Exploring the Potential of ChatGPT as a Substitute Teacher: A Case Study

Shatha Al Ghazali[1,*], Nazar Zaki[1], Luqman Ali[1], and Saad Harous[2]

[1]Department of Computer Science and Software Engineering, College of Information Technology, UAEU, UAE
[2]Department of Computer Science and Software Engineering, University of Sharjah, Sharjah, UAE
Email: 202090002@uaeu.ac.ae (S.A.G.); nzaki@uaeu.ac.ae (N.Z.); 201990024@uaeu.ac.ae (L.A.); harous@sharjah.ac.ae (S.H.)
*Corresponding author
Manuscript received June 9, 2023; revised July 3, 2023; accepted July 14, 2023; published February 18, 2024

***Abstract*—This study explores the potential of utilizing ChatGPT, an AI-powered chatbot, as a substitute teacher in the context of teaching chemistry to eleventh-grade students following the American curriculum in the United Arab Emirates. The study aims to assess the effectiveness of ChatGPT in engaging students and facilitating learning across different cognitive domains. The research employs Bloom's taxonomy to categorize learning objectives and evaluate student performance in the areas of knowing, applying, and reasoning. The study involved a sample of 39 students divided into an experimental group and a control group. Findings reveal that while ChatGPT shows promise in certain aspects, such as knowledge recall and reasoning skills, it faces challenges related to student engagement and completion rates. Analysis of students performance in the post class assessment shows that the average percentage of students who's answers were evaluated as good and outstanding according to the evaluation rubric and across all three cognitive domains, is almost the double in the control group in comparison to the experiment group, 15% and 7.5% respectively. The study also highlights the importance of balancing AI-powered tools with traditional teaching methods.**

***Keywords*—AI-powered chatbot, ChatGPT, cognitive domains, educational technology, substitute teacher**

I. Introduction

In the context of educational systems, class substitution refers to the practice of employing a substitute teacher to cover a class when the regular teacher is unavailable. Common reasons for this include illness, professional development attendance, or personal time off. Substitute teachers are responsible for maintaining classroom management and instruction, often needing to adapt to the regular teacher's lesson plans or create their own [1–3]. Effective substitutes must possess strong skills in classroom management, communication, and improvisation, as they frequently encounter unfamiliar situations and need to adapt quickly to new groups of students [1–3].

During class substitution, students may face challenges that could result in disengagement and disruptive behaviors. These challenges include difficulty comprehending concepts due to teacher changes [1], lack of personalized learning leading to disengagement [4], and the feeling of boredom, which can potentially lead to harsh tendencies and disruptive behaviors towards other students. [5].

The implementation of ChatGPT in schools has sparked debate among educators and experts. While some highlight its potential to enhance teaching and learning, others express concerns about its role in facilitating cheating. Innovative educators have found ways to incorporate ChatGPT into their teaching methods, such as utilizing ChatGPT to develop customized reading materials that cater to students' requirements. Moreover, it can automate the process of generating questions and tests. By comprehending natural language and generating text, ChatGPT can produce tailored questions and tests that align with students' comprehension levels. This capability simplifies the task of teachers in creating pertinent assessments and saves time [6]. Other educators recommend employing ChatGPT as a starting point for students to generate answers or create articles tailored to their reading level [7]. ChatGPT can also generate teaching material such as syllabi, lesson plans, classroom rules, learning objectives, activity instructions, discussion prompts, substitute teacher plans, presentation scripts, and classroom management tips [8]. Additionally, by engaging students in critiquing and refining ChatGPT responses, teachers can help improve students' writing and critical thinking skills [9].

Numerous studies have been conducted to explore the use of chatbots in teaching and learning contexts. The study [10] presents a case study exploring the use of a chatbot as a conversational agent to enhance the learning experience of pre-university students. The chatbot was designed to offer personalized support to students and help them stay on track with their studies. The study discovered that the chatbot effectively improved student engagement and performance while reducing teachers' workload. The paper proposes that chatbots can be an effective tool for enhancing learning experiences in pre-university education.

Lee and Yeo [11] elaborate on the development of a chatbot aimed at supporting responsive teaching practices in mathematics education. The study outlines the chatbot's development process, which utilized a GPT-2 model and was designed to provide personalized feedback to mathematics teachers on their pedagogical practices. The paper also discusses the chatbot's evaluation, which revealed its effectiveness in improving teacher practices and reducing workload. The authors suggest that AI-based chatbots, like the one developed in the study, have the potential to support teachers' professional development in mathematics education.

Sáez-Manzanares *et al.* [12] examine university students' perceived satisfaction with chatbots as a tool for self-regulated learning. The study involved a survey of university students who had used a chatbot to support their learning and found that the majority were satisfied with the chatbot, considering it a useful tool for self-regulated learning. The paper posits that chatbots have the potential to support self-regulated learning by providing personalized feedback, guidance, and support to students.

