MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2025, 21(2), em2582
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/15932

Evaluating artificial intelligence large language models' performances in a South African high school chemistry exam

Samuel Jere [1*]

[1] University of Venda, Thohoyandou, SOUTH AFRICA

Received 23 August 2024 ▪ Accepted 05 December 2024

Abstract

Gemini, ChatGPT Plus, and Claude 3.5 Sonnet are artificial intelligence (AI) chatbots with potential in education. Their capabilities, such as acting as virtual teaching assistants, offering personalized responses to learners' queries, and summarizing content, make them versatile tools with the potential to assist learners. The chemistry section of physical sciences in South Africa is often considered challenging, and learners could benefit from virtual teaching assistants to supplement traditional instruction. However, little is known about AI chatbots' abilities in solving high school chemistry problems. This descriptive case study examined the capabilities of Gemini, Claude 3.5 Sonnet, and ChatGPT Plus in accurately answering questions from the final grade 12 physical sciences chemistry exam in South Africa. The conceptual framework that guided the study was Bloom's taxonomy of educational objectives. The responses were rigorously evaluated using the same criteria and rubrics applied to the candidates that year, ensuring a fair and robust comparison. The findings were that ChatGPT Plus performed at 47%, Gemini at 51% and Claude 3.5 Sonnet at 65%. All chatbots performed above the average performance of the candidates who sat for the paper that year, which was 46%. This has significant implications for policymakers, teachers, and learners regarding integrating large language models in teaching physical sciences and exam preparation.

Keywords: large language models, physical sciences, chemistry, performance, grade 12 physical sciences chemistry exam

INTRODUCTION

The chat generative pre-trained transformer (ChatGPT), developed by OpenAI, Claude 3.5 Sonnet, created by Anthropic and Gemini, developed by Google, are artificial intelligence (AI), large language models (LLMs) trained on vast data. They can interact with users by answering questions, engaging in conversations, and offering text explanations. Their capabilities have led to widespread application in various fields, including education. In the context of high school education, they can explain complex chemistry concepts, provide additional practice problems, offer personalized feedback and can be used in automatic scoring of learners' work (Chiu et al., 2023; Grassini, 2023; Lee et al., 2024). The possibility that AI chatbots like ChatGPT and Gemini could greatly improve comprehension and achievement in difficult areas such as chemistry is not merely speculative but can enhance conventional classroom teaching. Learners can use them as personal teaching assistants, improving their academic performance and fostering a deeper understanding of the subject matter. This hopeful perspective on the future role of AI in education goes beyond using them as mere memorization aids, suggesting a transformative shift in our teaching and learning methods.

ChatGPT has evolved from GPT 1 to GPT 3.5 from 2019 to 2022 (Bahrini et al., 2023). The current version is GPT 4, and ChatGPT Plus is a paid version based on GPT 4. This AI chatbot was refined in this evolution and enhanced its capabilities (Tong et al., 2023). Similarly, Claude was developed by Anthropic via several iterations, and currently, Claude 3 is the most recent model upon which Claude 3.5 Sonnet is based. On the other hand, Google AI developed its natural language processing model, which debuted in March 2023 as Bard

✉ samuel.jere@univen.ac.za **(*Correspondence)**

Contribution to the literature

- AI LLMs are becoming widely available, and learners are utilizing these in their studies despite little being known about their abilities to offer accurate answers to high school chemistry questions.
- The study provided empirical evidence for the accuracy of three LLMs when generating answers to a high school chemistry examination question paper.
- The study has implications for the integration of LLMs in chemistry education.

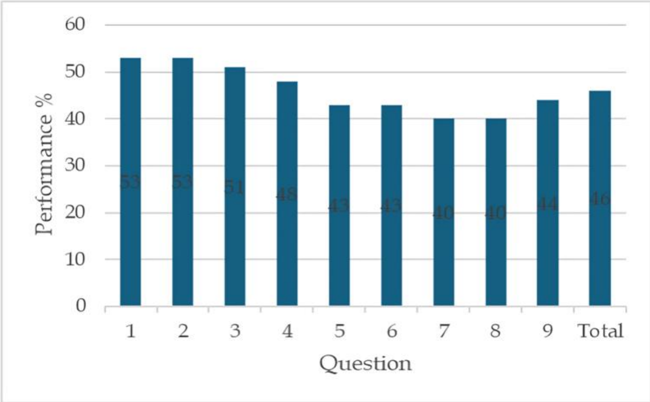


Figure 1. Average performance of candidates per question in chemistry in November 2023 (DBE, 2023)

and was upgraded to Gemini in February 2024 (Carlà et al., 2024). These three LLMs have similar capabilities. Hence, they have the potential to assist learners in ameliorating the challenges they encounter in conceptual understanding of chemistry. Determining their capabilities and limitations can assist teachers in integrating them into chemistry education.

In South Africa, chemistry and physics are offered as a part of the physical sciences. Learners who decide to study physical sciences take the grade 12 final examinations, which comprise two papers–physics and chemistry. The chemistry paper tends to be difficult for both learners and teachers. Although the candidates who sat for the chemistry paper in 2023 performed poorly in all questions (see **Figure 1**), examiners singled out questions on acids and bases, galvanic cells and electrolytes (Department of Basic Education [DBE], 2023a). The overall performance of candidates for the chemistry paper was 46% (**Figure 1**).

Anecdotal evidence suggests that learners depend on ChatGPT as a virtual teaching assistant in learning chemistry, although its capabilities have not been confirmed. There is a lack of empirical research assessing the accuracy of answers provided by AI chatbots at the high school level. A literature search found no studies that have assessed these chatbots within the South African context. This is the gap that this study attempted to fill. It is, therefore, important to examine the accuracy of the responses of the chatbots to chemistry questions. The study aimed to determine the types of questions that Gemini, Claude and ChatGPT can answer accurately in grade 12 chemistry, study the types of questions the chatbots struggle to answer and compare the performance of the three chatbots with the performance of the candidates that year.

LITERATURE REVIEW

As LLMs, such as ChatGPT, become widely available, it is necessary to determine their capabilities in education to find out if they can be meaningfully integrated into teaching and learning. These chatbots have the potential to be used as teaching assistants, educators or trainers, and in a classroom situation, they can assist the teacher by freeing time that can be used for other tasks (Pérez et al., 2020). Examining the capabilities of the chatbots in responding to chemistry questions is necessary to determine their accuracy in chemistry education and evaluation (Fergus et al., 2023).

Extant literature reveals that learners are using ChatGPT to obtain feedback on their queries in various learning areas, seek clarifications on difficult concepts and assist them in homework and assignments (Albadarin et al., 2024). Doğru (2023) goes further to suggest that learners should leverage the capabilities of chatbots in preparing for examinations. This may be reasonable given that research has confirmed the role of ChatGPT as useful in providing content, evaluating learners' work, and assisting teachers in various situations, while teachers consider it important for making choices regarding pedagogical approaches and making learning more engaging and active (Jeon & Lee, 2023; Jere et al., 2024). Due to the multifaceted roles that ChatGPT can play in education, there is a need for more research to understand the potential benefits and limitations of ChatGPT and other chatbots in education in general and specifically in learning chemistry.

In chemistry, leveraging AI chatbots can enhance learners' abilities to solve problems by offering explanations, detailed step-by-step solutions, and engaging supplementary questions about problems, which assist learners in understanding the rationale of solutions, enhancing their critical and innovative thinking abilities (Kasneci et al., 2023). Educators can leverage LLMs to craft individualized learning paths for their pupils. These chatbots can assess learners' writings and answers, deliver customized responses and recommend resources that match the learners' unique educational requirements, which assist in freeing up teachers' time and energy to concentrate on other teaching activities like developing captivating and interactive activities (Kasneci et al., 2023).

Some studies have focused on assessing ChatGPT's performance in answering chemistry questions at various levels of education. For example, Fergus et al. (2023) investigated the abilities of ChatGPT to answer questions in year-end exams in first and second-year pharmacy courses at the university level. Their study revealed that ChatGPT could competently answer questions at the recall and comprehension levels on Bloom's taxonomy of educational objectives. At the same time, it struggled to answer questions at the application and analysis level. ChatGPT could not answer questions where chemical structures were presented as figures. They also found that ChatGPT could not answer questions requiring drawing chemical structures or plotting graphs.

