Hassan, P. Neal. ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an

updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. Australasian J. Eng. Educ. **2024**, online first. DOI 10.1080/22054952.2024.2372154.

[53] F. M. Watts, A. J. Dood, G. V. Shultz, J.-M. G. Rodriguez, Comparing student and generative artificial intelligence chatbot responses to organic chemistry writing-to-learn assignments. *J. Chem. Educ.* **2023**, 100, 3806-3817. DOI 10.1021/acs.jchemed.3c00664

[54] K. Hallal, R. Hamdan, S. Tlais, Exploring the potential of AI-Chatbots in organic chemistry: An assessment of ChatGPT and Bard. *Comput. Educ.: Artif. Intel.* **2023**, 5, 100170. DOI 10.1016/j.caeai.2023.100170

[55] B. Leite, Generative Artificial Intelligence in chemistry teaching: ChatGPT, Gemini, and Copilot's content responses. *J. Appl. Learn. Teach.* **2024**, 7(2), 1-15. DOI 10.37074/jalt.2024.7.2.13

[56] W. J. D. Nascimento Jr, C. Morais, G. Girotto Jr, Enhancing AI Responses in Chemistry: Integrating Text Generation, Image Creation, and Image Interpretation through Different Levels of Prompts. *J. Chem. Educ.* **2024,** 101, 3767-3779. DOI 10.1021/acs.jchemed.4c00230

[57] N. Rana, N. Katoch, AI for biophysical phenomena: A comparative study of ChatGPT and Gemini in explaining liquid–liquid phase separation. *Appl. Sci.* **2024**, 14, 5065. DOI 10.3390/app14125065

[58] G. Rossettini, L. Rodeghiero, F. Corradi, C. Cook, P. Pillastrini, A. Turolla, G. Castellini, S. Chiappinotto, S. Gianola, A. Palese, Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. *BMC Med. Educ.* **2024**, 24, 694. DOI 10.1186/s12909-024-05630-9

[59] A. Meyer, A. Soleman, J. Riese, T. Streichert, Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. *Clin. Chem. Lab. Med.* **2024**, online first. DOI 10.1515/cclm-2024-0246

[60] S. Saeedi, M. Aghajanzadeh, Investigating the role of artificial intelligence in predicting perceived dysphonia level. *Eur. Arch. Otorhinolaryngol.* **2024**, online first. DOI 10.1007/s00405-024-08868-7

[61] E. Kaba, H. M. Bülbül, G. Burakgazi, A. T. Varlık, N. Hürsoy, M. Solak, S. Tabakoğlu, F. B. Çeliker. Large language models on magnetic resonance imaging safety-related questions: accuracy of ChatGPT-3.5, ChatGPT-4, Gemini, and Perplexity. *Curr. Res. MRI* **2024**, 3, 16-19. DOI 10.5152/CurrResMRI.2024.24095

[62] M. Shukla, I. Goyal, B. Gupta, J. Sharma, A comparative study of ChatGPT, Gemini, and Perplexity. *Int. J. Innov. Res. Comput. Sci. Technol.* **2024**, 12(4), 10-15. DOI 10.55524/ijircst.2024.12.4.2

[63] C. Sirisathitkul, Slow writing with ChatGPT: Turning the hype into a right way forward. *Postdigit. Sci. Educ.* **2024**, 6, 431-438. DOI 10.1007/s42438-023-00441-5

[64] A. Brailas, Postdigital duoethnography: An inquiry into human-artificial intelligence synergies. *Postdigit. Sci. Educ.* **2024**, 6, 486-515. DOI 10.1007/s42438-024-00455-7

[65] F. F.-H. Nah, R. Zheng, J. Cai, K. Siau, L. Chen, Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration, *J. Inform. Technol. Case Appl. Res.* **2023**, 25, 277-304. DOI 10.1080/15228053.2023.2233814

[66] T. Adiguzel, M. H. Kaya, F. K. Cansu, Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemp. Educ. Technol.* **2023**, 15, ep429. DOI 10.30935/cedtech/13152

[67] H. Crompton, D. Burke, The educational affordances and challenges of ChatGPT: State of the field. *TechTrends* **2024**, 68, 380-392. DOI 10.1007/s11528-024-00939-0

[68] A. O. Ifelebuegu, P. Kulume, P. Cherukut, Chatbots and AI in Education (AIEd) tools: The good, the bad, and the ugly. *J. Appl. Learn. Teach.* **2023**, 6, 332-345. DOI 10.37074/jalt.2023.6.2.29

[69] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines. *Nature Machine Intell.* **2019**, 1, 389-399. DOI 10.1038/s42256-019-0088-2

[70] T. Petricini, What would Aristotle do? Navigating generative artificial intelligence in higher education. *Postdigit. Sci. Educ.* **2024**, 6, 439-445. DOI 10.1007/s42438-024-00463-7