Liu *et al.* [13] analyze the interaction between children and an AI chatbot designed to promote reading. The study involved a group of children interacting with the chatbot over

a period and found that the chatbot positively impacted their interest in reading. The paper discusses the chatbot's design and development, as well as the evaluation of its influence on children's reading interests. The authors suggest that chatbots have the potential to enhance children's reading interests and support literacy development.

Guo *et al.* [14] investigate the use of chatbots to support English as a foreign language (EFL) students' argumentative writing. The study entailed developing a chatbot designed to provide scaffolded support to EFL students as they wrote argumentative essays and evaluating its effectiveness. The study concluded that the chatbot was effective in offering personalized support and feedback to students, leading to improved writing performance. The authors propose that chatbots have the potential to support EFL students' writing development and enhance their language learning experiences.

Although the mentioned studies have provided insights into the benefits of using chatbots in educational settings, the common factor across these studies is that learners were receiving a direct guidance from the subject teacher or instructor. Therefore, there is a need for more extensive exploration and investigation of such tools, especially given the emergence of powerful AI-based solutions like ChatGPT. This would help better understand their full potential and limitations in various educational contexts.

The main contribution of this article is to conduct an empirical study examining the potential of AI-based resources, such as ChatGPT, to enhance the quality of education during substitute teaching sessions where the subject teacher is unavailable.

II. Method

The aim of this study is to explore the potential of utilizing ChatGPT, an AI-powered chatbot, as a substitute teacher during teacher absences. In particular, the experiment focuses on teaching a chemistry topic to students attending a school that adheres to the American curriculum in the United Arab Emirates. The following sections will elaborate on the participants, the experimental design, and the specific topic chosen for this study.

A. Participants

The study involved 39 eleventh-grade students, who were divided into two groups—an experimental group consisting of 20 students and a control group comprising 19 students. Both groups included students with varying abilities and were composed solely of male students from diverse backgrounds.

B. The Topic

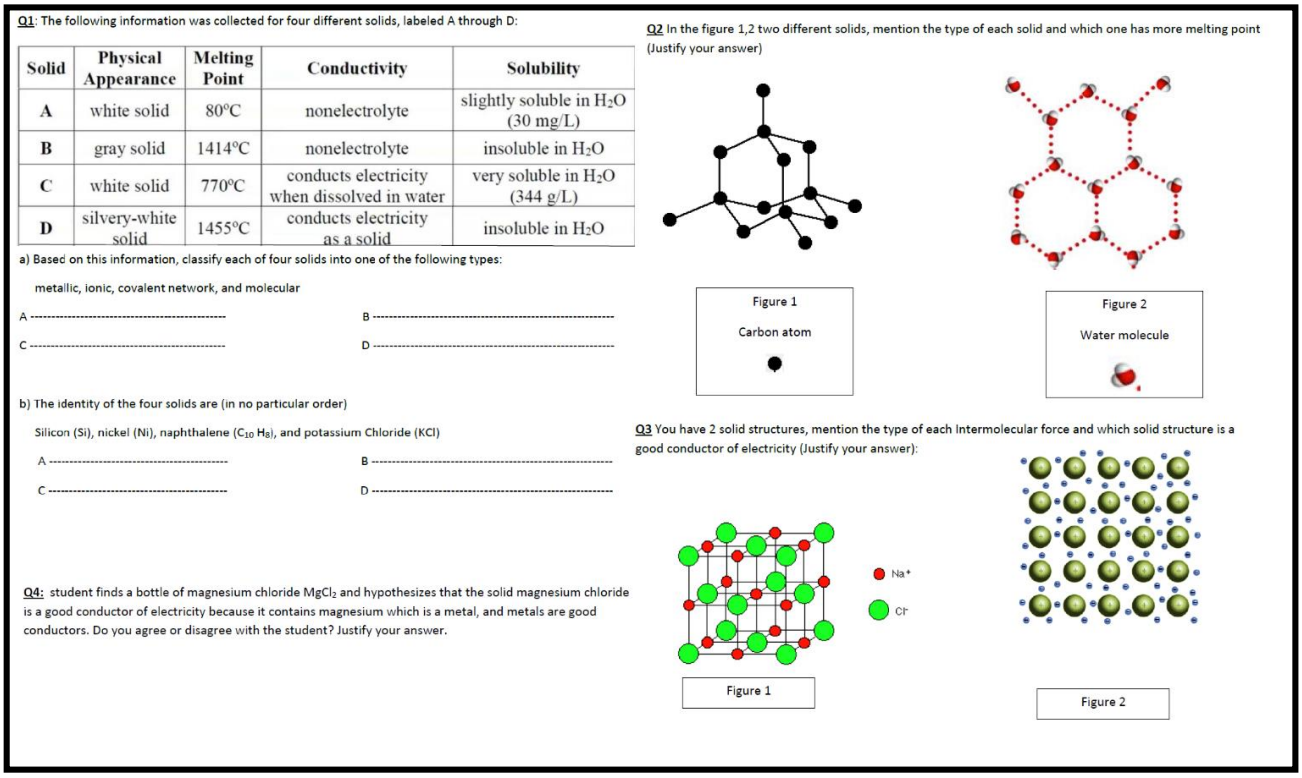
The lesson, a part of the grade 11 Chemistry curriculum, centers on the properties of solids. It has two primary objectives: a) comparing various types of solids, and b) identifying different crystalline solids. This class period represents the fourth session devoted to this topic, following three earlier lessons that covered related subject matter. Earlier in the week, students were introduced to the topic. In the first lesson, they learned to differentiate between solids, liquids, and gases based on molecular arrangements. The following lesson taught them to distinguish between solids, liquids, and gases according to 1) viscosity, 2) vapor pressure, and 3) changes in states. In the third lesson, students learned to differentiate between intramolecular and intermolecular forces, as well as define distinct types of intermolecular forces.