Daher et al. (2023) studied the capabilities of ChatGPT in answering chemistry questions in a college course on material science. Like Fergus et al. (2023), they found that ChatGPT could satisfactorily answer lower-order questions at the remembering level in Bloom's taxonomy. Beyond remembering, ChatGPT faced significant challenges in questions involving representation and requiring depth. Leite (2024) investigated the performance of Copilot, Gemini and ChatGPT in defining some chemistry concepts. The study revealed that these chatbots could generate comprehensive responses to the chemistry questions asked, implying that these chatbots have great potential in elucidating chemistry concepts. In light of the above studies, this research was carried out to evaluate the performance of Google Gemini, Claude 3.5 Sonnet and ChatGPT Plus in answering the grade 12 final examination question paper and compare the responses of the three chatbots.

While Gemini, Claude and ChatGPT are LLMs, differences in their performance in the chemistry exams would likely be observed. This is due to the differences in their architectures and the data upon which they were trained. On the one hand, ChatGPT is based on the pre-trained transformer architecture, while on the other, Gemini utilizes Google's language model dialogue-neutral architecture (Makrygiannakis et al., 2024). The differences in architectures and type and quantity of training data have been cited as possible causes of differences in the responses of these chatbots to the same queries (Makrygiannakis et al., 2024). Hence, it can be expected that the three chatbots may have different capabilities in answering chemistry problems.

This study aimed to determine the accuracy of Gemini, Claude 3.5 Sonnet and ChatGPT Plus in responding to high school chemistry questions. To achieve this aim, the following research questions were raised: How accurate was Gemini, Claude 3.5 Sonnet and ChatGPT Plus in responding to high school chemistry questions? Which types of questions were the chatbots more competent to answer, and which were more challenging? In seeking to answer these questions, the researcher was guided by Bloom's taxonomy of educational objectives as the conceptual framework.

Conceptual Framework: Bloom's Taxonomy

The study was guided by Bloom's taxonomy of educational objectives (Bloom et al., 1956). The cognitive domain in Bloom's taxonomy is represented by lower or higher-order thinking. Anderson and Krathwohl (2001) revised and updated the original Bloom's taxonomy in light of the criticisms raised against it. In the revised version, lower-order thinking consists of those objectives at the remembering, understanding and applying levels, while higher-order thinking involves analyzing, evaluating and creating (Hutton-Prager, 2018).

Examinations, such as the chemistry exam paper in this study, are assessment tools used to measure the extent to which teaching/learning objectives were achieved. Therefore, the examination questions align with learning objectives (Bonaci et al., 2013). The questions in these examinations fall under the cognitive domain in Bloom's taxonomy. The revised version of Bloom's taxonomy has two dimensions–knowledge levels and cognitive processes (Anderson & Krathwohl, 2001). The knowledge dimension has four levels. These levels, from basic to complex, are factual, conceptual, procedural, and metacognitive. The cognitive domain comprises 19 cognitive processes under the six cognitive levels (Anderson & Krathwohl, 2001).

It would, therefore, be expected that the questions can be placed in the four knowledge levels and classified into the 19 cognitive processes. The questions in the *remembering* cognitive level measure the cognitive processes that require *identifying* and *retrieving* knowledge. Questions requiring *understanding* involve mental processes such as *inferring*, *comparing*, and *explaining*, to mention a few. *Applying* involves mental processes requiring learners to *execute* or *implement* a procedure. Analyzing involves cognitive processes such as *differentiating* and *organising* while *evaluating* requires *checking* and *critiquing,* and finally, *creating* involves thinking processes like *generating*, *planning* and *producing* (Anderson & Krathwohl, 2001). The examination questions in this study were classified into the revised Bloom's taxonomy, and the findings were discussed using this conceptual framework.

MATERIALS AND METHODS

Research Design

The study used an interpretive research approach, utilizing inductive methods, beginning with data and attempting to generate theory from the studied phenomena (Bhattacherjee, 2012). Bhattacherjee (2012) states that interpretive research primarily relies on qualitative data, although it does incorporate quantitative data to a lesser degree when needed. This was a case study in which descriptive statistics were used to assess the performances of ChatGPT Plus, Claude 3.5 Sonnet and Google Gemini in the chemistry examination. Qualitative data analysis was used to interpret the responses of these AI technologies to support or refute the results in the descriptive statistics. In the qualitative phase, purposive sampling was used to select the analyzed questions (**Appendix A**). In selecting these questions, consideration was given to those that reflect the chatbots' below-average, average, and above-average performance.

Context of the Study

In South Africa, chemistry is offered as part of physical sciences. It is studied from grade 10 to grade 12 under three themes: matter and materials, chemical change and chemical systems. Grade 12 mainly focuses on introducing chemistry concepts under various topics, such as organic molecules, rate and extent of reaction, chemical equilibrium and electrochemistry. In the final grade 12 exam, used to admit candidates to various tertiary courses, chemistry is offered as a separate paper from physics in the physical sciences. Candidates sit for 3 hours for this paper, worth 150 marks. The paper consists of ten multiple-choice items worth 20 marks; the rest are structured questions.

Procedure

The researcher subscribed to ChatGPT Plus to be able to carry out the investigation and created accounts with Claude 3.5 Sonnet and Google Gemini. A November 2023 final exam chemistry paper and marking guidelines were downloaded from the DBE (2023b) website. The researcher uploaded the entire question paper into the ChatGPT Plus, Claude 3.5 Sonnet and Gemini applications. Prompt engineering was used to instruct the LLMs in answering the question paper.

According to Giray (2023), effective prompting of LLMs involves providing the LLM with instructions, context, input data, and output indicators. The *instruction* was for the LLMs to answer all questions, expressing the answers in plain language. The *context* provided to the LLMs was that they had to answer the questions as a candidate in the examination, the *input data* were all the questions in the paper, and the *output indicators* involved the LLMs providing step-by-step solutions in their answers.

The responses generated by the chatbots were printed, and three copies were made for each chatbot. The question paper, ChatGPT's responses, and the DBE (2023a) marking guides were sent to two faculty members who specialized in chemistry. They were requested to use the marking guides to assess the responses from ChatGPT. They completed this task in three days, and the researcher received the results. The researcher used their results to determine the inter-rater reliability (IRR).

The responses generated by the chatbots were discussed under the different cognitive levels in Bloom's taxonomy of educational objectives to determine the types of questions that ChatGPT could answer and those it struggled with. The accuracy of the three chatbots was determined by calculating the percentage of responses deemed to be correct.

Research Instrument

The research instrument was the November 2023 chemistry paper from the DBE (2023b) in South Africa. It consisted of nine questions. Question 1 had ten multiple-choice questions from various topics. The remaining questions were structured in most cases, requiring short answers. Questions two to four were from the organic chemistry section of the physical sciences curriculum document. Question five was based on the rate and extent of reaction, question six on chemical equilibrium, question seven on acid-base chemistry and questions eight and nine on electrochemical reactions. As the examination board standardized this instrument during the development phase, there was no need to validate the instrument. The instrument's validity and reliability were assumed to be acceptable.

Inter-rater reliability of the classification of questions into the levels in Bloom's taxonomy

While the question paper had nine questions, these questions had sixty-eight sub-questions. The researcher enlisted the services of two high school physical sciences teachers to classify these sixty-eight sub-questions into the six levels of Bloom's taxonomy of educational objectives (Krathwohl, 2002). The researcher discussed Bloom's taxonomy with each rater separately to ensure a common understanding of the rating process. The questions were rated using a six-point Likert scale where remembering was represented by one, understanding, two, applying, three, analyzing, four, evaluating, five and creating six. The researcher then requested each rater to complete the rating process. It took each rater one day to complete the rating process. The statistical package for social sciences (SPSS) version 29 was then used to assess the IRR of the ratings from the two raters. The IRR was 0.920, $p < .05$, as displayed in **Table 1**. This implies that there was substantial agreement between the two raters.

The researcher and the two raters met to discuss the areas where there were differences, and guided by the

Table 1. IRR of classification of questions into Bloom's taxonomy

	Statistics
N	68
Kappa value	0.920
p	< .001

Table 2. IRR values for marks obtained by two assessors for Gemini, ChatGPT Plus, and Claude 3.5 Sonnet in structured questions

	Gemini	ChatGPT Plus	Claude 3.5 Sonnet
N	58	58	58
Kappa value	0.880	0.877	0.882
p	< .001	< .001	< .001

initial guidelines for classification, the differences were resolved. The final classifications of the questions into the cognitive levels of Bloom's taxonomy are displayed in the results section.

Inter-rater reliability of the assessment of the responses from the chatbots

To determine the IRR, the SPSS version 29 was used to evaluate Cohen's kappa. All sub-questions, excluding multiple choice items in the chemistry paper (N = 58) were used to determine IRR. In calculating kappa, the marks assigned by the raters for each chatbot for each sub-question were compared to determine the level of agreement between the two raters. For Google Gemini, kappa was found to be 0.880, $p < .05$; for ChatGPT Plus, kappa was 0.877, p < .05; and for Claude 3.5 Sonnet, kappa was 0.882, $p < .05$. The results are presented in **Table 2**. These results indicate strong IRR (Landis & Koch, 1977).