C. The Experiment

The school's IT team created ChatGPT login credentials for the students. The class was divided into two groups: Group 1, a control group led by the subject teacher, and Group 2, an experimental group guided by ChatGPT. Due to safety precautions, a substitute teacher from a different subject area supervised the students in the experimental group. Both classes began simultaneously and lasted for approximately 35 minutes, within a 50-minute class period. In the control class, the teacher utilized planned learning resources, including both paper-based and digital activities. In the experimental class, the supervising teacher gave a brief overview of how the class period would be structured. She allocated 35 minutes for students to explore information related to the topic and 10 minutes to answer a paper. Additionally, she shared the lesson objectives outlined in the subject teacher's lesson plan and provided the assessment rubric. The students were informed by the teacher that they could use ChatGPT to seek clarification and ask questions pertaining to the lesson objectives. However, the teacher didn't provide any specific guidelines on how to generate specific prompts using ChatGPT. According to the supervising teacher, some students were trying to find some diagrams and data tables that show the comparisons. Other students got some satisfying answers for their questions while some other students received lengthy reading texts. It is worth noting that the teacher didn't interfere in any subject-related matters since Chemistry is not her major. In both groups, it was made clear to the students that they would be evaluated at the end of the class. During the final 10 minutes of both classes, students were given a paper-based assessment without access to any digital resources to evaluate their comprehension of the topic.

D. The Assessment

The paper-based assessment consists of four main questions (Q1–Q4), with Q1 and Q2 further divided into sub-questions that encompass the cognitive domains of knowing, applying, and reasoning. In Question 1, information is provided about four different solids and students are asked to classify their types and identity. Question 2 provides figures of two different solids, and requires students to mention with justifications, the type of each solid and which one has more melting point. In Question 3, the students are provided with the structure of two solids and are asked to mention with justification, the type of each Intermolecular force and which solid structure is a good conductor of electricity. Finally, Question 4 provides students with a word problem about a hypotheses that the solid magnesium chloride is a good conductor of electricity because it contains magnesium which is a metal, and metals are good conductors. The students are asked to provide a justification if they agree or disagree with the hypotheses. Fig. 1 shows these questions.



Q1: The following information was collected for four different solids, labeled A through D:

Solid	Physical Appearance	Melting Point	Conductivity	Solubility
A	white solid	80°C	nonelectrolyte	slightly soluble in H_2O (30 mg/L)
B	gray solid	1414°C	nonelectrolyte	insoluble in H_2O
C	white solid	770°C	conducts electricity when dissolved in water	very soluble in H_2O (344 g/L)
D	silvery-white solid	1455°C	conducts electricity as a solid	insoluble in H_2O

a) Based on this information, classify each of four solids into one of the following types:

metallic, ionic, covalent network, and molecular

A ---------- B ----------

C ---------- D ----------

b) The identity of the four solids are (in no particular order)

Silicon (Si), nickel (Ni), naphthalene ($C_{10}H_8$), and potassium Chloride (KCl)

A ---------- B ----------

C ---------- D ----------

Q4: student finds a bottle of magnesium chloride $MgCl_2$ and hypothesizes that the solid magnesium chloride is a good conductor of electricity because it contains magnesium which is a metal, and metals are good conductors. Do you agree or disagree with the student? Justify your answer.

Q2 In the figure 1,2 two different solids, mention the type of each solid and which one has more melting point (Justify your answer)

Q3 You have 2 solid structures, mention the type of each Intermolecular force and which solid structure is a good conductor of electricity (Justify your answer):

Fig. 1. A sample questions from the paper-based assessment in the cognitive domains of knowing, applying, and reasoning.