The researcher and the two assessors met to discuss the results. The marking guideline was used to resolve the differences in their ratings. The final results are shown in **Table 3**.

RESULTS

Table 3 shows the classifications of the questions in the paper according to Bloom's taxonomy of educational objectives (Krathwohl, 2002). The classification in **Table 3** is primarily based on the cognitive processes in the revised Bloom's taxonomy (Krathwohl, 2002). The categories of the knowledge dimension which intersect with these cognitive processes include factual knowledge, conceptual knowledge and procedural knowledge (Wilson, 2016). In this study, the factual knowledge dimension was associated with the cognitive processes of *remembering* and *understanding*. The conceptual knowledge dimension was linked to *applying* and *analyzing,* while the procedural knowledge dimension was regarded as intersecting with *analyzing* and *creating*. **Table 3** also gives the percentage of marks each of the three chatbots obtained for the different cognitive levels in Bloom's taxonomy.

The results in **Table 3** indicate that Claude 3.5 Sonnet was more accurate at answering the chemistry questions, obtaining an overall mark of 65%. On the other hand, Gemini (51%) and ChatGPT Plus (47%) appeared to have similar accuracy overall. The performance of the three chatbots was above the average performance of the candidates who sat for this paper, which was 46% (DBE, 2023a). The responses generated by all the chatbots were outstanding at the remembering level in Bloom's taxonomy. All the responses generated at that level were satisfactory. This could be because the questions at the remembering level required chatbots to retrieve this information from the large data store upon which they were trained. Chatbots faced significant challenges regarding questions at all higher levels beyond remembering. The chatbot could only generate half of the responses at the understanding and analysis levels. ChatGPT Plus and Gemini could not get more than half of the questions correct in the application, evaluation and creation levels.

Remembering

The questions from the topic of organic molecules at the remembering level required learners to be able to define organic compounds (2.1), boiling point (3.1), cracking (4.1.1), and positional isomers (4.2.1). The questions are shown in **Figure 2**.

For the definition of an organic compound, the responses generated by the chatbots were:

Table 3. Marks obtained by chatbots per cognitive level in Bloom's taxonomy

Domain-specific cognitive level	Questions	Total marks	Gemini marks	ChatGPT Plus marks	Claude AI marks
Level 1: Remembering	1.1, 1.2. 1.3, 2.1, 3.1, 4.1.1, 4.2.1, 5.1, 6.1, 6.2, 7.1, 8.1.2, 8.3.1, 8.3.2, 8.3.3, & 9.1	28	27 (96%)	28 (100%)	27 (96%)
Level 2: Understanding	1.5, 1.9, 2.3.1, 2.3.2, 2.3.3, 2.3.4, 3.4.1, 4.2.2, 4.2.5, 4.2.7, 5.2, 8.1.1, & 9.3	19	14 (74%)	9 (35%)	14 (74%)
Level 3: Applying	1.4, 1.8, 1.10, 2.2.1, 2.2.2, 2.3.5, 3.5.1, 3.5.2, 4.1.2, 4.1.3, 4.2.3, 4.2.4, 4.2.6, 5.3, 5.4, 5.5, 5.6, 6.3.2, 6.4, 7.2.1, 7.2.2, 9.2, & 9.4	63	25 (40%)	22 (35%)	32 (51%)
Level 4: Analyzing	1.6, 1.7, 2.4.1, 2.4.2, 3.2, 3.3, 3.4.2, 3.6, 6.3.1, 6.6, 7.3, 8.2, 8.4, & 9.5	34	11 (32%)	12 (35%)	19 (56%)
Level 5: Evaluating	6.5.2	3	0 (0%)	0 (0%)	3 (100%)
Level 6: Creating	6.5.1	3	0 (0%)	0 (0%)	2 (67%)
Total		150	77 (51%)	71 (47%)	97 (65%)

Note. IRR Cohen's kappa for assessment of responses: Gemini - 0.880, $p < .05$; ChatGPT Plus - 0.877, $p < .05$; & Claude 3.5 Sonnet - 0.882, $p < .05$.

2.1 Define the term organic compound.

3.1 Define the term boiling point.

4.1.1 Define cracking.

4.2.1 Write down the definition of positional isomers.

Figure 2. Organic chemistry questions at the remembering level (DBE, 2023)

> An organic compound is any chemical compound that contains carbon atoms bonded to hydrogen atoms in chains or rings and may also contain other elements such as oxygen, nitrogen, sulfur, etc. (ChatGPT Plus).

> Organic compounds are chemical compounds that primarily contain carbon atoms bonded together with other elements, most commonly hydrogen, oxygen, nitrogen, and sulfur. They form the basis of life and are found in various natural and synthetic substances (Gemini).

> An organic compound is a molecule that contains carbon, typically bonded to hydrogen, and often including other elements such as oxygen, nitrogen, sulfur, etc. (Claudi 3.5 Sonnet).

The marking guide required candidates to define an organic compound as a molecule/compound containing carbon atoms. Therefore, the chatbot's responses were more elaborate than the marking guide. The same trend of providing elaborate definitions was observed for all questions at the remembering level. For example, for question 4.2.1, they generated the following responses for the definition of positional isomer:

> Positional isomers are compounds with the same molecular formula but different positions of functional groups or substituents on the parent chain (ChatGPT Plus).

> Positional isomers are compounds with the same molecular formula but different positions of a functional group (Gemini).

> Positional isomers are compounds with the same molecular formula but different arrangements of atoms along the carbon chain or different positions of functional groups (Claude 3.5 Sonnet).

These responses were the same as the marking guideline expected candidates to answer the question. The chatbots were adept at generating accurate responses to the questions at the remembering level in Bloom's taxonomy, as reflected in **Table 3**. The accuracy of Gemini and Claude is reflected by obtaining 96% (27 out of the 28 marks) correct for the total marks at the remembering level, while ChatGPT Plus obtained 100% (**Table 3**). The implication is that chatbots are useful tools for learners to revise questions that require recalling factual information. This finding is supported by the descriptive statistics, which showed that the chatbots generated accurate responses to questions that required remembering.

Understanding

The questions that assessed understanding in question 2 and question 3 were 2.3.1, 2.3.2, 2.3.3, 2.3.4, question 3.4.1 (**Figure 3**).

The chatbots struggled to generate accurate responses to questions requiring understanding. These difficulties were noted to be prevalent in questions in

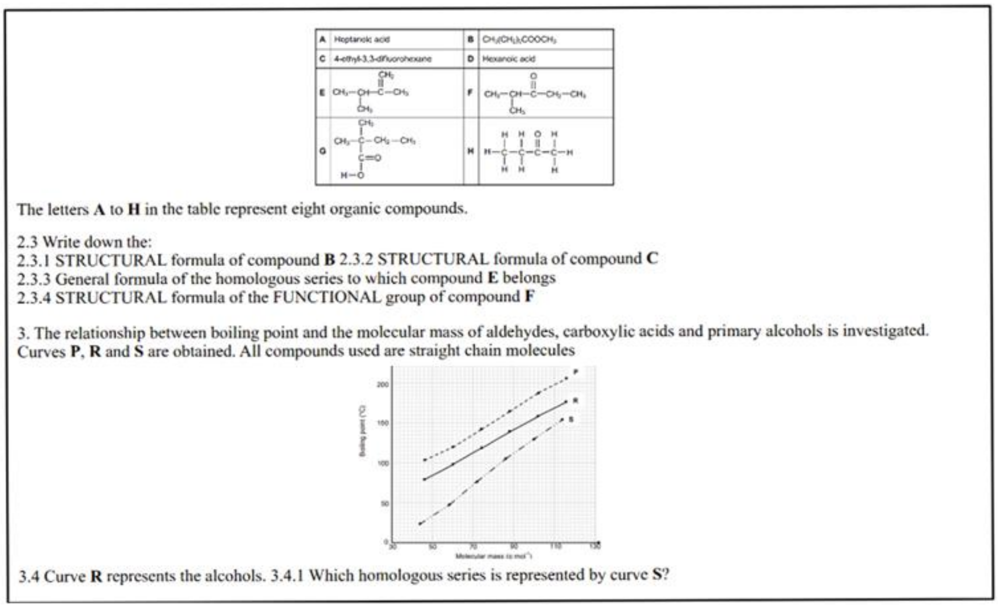


Figure 3. Organic chemistry questions in question 2 and question 3 requiring understanding (DBE, 2023)

4.1 The Compound, C_6H_{14}, undergoes complete combustion. 4.1.3 Using MOLECULAR FORMULAE, write down the balanced equation for this reaction.

4.2 Consider the equations for reactions **I** to **III** below. **A** and **B** represent organic compounds that are POSITIONAL ISOMERS. **X** is an inorganic product.

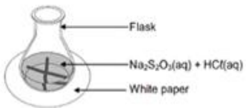
I	$CH_3CH_2CHCHCH_3 + HC\ell \rightarrow$ **A** + **B**
II	**A** $\xrightarrow[\Delta]{H_2O}$ $CH_3CH_2CH_2CH(OH)CH_3$ + **X**
III	$CH_3CH_2CH_2CH(OH)CH_3 \longrightarrow CH_3CH_2CHCHCH_3 + H_2O$

Write down the: 4.2.2 Type of reaction represented by reaction **I**

Compound **A** can be converted directly to the organic product of reaction **III**.

4.2.7 Write down TWO terms that describe this type of reaction.

Figure 4. Organic chemistry questions in question 4 at the understanding level (DBE, 2023)



The reaction between EXCESS dilute hydrochloric acid and sodium thiosulphate is used to investigate factors that influence reaction rate.
$Na_2S_2O_{3(aq)} + 2HC\ell_{(aq)} \rightarrow 2NaC\ell_{(aq)} + S_{(s)} + H_2O_{(\ell)} + SO_{2(g)}$

The concentration of $HC\ell(aq)$ used is 1 mol·dm^{-3} . The same volume of $HC\ell(aq)$ is used in each run. The time taken for the cross on the paper under the flask to become invisible is measured. The table below summarises the reaction conditions and results of the experiment.

RUN	VOLUME $Na_2S_2O_3(aq)$ (cm^3)	VOLUME $H_2O(\ell)$ ADDED (cm^3)	CONCENTRATION $Na_2S_2O_3(aq)$ (mol·dm^{-3})	TIME (s)
1	50	0	0,13	20,4
2	40	10	0,10	26,7
3	30	20	**P**	33,3

5.2 Write down the independent variable for this investigation.

5.3 Calculate the value of **P** in the table.

5.4 When 0,21 g of sulphur has formed in Run 1, the cross becomes invisible. Calculate the average reaction rate with respect to sodium thiosulphate, $Na_2S_2O_{3(aq)}$, in g·s^{-1} .

Another investigation is performed at different temperatures. 5.5 Sketch the Maxwell-Boltzmann distribution curve for the reaction at 20 °C. Label this curve as **A**. On the same set of axis, draw the curve that will be obtained at 35 °C and label it as **B**.

5.6 Explain the effect of temperature on reaction rate in terms of the collision theory.

Figure 5. Rate and extent of reaction questions at the understanding (5.2) and applying level (5.3, 5.4, and 5.6) (DBE, 2023)

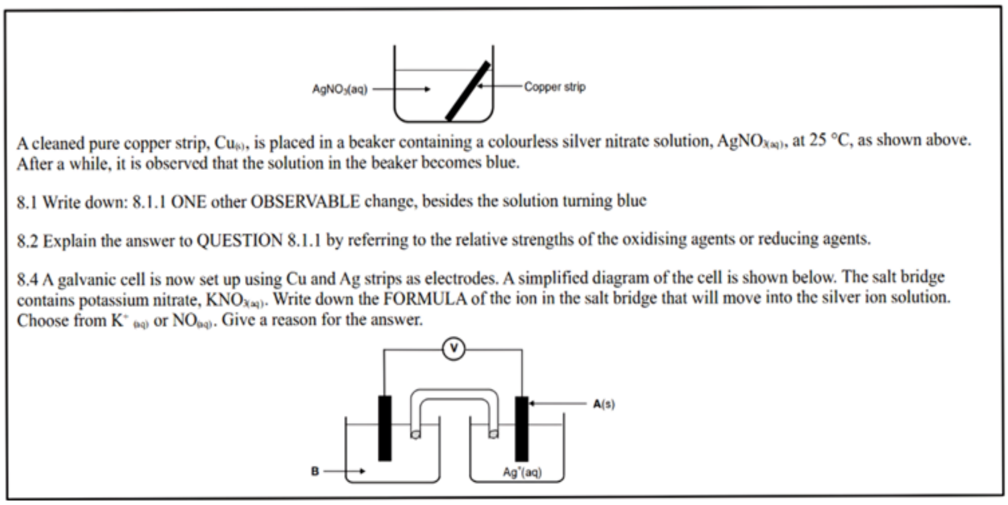
organic chemistry. ChatGPT Plus failed to generate accurate responses to all the sub-questions under question 2 and question 3.

The trend of inaccurate responses to organic questions was also observed in answers from Claude and Gemini. Both chatbots could not write down completely accurate structural formulae for organic molecules as was required in question 2.3.1 and could not understand structural formulae to derive the general formula as in 2.3.3. They could not draw the structural formula of a functional group of a given organic molecule as in 2.3.4. The study could not verify whether the failure to provide accurate responses could have been caused by the chatbots not understanding the structures presented in a table. Confirming whether the chatbots could have accurately answered the questions if only words were used was beyond the present study.

Further questions that needed understanding included questions 4.1.3, 4.2.2, and 4.2.7 (**Figure 4**). ChatGPT and Claude could generate accurate responses to question 4.1.3 on writing balanced equations for the complete combustion of C_4H_{14}, which may be attributed to the data upon which they were trained, while Gemine could not. In summary, the chatbots' failure to generate accurate responses to questions on organic chemistry suggests that they have limited capabilities in this part of chemistry.

In question 5, on the rate and extent of reaction, candidates were provided with a balanced chemical reaction of sodium thiosulphate and hydrochloric acid and a description of how the reaction was carried out. In 5.2 (**Figure 5**), candidates were asked to write down the independent variable for the investigation. All chatbots provided an accurate response by identifying the concentration of sodium thiosulphate as the independent variable. While it may appear like the chatbots could understand the information provided in a table, it is possible that the questions provided sufficient information for the chatbots to work with to produce accurate responses. Further research is needed to verify if chatbots can interpret tables accurately.

Question 8.1.1 (**Figure 6**) was based on electrochemical reactions, and a beaker in which an aqueous silver nitrate solution was reacting with a copper strip was provided. Candidates were asked to provide one other observation apart from the color change of the solution to blue. All chatbots accurately identified the formation of a solid silver deposit on the



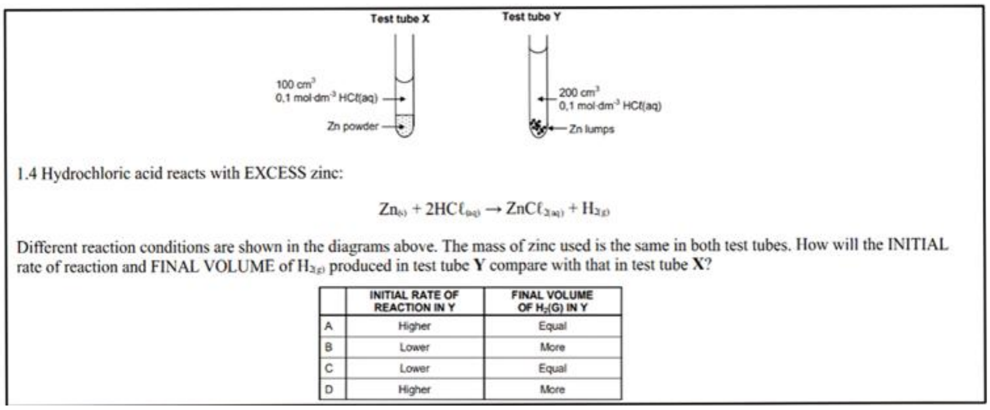
A cleaned pure copper strip, $Cu_{(s)}$, is placed in a beaker containing a colourless silver nitrate solution, $AgNO_{3(aq)}$, at 25 °C, as shown above. After a while, it is observed that the solution in the beaker becomes blue.

8.1 Write down: 8.1.1 ONE other OBSERVABLE change, besides the solution turning blue

8.2 Explain the answer to QUESTION 8.1.1 by referring to the relative strengths of the oxidising agents or reducing agents.

8.4 A galvanic cell is now set up using Cu and Ag strips as electrodes. A simplified diagram of the cell is shown below. The salt bridge contains potassium nitrate, $KNO_{3(aq)}$. Write down the FORMULA of the ion in the salt bridge that will move into the silver ion solution. Choose from $K^{+}_{(aq)}$ or $NO_{3(aq)}^{-}$. Give a reason for the answer.