Table 1 represents the paper-based assessment questions and the corresponding cognitive domains (Knowing, Applying, Reasoning)

Table 1. Paper-based assessment questions and cognitive domains

	Knowing	Applying	Reasoning
Q1-a	√		
Q1-b:		√	
Q2-a		√	
Q2-b			√
Q3-		√	
Q4-			√

To assess the quality of students' responses, the teacher utilized a rubric, which helps stay focused on the goals as educators. Rubrics assist teachers in defining their learning objectives, developing instructional strategies that align with those objectives, effectively communicating the goals to students, providing feedback based on their progress towards the goals, and evaluating final outcomes in relation to the extent to which the goals were achieved. Teachers employ rubrics before, during, and after delivering instruction, and the advantages are abundant. Instructional rubrics play a crucial role in enhancing the clarity of the teaching methods [15]. Rubrics are adaptable and can be employed to assess a wide range of student work, such as essays, reports, speeches, creative projects, and science experiments [16]. By utilizing rubrics, teachers can provide comprehensive and targeted feedback to students, enhancing the effectiveness of the assessment and facilitating timely feedback delivery [17].

For effective use of rubrics, it is crucial to ensure that they are clear, specific, and relevant to the given assignment. Introducing the rubric to students at the outset of the assignment enables them to be aware of the evaluation criteria [18]. Additionally, rubrics can be modified or adjusted to align with specific learning objectives and standards [19]. Table 2 provides an overview of the rubric employed to evaluate students' responses to the given assessment.

Table 2. Rubric for evaluating students' responses in the assessment

Domain	Outstanding	Good	Acceptable	No answer
Knowledge	The student is able to connect between the properties of solids and different given compounds.	The student is able to compare different solid properties to mention the type of solid.	The student is able to use solid properties to determine the type of solid.	The student left the question unanswered.
Application	The student is able to compare the strength of intermolecular forces for each structural formula.	The student is able to mention the type of intermolecular forces in each structural formula.	The student is able to use a structural formula of different solids to determine the type of solid.	The student left the question unanswered.
Reasoning	The student is able to explain the structural formula, mention the type of intermolecular forces in each, and relate them with solid properties (melting point, electric conductivity).	The student is able to use the type of intermolecular forces in each formula to compare their melting point.	The student is able to use the type of solid to mention the melting point of different structural formulas.	The student left the question unanswered.

E. Bloom's Taxonomy to Analyze, Compare, and Assess the Effectiveness of Different Teaching Methods

Bloom's taxonomy is used in this study as it offers a systematic approach to analyze, compare, and assess the effectiveness of different teaching methods across various cognitive domains, thereby contributing to a better

understanding of the learning outcomes and potential improvements in teaching practices. It is used in this study for several reasons:

1) *Framework for cognitive complexity*: Bloom's taxonomy provides a structured framework to classify learning objectives into different levels of cognitive complexity. This allows for a systematic approach to analyzing students' understanding and performance in various cognitive domains.
2) *Comparison of teaching methods*: By employing Bloom's taxonomy, the study can compare the effectiveness of ChatGPT and traditional teaching methods across the different cognitive domains. This helps to identify strengths and weaknesses in each approach and provides insights into their potential impact on student learning.
3) *Design of assessment*: Bloom's taxonomy guides the creation of assessment questions that target specific cognitive domains. This ensures that the evaluation is well-balanced and covers a range of cognitive levels, providing a comprehensive understanding of students' learning outcomes.
4) *Interpretation of results*: The taxonomy enables a more nuanced interpretation of the study's results by categorizing students' responses according to their cognitive complexity. This aids in determining the extent to which each teaching method is effective in fostering higher-order thinking skills.

III. Results

Student responses were evaluated using four rating categories: outstanding, good, acceptable, and no answer, as depicted in Table 3. To compare the performance of students in each group, we first analyzed the percentage of answered questions within each cognitive domain for both groups relative to the total number of questions. Considering that there are four questions (including 4 sub-questions) in the assessment, the total number of responses in the experimental and control groups should be 120 and 114, respectively. Table 3 presents these findings.

Table 3. Comparison of student responses in cognitive domains: ChatGPT vs. traditional teaching

	Knowing	Applying	Reasoning	No Answer	Total
ChatGPT	15 (12.5%)	28 (23.33%)	10 (8.33%)	67 (55.83%)	120
Traditional Teaching	18 (15.79%)	49 (42.98%)	20 (17.54%)	27 (23.68%)	114

To gain deeper insights into the quality of students' performance across each cognitive domain, we examined the students' responses in both groups that were rated as "Good" and "Outstanding" based on the evaluation rubric. This analysis allowed us to better understand the effectiveness of ChatGPT and traditional teaching methods in fostering higher levels of understanding and application of knowledge. Table 4 showcases the results of this comparative analysis, highlighting the distribution of "Good" and "Outstanding" responses within each cognitive domain for both the ChatGPT and teacher-led groups.