Figure 6. Electrochemical reaction questions at the remembering (8.1.1) and analysis levels (8.2 and 8.4) (DBE, 2023)



1.4 Hydrochloric acid reacts with EXCESS zinc:

$$Zn_{(s)} + 2HC\ell_{(aq)} \rightarrow ZnC\ell_{2(aq)} + H_{2(g)}$$

Different reaction conditions are shown in the diagrams above. The mass of zinc used is the same in both test tubes. How will the INITIAL rate of reaction and FINAL VOLUME of $H_{2(g)}$ produced in test tube **Y** compare with that in test tube **X**?

	INITIAL RATE OF REACTION IN Y	FINAL VOLUME OF H_2(G) IN Y
A	Higher	Equal
B	Lower	More
C	Lower	Equal
D	Higher	More

Figure 7. Multiple choice question assessing applying (DBE, 2023)

copper strip as the other observation. The chatbots needed to interpret the information in the diagram to answer the question accurately. Their ability to provide accurate responses could not ascertain if the chatbots could comprehend the chemistry information provided in diagrams. This is due to sufficient information provided in the question, which could enable the chatbot to provide accurate responses. Further research is required to confirm if chatbots can comprehend information in chemistry diagrams.

From the responses generated by the chatbots, while they appear to face challenges in answering questions in organic chemistry, their responses in other chemistry sections were satisfactory at the understanding level of Bloom's taxonomy. Severe limitations were noted in understanding the various kinds of formulae, such as structural formulae used in organic chemistry. This is reflected by the accuracy of the chatbots in the percentage of correct responses they obtained (**Table 3**). Gemini and Claude 3.5 Sonnet obtained 74%, while ChatGPT Plus obtained 35% correct responses at the understanding level.

Applying

ChatGPT Plus and Gemini experienced considerable difficulties regarding questions that required applying chemical knowledge to novel situations, although Claude 3.5 Sonnet was more accurate. For example, in question 1.4, the chatbots were given a diagram in **Figure 7**. They were told that the mass of zinc was the same in both test tubes and that zinc was in excess. They were then asked to describe how the initial reaction rate and final volume of $H_{2(g)}$ produced in test tube Y compared with that in test tube X.

The response from ChatGPT Plus was, as follows:

> Answer: C. Lower (in test tube Y); Equal (Zn lumps have a smaller surface area compared to Zn powder, resulting in a lower initial rate. However, the final volume of H_2 depends on the total amount of zinc reacting, which is equal.)

While the response to the first part of the question is that the reaction rate is lower in test tube Y, and the given reason was sound, the second part was incorrect. The second part required an understanding of limiting reactants. As zinc was in excess, the amount of $H_{2(g)}$ produced depended on the amount of $HCl_{(aq)}$.

Similarly, Gemini could correctly state that the initial reaction rate would be lower in Y, but it provided an incorrect response on the final volume of $H_{2(g)}$. This trend of failure to notice that the limiting reactant was HCl was observed in the response from Claude, which was, as follows:

> The correct answer is C) Lower, Equal. With zinc lumps vs powder, the surface area is lower so the initial rate will be lower. But final H_2 volume depends only on amount of zinc, which is equal.

Question 5 was on the rate and extent of reaction (see **Figure 5**). The 1 $mol.dm^{-3}$ of $HCl_{(aq)}$ was reacted with sodium thiosulphate at various concentrations. The time taken for the reaction was measured by the disappearance of an X on a white paper on which the flask with reactants was placed. In question 5.3, they were required to determine the value of **P** (see **Figure 5**). All chatbots were able to generate accurate responses to this question. ChatGPT Plus generated the following response:

> Concentration of $Na_2S_2O_3$ = 30/50 × 0.13 = 0.078 $mol.dm^{-3}$.

In question 5.4 (**Figure 5**), chatbots were informed that 0.21 g of sulfur was formed in run 1, and the cross disappeared. They were then required to calculate the average reaction rate with respect to sodium thiosulphate in $g.s^{-1}$. Only ChatGPT Plus could generate an accurate and complete response to this question. It responded, as follows:

> Find the moles of sulfur formed:
>
> Moles of sulfur formed = mass of sulfur/molar mass of sulfur
>
> Moles of sulfur = 0.21/32.07 = 0.00654 mol
>
> Determine the moles of $Na_2S_2O_3$ reacted:
>
> From the balanced equation, 1 mole of $Na_2S_2O_3$ reacts to produce 1 mole of sulfur
>
> Therefore, moles of $Na_2S_2O_3$ reacted = 0.00654 mol
>
> Calculate the mass of $Na_2S_2O_3$ reacted:
>
> Mass of $Na_2S_2O_3$ = 0.00654 X 158.11 $g.mol^{-1}$=1.03 g
>
> Calculate the average reaction rate:
>
> Average reaction rate = mass of $Na_2S_2O_3$ reacted/time
>
> Average reaction rate =1.03 g/20.4 s = 0.0505 $g.s^{-1}$

Question 5.5 (**Figure 5**) required chatbots to sketch a Maxwell-Boltzmann distribution curve for the reaction at 20 °C and another curve on the same set of axes for the same reaction at 35 °C. Gemini generated the curve in **Figure 8**. **Figure 8** was just retrieved from a website, and the chatbot did not attempt to address the question. The response was deemed inaccurate.

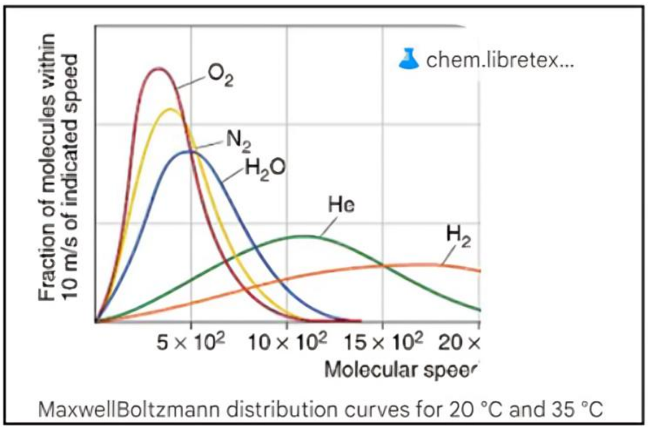


Figure 8. Gemini's answer to question 5.5 (Source: Authors' own elaboration, using Google Gemini)

Claude 3.5 Sonnet did not attempt to draw the curves, implying that the chatbot cannot. However, instead, it offered the following attempt, which was regarded as partially correct:

> I would sketch two curves here, with curve B (35°C) shifted slightly to the right of curve A (20°C), having a lower peak but extending further along the x-axis.

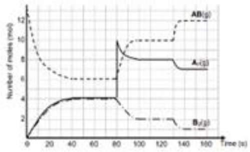
Similarly, ChatGPT did not attempt to draw the curve and offered the following response, also regarded as partially correct:

> A proper curve should show curve B shifted to the right and lower peak than curve A indicating more molecules with higher kinetic energy.

The last sub-question, 5.6 in **Figure 5**, required chatbots to explain the effect of temperature on reaction rate. All chatbots offered complete and accurate answers to this question. For example, Gemini responded, as follows:

> According to the collision theory, for a reaction to occur, reactant particles must collide with sufficient energy (activation energy) and with the correct orientation. Increasing the temperature increases the average kinetic energy of the particles. This leads to a higher proportion of particles with enough energy to overcome the activation energy barrier, resulting in more frequent and successful collisions. Consequently, the reaction rate increases with an increase in temperature.

It was observed that the chatbots displayed elaborate explanations of key chemistry concepts in questions such as 5.6. Some chatbots, such as ChatGPT Plus, could offer accurate responses to questions requiring calculations, such as question 5.4. Notable limitations



Consider the following hypothetical reaction reaching equilibrium in a 4 dm³ closed container at 150 °C.

$$2AB_{(g)} \rightleftharpoons A_{2(g)} + B_{2(g)}$$

The above graph shows the changes in the amounts of reactants and products over time.

6.3 A change was made to the equilibrium mixture at t = 80 s. 6.3.1 Write down the change made at t = 80 s.

6.5 At t = 130 s the temperature of the system is decreased to 100 °C.

6.5.1 Draw a potential energy diagram for this reaction.

6.5.2 Will the equilibrium constant, K_c, at 100 °C be GREATER THAN, LESS THAN or EQUAL TO the K_c at 150 °C? Explain the answer.

6.6 The initial reaction now takes place in the presence of a catalyst at 150 °C. Describe the changes that will be observed on the graph between t = 0 s and t = 60 s.

Figure 9. Questions on the analyzing level in Bloom's taxonomy in question 6 (DBE, 2023)

were observed with all the chatbots responding to questions requiring drawing graphs, such as in question 5.5. It appears that chatbots cannot completely understand the information provided in tables and figures, leading to the inability to offer accurate responses. This led to the chatbots obtaining much lower accuracy, as indicated in **Table 3**, where Gemini had 40%, ChatGPT 35% and Claude 3.5 Sonnet 51%.

Analyzing

Questions at the analyzing level in Bloom's taxonomy proved challenging for the chatbots. To reflect on the challenges and how the chatbot attempted to answer these questions, questions 6.3.1, 6.6, 8.2, 8.4, and 9.5 were used. Question 6 was based on chemical equilibrium (**Figure 9**). Chatbots were given a hypothetical reaction that reached equilibrium in a 4 dm^3 closed container at 150 °C.

In question 6.3, the chatbots were informed that a change was made to the equilibrium mixture at t = 80 s. They were then asked to state the change that was made. Claude 3.5 Sonnet generated the following response:

> The change made was the addition of $A_{2(g)}$ to the system.

This was the accurate response required. Gemini and ChatGPT Plus could not generate accurate responses, providing evidence that they could not analyze and understand the information in the diagram.

In question 6.6 (**Figure 9**), the chatbots were asked to describe the changes observed in the graph between t = 0 s and t = 60 s if the reaction occurred in the presence of a catalyst at the same temperature. All chatbots generated accurate and elaborate responses. For example, Gemini generated the following response:

> A catalyst increases the rate of both the forward and reverse reactions without affecting the equilibrium position. Therefore, the graph will show that the system reaches equilibrium faster than without the catalyst. The shape of the curves representing the concentrations of A_2, B_2, and AB will remain the same, but they will reach their equilibrium values sooner.

While this may suggest that the chatbots have some understanding of the information provided in diagrams such as graphs, it does not confirm this ability, as they could have inferred this understanding from the wording of the questions.

In question 8.4, shown in **Figure 6**, based on galvanic cells in electrochemical reactions, the chatbots were provided with a galvanic cell diagram with copper and silver strips as electrodes. The question informed them that the salt bridge contained potassium nitrate $KNO_{3(aq)}$. The question required them to identify the ion in the salt bridge that would migrate into the silver ion solution.

All three chatbots failed to generate the correct response. For example, ChatGPT Plus generated the following response:

> NO_3^- (aq) moves to maintain charge balance as Ag^+ ions are reduced.

The silver ions would be reduced in the silver electrode, which was correctly stated. ChatGPT Plus could not understand that when silver ions are removed from the solution due to the reduction reaction, there would be a negative charge build-up, requiring K^+ ions from the salt bridge to maintain electrical neutrality.

Although the chatbots could accurately answer some questions (e.g., question 6.6), it was evident that they were not yet adept at providing accurate solutions for questions in higher order thinking skills such as analyzing. Claude 3.5 Sonnet was the only chatbot that obtained more than half of the marks at the analyzing level of Bloom's taxonomy, with the other chatbots getting less than half, as shown in **Table 3**.

Evaluating and Creating

Based on chemical equilibrium, one sub-question in question 6.5 was classified under evaluating, and the other sub-question under creating. These questions were 6.5.1 (creating) and 6.5.2 (evaluating). These questions are shown in **Figure 9**. Question 6.5 stated that for the reaction depicted in **Figure 9**, the temperature was decreased from 150 °C to 100 °C at time t = 130 s. Question 6.3.1 required the chatbots to draw a potential energy diagram for the reaction. All three chatbots could not draw the required graph, which provided further evidence that they currently do not have this ability. ChatGPT and Claude offered descriptive explanations, but Gemini failed completely. The responses from the chatbots were, as follows:

> I would draw a diagram showing reactants on the left, products on the right, with an activation energy hump in between. The products would be slightly higher in energy than the reactants, indicating an endothermic reaction (Claude).

> The diagram should show reactants, activation energy, and products with appropriate labels (ChatGPT Plus).

> To draw a potential energy diagram, we need information about the activation energy of the forward and reverse reactions. The information is not provided in the question (Gemini).

The response generated by Claude was the only accurate response but was considered partially correct as it could not draw the diagram. The other responses were inaccurate. Question 6.5.2 required the chatbots to evaluate whether the equilibrium constant, K_c at 100 °C, would be greater than, less than or equal to the K_c at 150 °C and to explain the answer (**Figure 8**). The responses generated by the chatbots were, as follows:

> K_c at 100°C will be LESS THAN K_c at 150°C. Explanation: The forward reaction is endothermic (products higher energy than reactants). Decreasing temperature favors the exothermic reaction (reverse in this case), shifting the equilibrium towards reactants and decreasing K_c, (Claude).

> Note: To determine the effect of temperature on K_c, we need to know whether the reaction is exothermic or endothermic. Please indicate whether the reaction is exothermic or endothermic. Once we know this, we can apply Le Chatelier's principle to determine the effect of temperature on K_c (Gemini).

> Greater than if the reaction is exothermic, as decreasing temperature favors the formation of products (ChatGPT).

As seen from the three responses above, only Claude could answer the question accurately. Evaluating and creating are higher-order thinking skills. The responses from the three chatbots seem to suggest that Claude was a more versatile application that had some ability to respond accurately to chemistry questions at the high levels of Bloom's taxonomy of educational objectives. Gemini and ChatGPT Plus could not perform at the same level as Claude.

DISCUSSION

The outcomes of this study are consistent with the results from other scientific research areas. The study has provided empirical evidence that while the investigated chatbots could satisfactorily answer questions at the remembering level in chemistry, they struggled to answer higher-order questions. In this study, Claude 3.5 Sonnet had an overall mark of 65%, followed by Gemini with a mark of 51%, and ChatGPT Plus could only obtain 47%. Antaki et al. (2023) investigated ChatGPT Legacy and ChatGPT Plus's performance in answering exam questions in ophthalmology. Their findings were that the ChatGPT Plus performed better than the legacy model. ChatGPT Plus generated 59.4% correct responses in the basic and clinical science course, while the legacy model obtained 55.8% correct responses. In the section on ophtho-questions, the legacy model achieved 49.2%, while ChatGPT Plus obtained 55.8% correct responses. Both models of ChatGPT performed better in questions requiring low-order thinking and had lower performance in questions requiring higher-order thinking, similar to what was observed in this study. Similarly, Kung et al. (2023) evaluated ChatGPT's performance in a USA medical examination and found it to perform at 60% accuracy with good comprehension and valid reasoning on clinical matters.

The finding that ChatGPT defines chemical concepts accurately aligns with the research by Leite (2024), who found that chatbots such as ChatGPT provide comprehensive explanations of chemical concepts. This is likely due to their extensive training on publicly available data. Our study further revealed difficulties the investigated chatbots encountered for chemical problems requiring comprehension and application. Some studies have also demonstrated that ChatGPT cannot respond satisfactorily to chemical problems requiring understanding and application (Daher et al., 2023; Fergus et al., 2023; Yik & Dood, 2024). This finding implies that learners dependent on these chatbots should use the chatbots sparingly or alternatively use them together with traditional sources of information and must always check the accuracy of responses provided by the chatbots. While Yik and Dood (2024) noted that only a quarter of explanations of reaction mechanisms by ChatGPT were accurate and convincingly achieved, they also observed that the responses of inaccurate explanations could be improved through enhanced prompt engineering. This suggests that learners using ChatGPT and other LLMs should have prompt engineering skills and can use these skills to improve the accuracy of the responses generated by the chatbots.

In the section on organic chemistry, the language models were found to have considerable difficulties in naming organic molecules, given structural formulae or drawing structural formulae, and analyzing these structures. This finding aligns with Hallal et al. (2023), whose study revealed that ChatGPT had problems with IUPAC naming organic compounds and converting between various kinds of organic structural formulae. Therefore, learners cannot depend on responses generated by the chatbots when studying organic chemistry problems.

Limitations of the Study

While the study provided empirical evidence of the current performance of the three chatbots in their ability to generate accurate responses to the grade 12 final examination chemistry paper in South Africa, it had some limitations worth noting. Firstly, only one examination paper was used as the research instrument to conduct an in-depth qualitative data analysis. Providing the three chatbots with a different chemistry paper from the same examination board may result in different results. Secondly, the results only apply to the chatbots' performance during the research period when the study was done. The chatbots are undergoing continuous improvement, and these results may not apply in the future. No attempt was made to enhance prompt engineering, which could have resulted in different results. Learners using the chatbots to revise their work may use prompt engineering to obtain more accurate results.