Table 4. Comparison of "good" and "outstanding" student responses in cognitive domains: ChatGPT vs. traditional teaching

	Knowing	Applying	Reasoning
ChatGPT	7/20 (35%)	14/60 (23.33%)	6/40 (15%)
Traditional Teaching	14/19 (73.68%)	28/57 (49.13%)	10/38 (26.32%)

A. Knowing

According to Bloom's taxonomy, the knowing domain is the first of six levels of cognitive complexity. This domain involves the ability to recall or recognize information, facts, or concepts and serves as the foundation for higher levels of cognitive complexity, providing the basis for comprehension, application, analysis, synthesis, and evaluation. In the knowledge level of Bloom's Taxonomy, the primary purpose of questions is to assess whether a student has acquired specific information from the lesson. Some question verbs used at this level include define, list, state, distinguish, name, and ask about who, when, where, and what [20]

Bloom's taxonomy is frequently employed in curriculum design, learning activities, and assessments to structure learning objectives. While the knowing domain is considered the lowest level of critical thinking skills, it is necessary to begin at this level to build more complex thinking and tasks [20].

In this study, the teacher included one question from the "knowing" domain. In the experimental group, 5 out of the 20 students left this question unanswered, while in the control group, only one student left the question empty. Furthermore, according to the evaluation rubric, the quality of the answers in the control group was better than that in the experimental group. Table 5 displays the students' responses in both groups for the question targeting the "knowing" domain. According to student feedback, nearly 47% of the responses provided by the experimental group were rated as good or outstanding. In contrast, the control group demonstrated a higher performance with approximately 77% of their answers achieving the same rating.

Table 5. Comparison of student responses in the knowing domain: ChatGPT vs. traditional teaching

	Outstanding	Good	Acceptable	No Answer	Total
ChatGPT	1 (5%)	6 (30%)	8 (40%)	5 (25%)	20
Traditional Teaching	2 (10.53%)	12 (63.16)	4 (21.05%)	1 (5.26%)	19

B. Applying

The "applying" domain is the third level in the hierarchy of cognitive domains as outlined in Bloom's taxonomy [20] and illustrated in Fig. 1. This category involves the use of acquired knowledge, critical thinking, and problem-solving skills to address new or hypothetical problems within real-world contexts.

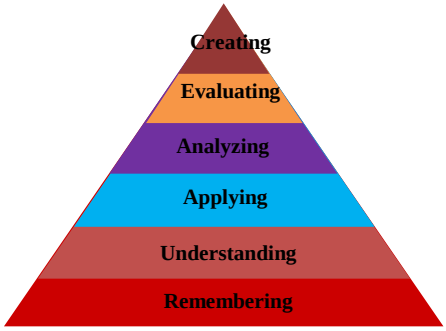


Fig. 2. The hierarchy of cognitive domains outlined in Bloom's taxonomy.

Application questions require students to actively apply the knowledge they have learned. They are presented with a problem or situation where they need to use the information from the class to develop a practical solution. Question verbs commonly used in application questions include: how could x be used to y? How might you demonstrate, utilize, modify, design, explain, or apply x to situations y? [21]. The application domain, in particular, encourages students to transfer and apply their acquired knowledge to diverse situations.

Employing the application domain in teaching methodologies enables educators to stimulate higher-order thinking skills among their students. This practice fosters critical thinking and problem-solving skills, thereby enhancing the overall learning experience [20].

The assessment used in this study comprised three questions designed to evaluate the application domain. Consequently, the total number of responses in the experimental group should amount to 60, while in the control group, it should be 57. Following the teacher's evaluation of the students' answers, it was found that the experimental group had 32 unanswered responses, while the control group had 8 unanswered responses. Table 6 presents the students' responses to these questions.

Table 6. Comparison of student responses in the applying domain: ChatGPT vs. traditional teaching

	Outstanding	Good	Acceptable	No Answer	Total
ChatGPT	3 (5%)	11 (18.33%)	14 (23.33%)	32 (53.33%)	60
Traditional Teaching	4 (7.02%)	24 (42.11%)	21 (36.84%)	8 (14.04%)	57

After excluding the unanswered responses, it was observed that in the experimental group, 50% of the attempted responses were evaluated as Good and Outstanding. Similarly, approximately 57% of the responses in the control group received the same positive evaluation. It is worth noting that while both groups received comparable evaluations, the number of unattempt questions in the experimental group was four times greater than in the control group.

C. Reasoning

Bloom's Taxonomy is a well-known framework that categorizes educational learning objectives into different levels of complexity and specificity. The original taxonomy consists of six categories: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation [20]. The higher-order thinking skills are found in the top three levels, namely Analysis, Synthesis, and Evaluation, which collectively represent the reasoning domain. The reasoning domain emphasizes cognitive processes that involve analyzing, synthesizing, and evaluating information.