CONCLUSIONS AND IMPLICATIONS

The findings of this study showed that the three chatbots investigated could satisfactorily generate accurate responses at the lower-order thinking level of remembering in Bloom's taxonomy. Beyond remembering, the chatbots could not generate accurate responses to the examination questions in most of the questions. This trend compares favorably with recent studies that showed that chatbots obtain slightly above 50% in most exams in science-related fields. These results suggest that grade 12 learners cannot rely on these chatbots to revise more challenging questions. It is also possible that reliance on these chatbots can transmit misconceptions, which can become hard to overcome. A more practical guideline is that when candidates prepare for examinations, they can only use these chatbots in conjunction with traditional sources of information such as textbooks, and they would need to verify any answers generated by these chatbots. The chatbots are useful when candidates wish to verify facts based on low-order thinking skills such as basic facts or definitions. Learners should develop critical thinking skills to benefit from the chatbots as they need to evaluate their responses. Meaningful learning is likely to occur if learners use critical thinking skills and avoid regarding the responses from the chatbots as being absolutely accurate.

Funding: This study was partially supported by the University of Venda for open access.

Ethical statement: The author stated that no human or animal subjects were involved in this study. The study did not require an ethical clearance.

Declaration of interest: No conflict of interest is declared by the author.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the author.

REFERENCES

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education, 3*(1), Article 60. https://doi.org/10.1007/s44217-024-00138-2

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. Addison Wesley Longman, Inc.

Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science, 3*(4), Article 100324. https://doi.org/10.1016/j.xops.2023.100324

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). ChatGPT: Applications, opportunities, and threats. In *Proceedings of the 2023 Systems and Information Engineering Design Symposium*. https://doi.org/10.1109/sieds58326.2023.10137850

Bhattacherjee, A. (2012). Social science research: Principles, methods, and practices. *University of South Florida*. https://repository.out.ac.tz/504/1/Social_Science_Research-_Principles_Methods_and_Practices.pdf

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain*. David McKay Co. In: Inc.

Bonaci, C. G., Mustata, R. V., & Ienciu, A. (2013). Revisiting Bloom's taxonomy of educational objectives. *The Macrotheme Review, 2*(2), 1-9.

Carlà, M. M., Gambini, G., Baldascino, A., Boselli, F., Giannuzzi, F., Margollicci, F., & Rizzo, S. (2024). Large language models as assistance for glaucoma surgical cases: A ChatGPT vs. Google Gemini comparison. *Graefe's Archive for Clinical and*

Experimental Ophthalmology, 262, 2945-2959. https://doi.org/10.1007/s00417-024-06470-5

Chiu, T. K., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with artificial intelligence (AI) based chatbot. *Interactive Learning Environments, 32*(7), 3240-3256. https://doi.org/10.1080/10494820.2023.2172044

Daher, W., Diab, H., & Rayan, A. (2023). Artificial intelligence generative tools and conceptual knowledge in problem solving in chemistry. *Information, 14*(7), Article 409. https://doi.org/10.3390/info14070409

DBE. (2023a). 2023 diagnostic report book 1. *Department of Basic Education*. https://www.education.gov.za/Portals/0/Documents/Reports/Diagnostic%20Reports%202022/Diagnostic%20Report%202023%20Book%201.......pdf?ver=2024-02-27-144522-073

DBE. (2023b). Physical sciences paper 2. *Department of Basic Education*. https://www.education.gov.za/Curriculum/NationalSeniorCertificate(NSC)Examinations/2023NSCNovemberpastpapers.aspx

Doğru, S. (2023). ChatGPT-science education and instruction reshapes management. *Online Science Education Journal, 8*(1), 12-21.

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education, 100*(4), 1672-1675. https://doi.org/10.1021/acs.jchemed.3c00087

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51*(12), 2629-2633. https://doi.org/10.1007/s10439-023-03272-4

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), Article 692. https://doi.org/10.3390/educsci13070692

Hallal, K., Hamdan, R., & Tlais, S. (2023). Exploring the potential of AI-Chatbots in organic chemistry: An assessment of ChatGPT and Bard. *Computers and Education: Artificial Intelligence, 5*, Article 100170. https://doi.org/10.1016/j.caeai.2023.100170

Hutton-Prager, B. (2018). Utilizing a differentiation framework, Piagetian theories and Bloom's Taxonomy to foster experiential learning activities in chemical engineering. In B. I. Morsi, & O. M. Basha (Eds.), *Laboratory unit operations and experimental methods in chemical engineering*. IntechOpen. https://doi.org/10.5772/intechopen.75646

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*(12), 15873-15892. https://doi.org/10.1007/s10639-023-11834-1

Jere, S., Bessong, R., Mpeta, M., & Litshani, N. F. (2024). Pre-service physical sciences teachers' views on integrating ChatGPT into teaching: A case study. *Research Square*. https://doi.org/10.21203/rs.3.rs-4783356/v1

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., & Maningo, J. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health, 2*(2), Article e0000198. https://doi.org/10.1371/journal.pdig.0000198

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, Article 100213. https://doi.org/10.1016/j.caeai.2024.100213

Leite, B. S. (2024). Generative artificial intelligence in chemistry teaching: ChatGPT, Gemini, and Copilot's content responses. *Journal of Applied Learning and Teaching, 7*(2). https://doi.org/10.37074/jalt.2024.7.2.13

Makrygiannakis, M. A., Giannakopoulos, K., & Kaklamanos, E. G. (2024). Evidence-based potential of generative artificial intelligence large language models in orthodontics: A comparative study of ChatGPT, Google Bard, and Microsoft Bing. *European Journal of Orthodontics, 2024*, cjae017. https://doi.org/10.1093/ejo/cjae017

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education, 28*(6), 1549-1565. https://doi.org/10.1002/cae.22326

Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan, S., & Liu, Q. (2023). Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pacific*

Education Review, 25, 1379-1389. https://doi.org/10.1007/s12564-023-09913-6

Wilson, L. O. (2016). *Anderson and Krathwohl Bloom's taxonomy revised understanding the new version of Bloom's taxonomy*. https://www.quincycollege.edu/wp-content/uploads/Anderson-and-Krathwohl_Revised-Blooms-Taxonomy.pdf

Yik, B. J., & Dood, A. J. (2024). ChatGPT convincingly explains organic chemistry reaction mechanisms slightly inaccurately with high levels of explanation sophistication. *Journal of Chemical Education, 101*(5), 1836-1846. https://doi.org/10.1021/acs.jchemed.4c00235

APPENDIX A: QUESTIONS (Source: Department of Basic Education [DBE], 2023b)

QUESTION 1: MULTIPLE-CHOICE QUESTIONS

Various options are provided as possible answers to the following questions. Each question has only ONE correct answer. Choose the answer and write only the letter (A–D) next to the question numbers (1.1 to 1.10) in the ANSWER BOOK, e.g. 1.11 E.

1.1 Which ONE of the following represents a straight chain SATURATED hydrocarbon?

A C_5H_8

B C_5H_{10}

C C_6H_{12}

D C_6H_{14} (2)

1.2 Which ONE of the following is a SECONDARY alcohol?

A $C(CH_3)_3OH$

B $CH_3(CH_2)_3OH$

C $CH_3(CH_2)_2CHO$

D $CH_3CH_2CH(OH)CH_3$ (2)

1.3 Which ONE of the following is a HYDROLYSIS reaction?

A $CH_3CH_2Br + H_2O \rightarrow CH_3CH_2OH + HBr$

B $CH_3CH_2OH + HBr \rightarrow CH_3CH_2Br + H_2O$

C $CH_2CH_2 + H_2O \rightarrow CH_3CH_2OH$

D $CH_2CH_2 + H_2 \rightarrow CH_3CH_3$ (2)

1.4 Hydrochloric acid reacts with EXCESS zinc:

$$Zn(s) + 2HC\ell(aq) \rightarrow ZnC\ell_2(aq) + H_2(g)$$

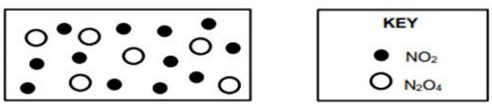
Different reaction conditions are shown in the diagrams below. The mass of zinc used is the same in both test tubes.

How will the INITIAL rate of reaction and FINAL VOLUME of $H_2(g)$ produced in test tube **Y** compare with that in test tube **X**?

	INITIAL RATE OF REACTION IN Y	FINAL VOLUME OF $H_2(G)$ IN Y
A	Higher	Equal
B	Lower	More
C	Lower	Equal
D	Higher	More

(2)

1.5 The diagram below represents a mixture of $NO_2(g)$ and $N_2O_4(g)$ molecules at equilibrium in a 1 dm^3 container at T °C.



The balanced equation for this reaction is:

$$2NO_2(g) \rightleftharpoons N_2O_4(g)$$

Which ONE of the following is TRUE for the value of the equilibrium constant, K_c, for the reaction at T °C?