At the analysis level, students are expected to move beyond mere information and application, and instead identify patterns that can be used to understand a problem. For instance, an English teacher might inquire about the underlying motives behind the protagonist's actions in a novel. This requires students to analyze the character and draw conclusions based on their observations. Question verbs commonly used at this level include differentiate, compare/contrast, identify differences between x and y, explore the impact or relationship between x and y, why, how, and what element of x is missing or needed. . In the synthesis stage, students are expected to utilize given facts to generate new ideas or make predictions. They may be required to draw from knowledge across different subjects and combine this information to reach a final outcome. For instance, when students are asked to imagine a new product or game, they are being prompted to integrate various elements. Question verbs commonly used in this stage include design, construct, generate, elaborate, envision, produce, modify, and compose, such as writing a short story and naming its accompanying components.. Lastly, at the highest level of Bloom's Taxonomy, evaluation tasks students with assessing information and reaching a conclusion, such as determining its value or the bias involved. Question verbs commonly used at this level include justify, assess, evaluate, judge x based on given criteria. Which option would be superior/preferred over alternative y? [21].

In this study, there were two questions specifically targeting the reasoning domain. Based on the students' responses, 30 responses were left unanswered in the experimental group, while the control group recorded 18 unanswered responses. Table 7 provides an overview of the evaluation of students' responses to the reasoning questions included in the assessment, assessing their comprehension of the topic.

Table 7. Evaluation of student responses in the reasoning domain: ChatGPT vs. traditional teaching

	Outstanding	Good	Acceptable	No Answer	Total
ChatGPT	3 (7.5%)	3 (7.5%)	4 (10%)	30 (75%)	40
Traditional Teaching	2 (5.3%)	8 (21.1%)	10 (26.3%)	18 (47.4%)	38

After excluding the unanswered responses, it is evident that the experiment group has a higher percentage of Good and Outstanding responses compared to the control group (60% and 50% respectively). However, it is important to note that the experimental group also has a greater number of unattempted responses in comparison to the control group. In other words, while the experiment group shows relatively better performance in terms of response quality, it also exhibits a higher number of unattempted responses compared to the control group.

IV. Discussion

The findings of this study provide valuable insights into the potential of utilizing ChatGPT, an AI-powered chatbot, as a substitute teacher in the context of teaching chemistry to eleventh-grade students following the American curriculum in the United Arab Emirates. The study employed Bloom's taxonomy to assess student performance across different cognitive domains, including knowing, applying, and reasoning.

In the knowing domain, both groups showed similar evaluations, although the experiment group had a higher number of unanswered questions. This suggests that while the experiment group had comparable understanding, they faced challenges in providing responses. On the other hand, the control group had the advantage of being able to express their understanding and needs to the subject teacher, who could provide individualized support according to each student's requirements. The implementation of cooperative learning strategies in the control group contributed to enhancing students' higher-order thinking skills [22]. For instance, the teacher implemented the "Quiz-Quiz-Trade" strategy in the starter. The students worked in pairs to quiz each other about the properties of coal. When the students are finished and able to coach each other into the right answer, they raise their hand to find another partner to quiz.

In the applying domain, both groups exhibited relatively good performance, with the control group having a slightly higher percentage of Good and Outstanding responses. However, it is worth noting that the experiment group had a significantly larger number of unanswered questions. This indicates that while the experiment group demonstrated potential in applying knowledge, they encountered difficulties in attempting all questions. Cooperative learning methods, known for promoting students' activeness, critical thinking, and creative thinking skills, were effective in the control group and could have contributed to their better performance [22]. The teacher implemented the "Time-Pair-Share" strategy to help students deep in their understanding of the application of the acquired knowledge. Students were provided with the figures of 3 different solids structures, Graphite, Diamond and Coal, and they were asked to determine the properties of these solids to be used in daily life. Students are allocated a specific time to think individually about the question, then, each student selected a partner nearby to share their thoughts or responses with their partner. In the sharing phase, each pair discussed their ideas with the rest of the class using Padlet.

Within the reasoning domain, the experiment group had a higher percentage of Good and Outstanding responses compared to the control group. However, the experiment group also had a larger number of unattempted responses. This suggests that the experiment group showed stronger comprehension and critical thinking skills but struggled with fully engaging with all questions.

The teacher who invigilated the ChatGPT group observed that students received well-descriptive explanations of the topic, sometimes exceeding their expectations. However, it was noted that some students started skimming the topic without going into depth, indicating the need for specific queries to obtain accurate responses. The absence of diagrams or images also posed a challenge, particularly for visual learners who rely on visual aids to comprehend concepts. The subject teacher recommended providing students with proper training on how to generate prompts in ChatGPT to enhance accuracy and understanding of the results.

These findings highlight significant factors to be taken into account when utilizing ChatGPT to support student learning during substitute classes, emphasizing its role as an aid rather than a replacement for a teacher. While the experiment group showed promise in certain aspects, such as higher percentages of quality responses, the higher number of unattempted questions raises concerns about engagement and completion rates. The study highlights the need for further investigation into optimizing the design and implementation of AI-powered chatbots to enhance student participation and completion rates. Additionally, it is crucial to provide students with training on how to formulate precise prompts when interacting with ChatGPT. This will enable them to receive clearer responses tailored to their individual comprehension levels. Cooperative learning strategies employed in the control group proved beneficial in promoting higher-order thinking skills.

Additionally, ethical considerations and potential limitations of using AI chatbots in educational settings should be taken into account. Privacy, data security, and the potential for bias are important areas that require careful attention to ensure a safe and inclusive learning environment. Recommendations include training students on how to effectively use ChatGPT, incorporating visual aids to support different learning styles, and providing comprehensive support to enhance their learning experiences.

This study provides valuable insights into the use of ChatGPT as a substitute teacher in chemistry education. It emphasizes the strengths and limitations observed in both the experiment and control groups, highlighting the importance of personalized support, cooperative learning strategies, engagement, and training for effective utilization of AI chatbots in education. By addressing these considerations, educators can harness the potential of AI-powered chatbots to

enhance teaching and learning experiences while ensuring a supportive and inclusive educational environment.

The study has several limitations that should be taken into consideration when interpreting the findings. Firstly, the sample size was relatively small, consisting of only 39 students from a single grade level and a specific educational context. This limited sample size raises concerns about the generalizability of the findings to a broader population.

Secondly, the study had a gender and background bias, as it included only male students from diverse backgrounds. This lack of gender diversity and potential bias in participant selection may limit the applicability of the findings to a more diverse student population.

Furthermore, the study focused on the American curriculum in the United Arab Emirates, which may not reflect other educational systems or subjects. The findings may not be directly transferable to different curricula or teaching environments.

Another limitation is the lack of long-term evaluation. The study did not explore the sustained impact of using ChatGPT as a substitute teacher over an extended period. Assessing the long-term effects on students' learning outcomes and engagement would provide a more comprehensive understanding of the benefits and limitations of AI-powered chatbots in educational settings taking into account the controversial nature of ChatGPT and its application in the educational field and the needs to align its use with the school system's policies and guidelines.

Additionally, the evaluation rubric used to assess student responses may have inherent subjectivity or limitations. Different evaluators may interpret "Good" or "Outstanding" responses differently, potentially introducing inconsistency in the assessment process.

Lastly, the study focused on a specific set of cognitive domains within Bloom's taxonomy, which may not fully capture the breadth of cognitive skills or other important aspects of student learning.

Considering these limitations, future research should aim to address these concerns by conducting studies with larger and more diverse samples, exploring different curricula and subjects, assessing long-term impacts, and employing robust evaluation methods to enhance the understanding of the potential of AI-powered chatbots as substitute teachers in various educational contexts.

V. Conclusion

In conclusion, this study highlights the potential of using ChatGPT, an AI-powered chatbot, as a substitute teacher in specific educational contexts. While it demonstrates promise in engaging students and facilitating learning, limitations such as student engagement and completion rates need to be addressed. Traditional teaching methods still have strengths, indicating the potential for a balanced approach that combines AI-powered tools with traditional teaching.

Moving forward, future research should focus on larger and more diverse studies to enhance generalizability. Long-term evaluation is necessary to understand the sustained impact of using ChatGPT as a substitute teacher on student learning outcomes and engagement. Optimization of AI-powered chatbots is vital to improve design and implementation, considering challenges related to student engagement and response quality. Establishing ethical guidelines and pedagogical strategies is crucial for addressing privacy concerns, mitigating biases, and maintaining a balance between human interaction and technology. Providing teacher training and support in utilizing AI-powered tools effectively is also important for enhancing pedagogical practices. By pursuing these future directions, educators and researchers can further explore the benefits, overcome limitations, and optimize the use of AI-powered chatbots as substitute teachers, ultimately enhancing the teaching and learning experiences in educational settings.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

SG and NZ conceived the idea, conducted the literature review, prepared the methodology, and designed the framework for the study. SG performed the experimental work, results analysis, and interpretation. SG and NZ collaborated on synthesizing the information and writing the manuscript. SH and LA provided critical revisions and guidance throughout the writing process. All authors approved the final version of the manuscript for submission.

References

[1] J. Liu, "America faces a substitute teacher shortage—and disadvantaged schools are hit hardest," *Brookings*, Oct. 21, 2020.

[2] Substitute Teaching Preparation: Classroom Management. *UniversalClass.com*. [Online]. Available: https://www.universalclass.com/articles/self-help/substitute-teaching-preparation-classroom-management.htm

[3] How to maintain control of a class as a substitute. *ThoughtCo*. [Online]. Available: https://www.thoughtco.com/classroom-management-tips-for-substitute-teachers-8286

[4] E. Okekwuo and U. Udeani, *Challenges Students Teachers Face during Teaching Practice*, 2021. doi: 10.13140/RG.2.2.30435.27688

[5] S. Pfattheicher, L. B. Lazarević, Y. A. Nielsen, E. C. Westgate, K. Krstić, and S. Schindler, "I enjoy hurting my classmates: On the relation of boredom and sadism in schools," *Journal of School Psychology*, vol. 96, pp. 41–56, Feb. 2023. doi: 10.1016/j.jsp.2022.10.008

[6] A. M. A. Ausat, B. Massang, M. Efendi, and Y. Riady, "Can Chat GPT replace the role of the teacher in the classroom: A fundamental analysis," vol. 05, no. 04, 2023.

[7] (Feb. 17, 2023). Five ways teachers can integrate ChatGPT into their classrooms today. *Penn Today*. [Online]. Available: https://penntoday.upenn.edu/news/five-ways-teachers-can-integrate-chatgpt-their-classrooms-today (accessed Mar. 27, 2023).

[8] T. Trust, J. Whalen, and C. Mouza, "Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education," vol. 23, no. 01, 2023.

[9] R. P D. Santos, "Enhancing chemistry learning with ChatGPT and bing chat as agents to think with: A comparative case study," *SSRN Journal*, 2023. doi: 10.2139/ssrn.4447416.

[10] N. H. Jaya, N. R. Mahyan, S. M. Suhaili, M. N. Jambli, and W. S. W. Ahmad, "Development of conversational agent to enhance learning experience: Case study in pre university," no. 02, 2021.

[11] D. Lee and S. Yeo, "Developing an AI-based chatbot for practicing responsive teaching in mathematics," *Computers & Education*, vol. 191, 104646, Dec. 2022. doi: 10.1016/j.compedu.2022.104646

[12] M. C. Sáiz-Manzanares, R. Marticorena-Sánchez, L. J. Martín-Antón, I. González Díez, and L. Almeida, "Perceived satisfaction of university students with the use of chatbots as a tool for self-regulated learning," *Heliyon*, vol. 9, no. 1, p. e12843, Jan. 2023. doi: 10.1016/j.heliyon.2023.e12843

[13] C.-C. Liu, M.-G. Liao, C.-H. Chang, and H.-M. Lin, "An analysis of children' interaction with an AI chatbot and its impact on their interest in reading," *Computers & Education*, vol. 189, 104576, Nov. 2022. doi: 10.1016/j.compedu.2022.104576

[14] K. Guo, J. Wang, and S. K. W. Chu, "Using chatbots to scaffold EFL students' argumentative writing," *Assessing Writing*, vol. 54, 100666, Oct. 2022. doi: 10.1016/j.asw.2022.100666

[15] H. G. Andrade, "Teaching with rubrics: The good, the bad, and the ugly," *College Teaching*, vol. 53, no. 1, pp. 27–30, 2005.

[16] (Apr. 05, 2023). How to use rubrics to assess your child. [Online]. Available: https://hslda.org/post/how-to-use-rubrics-to-assess-your-child

[17] A. A. Lipnevich, L. N. McCallen, K. P. Miles, and J. K. Smith, "Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment," *Instr Sci*, vol. 42, no. 4, pp. 539–559, Jul. 2014. doi: 10.1007/s11251-013-9299-9

[18] (Apr. 05, 2023). Use rubrics to communicate with students and assess their work. [Online]. Available: https://ifd.vanguard.edu/resources/using-rubrics-to-communicate-expectations-assess-student-work-and-provide-feedback/

[19] D. Allen and K. Tanner, "Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners," *LSE*, vol. 5, no. 3, pp. 197–203, Sep. 2006. doi: 10.1187/cbe.06-06-0168

[20] *Bloom's Taxonomy - Ebook*.

[21] G. N. R. Prasad, "Evaluating student performance based on bloom's taxonomy levels," *J. Phys.: Conf. Ser.*, vol. 1797, no. 1, 012063, Feb. 2021. doi: 10.1088/1742-6596/1797/1/012063

[22] R. Hasan, M. Lukitasari, S. Utami, and A. Anizar, "The activeness, critical, and creative thinking skills of students in the Lesson Study-based inquiry and cooperative learning," *JPBI*, vol. 5, no. 1, pp. 77–84, Mar. 2019. doi: 10.22219/jpbi.v5i1.7328

Copyright © 2024 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).