A $K_c = 24$

B $K_c > 1$

C $K_c = 1$

D $0 < K_c < 1$ (2)

1.6 A reaction is at equilibrium in a closed container according to the following balanced equation:

$$4CuO(s) \rightleftharpoons 2Cu_2O(s) + O_2(g)$$

The volume of the container is now increased while the temperature remains constant. A new equilibrium is reached.

Which ONE of the following combinations is CORRECT for the new equilibrium?

	CONCENTRATION OF O_2	NUMBER of MOLES OF O_2	EQUILIBRIUM CONSTANT (K_c)
A	Decreases	Remains the same	Increases
B	Remains the same	Decreases	Remains the same
C	Remains the same	Increases	Remains the same
D	Decreases	Increases	Remains the same

(2)

1.7 Nitric acid, $HNO_3(aq)$, and ethanoic acid, $CH_3COOH(aq)$, of equal volumes and concentrations are compared.

Consider the following statements regarding these solutions:

(i) They have different pH values.
(ii) Both have the same electrical conductivity.
(iii) Both solutions require the same number of moles of KOH(aq) for complete neutralisation.

Which of the above statement(s) is/are TRUE?

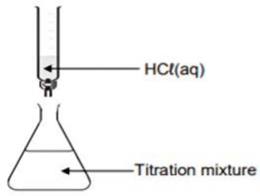
A (i) only

B (i) and (ii) only

C (i) and (iii) only

D (ii) and (iii) only (2)

1.8 The apparatus in the diagram below is used for the titration between HCℓ(aq) and KOH(aq).



In a titration, the learner accidentally exceeds the endpoint. Which ONE of the following will be TRUE for the titration mixture?

A $[H^+] > [OH^-]$ and pH < 7

B $[H^+] < [OH^-]$ and pH < 7

C $[H^+] < [OH^-]$ and pH > 7

D $[H^+] > [OH^-]$ and pH > 7 (2)

1.9 The following hypothetical standard reduction potentials relate to a galvanic cell:

$$X^{2+}(aq) + 2e^- \rightarrow X(s) \qquad E^\theta = +0{,}10 \text{ V}$$
$$Y^+(aq) + e^- \rightarrow Y(s) \qquad E^\theta = -0{,}10 \text{ V}$$

Consider the following statements for this galvanic cell:

(i) The emf of the cell is 0,20 V under standard conditions.
(ii) Electrode **Y** is the anode.
(iii) **X** is oxidised.

Which of the above statement(s) is/are TRUE for this galvanic cell?

A (i) only

B (i) and (ii) only

C (i) and (iii) only

D (ii) and (iii) only (2)

1.10 Which ONE of the half-reactions below will be the MAIN reaction at the ANODE during the electrolysis of CONCENTRATED $CuCℓ_2(aq)$?

A $Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$

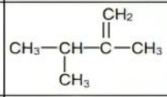
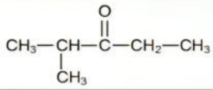
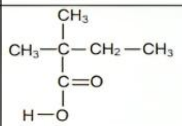
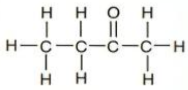
B $2H_2O(ℓ) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$

C $2H_2O(ℓ) \rightarrow O_2(g) + 4H^+(aq) + 4e^-$

D $2Cℓ^-(aq) \rightarrow Cℓ_2(g) + 2e^-$ (2)
[20]

QUESTION 2 (Start on a new page.)

The letters **A** to **H** in the table below represent eight organic compounds.

A	Heptanoic acid	**B**	$CH_3(CH_2)_3COOCH_3$
C	4-ethyl-3,3-difluorohexane	**D**	Hexanoic acid
E		**F**	
G		**H**	

2.1 Define the term *organic compound*. (1)

2.2 Write down the IUPAC name of compound:

2.2.1 **E** (2)

2.2.2 **H** (2)

2.3 Write down the:

2.3.1 STRUCTURAL formula of compound **B** (2)

2.3.2 STRUCTURAL formula of compound **C** (3)

2.3.3 General formula of the homologous series to which compound **E** belongs (1)

2.3.4 STRUCTURAL formula of the FUNCTIONAL group of compound **F** (1)

2.3.5 IUPAC name of the alcohol needed to produce compound **B** (2)

2.4 Write down the letter(s) of the compound(s) that:

2.4.1 Is a FUNCTIONAL isomer of compound **G** (1)

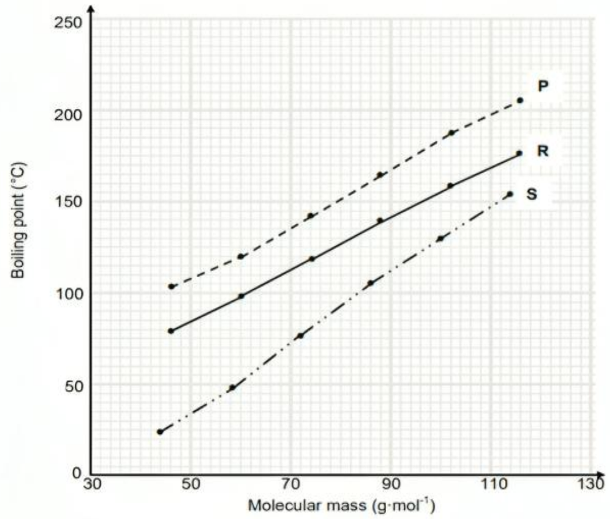
2.4.2 Are CHAIN isomers of each other (1)

[16]

QUESTION 3 (Start on a new page.)

The relationship between boiling point and the molecular mass of aldehydes, carboxylic acids and primary alcohols is investigated. Curves **P**, **R** and **S** are obtained. All compounds used are straight chain molecules.

GRAPH OF BOILING POINT VERSUS MOLECULAR MASS



3.1	Define the term *boiling point*.	(2)
3.2	Write down the conclusion that can be made for curve **P**.	(2)
3.3	Explain the answer to QUESTION 3.2 in terms of the structures of the compounds.	(2)
3.4	Curve **R** represents the alcohols.	
3.4.1	Which homologous series is represented by curve **S**?	(1)
3.4.2	Explain the answer to QUESTION 3.4.1 by referring to the strength of intermolecular forces.	(2)
3.5	For curve **R**, write down the:	
3.5.1	Molecular mass of the compound with a boiling point of 97 °C	(1)
3.5.2	IUPAC name of the compound in QUESTION 3.5.1	(2)
3.6	Two compounds, **A** and **B**, used in this investigation have a molecular mass of 74 $g \cdot mol^{-1}$. **A** has a boiling point of 118 °C and **B** a boiling point of 142 °C. Explain the difference in these boiling points by referring to the structures of these compounds.	(3) **[15]**

QUESTION 4 (Start on a new page.)

4.1 Consider the cracking reaction below.

$$C_{16}H_{34} \longrightarrow C_6H_{14} + C_6H_x + 2C_yH_z$$

4.1.1 Define *cracking*. (2)

4.1.2 Write down the values represented by **x**, **y** and **z** in the equation above. (3)

Compound C_6H_{14} undergoes complete combustion.

4.1.3 Using MOLECULAR FORMULAE, write down the balanced equation for this reaction. (3)

4.2 Consider the equations for reactions **I** to **III** below.

A and **B** represent organic compounds that are POSITIONAL ISOMERS. **X** is an inorganic product.

I	$CH_3CH_2CHCHCH_3 + HC\ell \rightarrow$ **A** + **B**
II	**A** $\xrightarrow[\Delta]{H_2O}$ $CH_3CH_2CH_2CH(OH)CH_3$ + **X**
III	$CH_3CH_2CH_2CH(OH)CH_3 \longrightarrow CH_3CH_2CHCHCH_3 + H_2O$

Write down the:

4.2.1 Definition of *positional isomers* (2)

4.2.2 Type of reaction represented by reaction **I** (1)

4.2.3 STRUCTURAL formula of compound **B** (3)

4.2.4 Formula of **X** (1)

4.2.5 Inorganic reagent for reaction **III** (1)

Compound **A** can be converted directly to the organic product of reaction **III**.

4.2.6 Besides heat, write down the reaction condition needed for this conversion. (1)

4.2.7 Write down TWO terms that describe this type of reaction. (2)
[19]

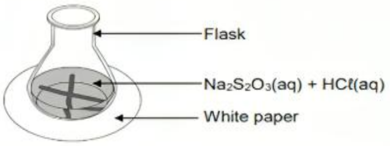
QUESTION 5 (Start on a new page.)

The reaction between EXCESS dilute hydrochloric acid and sodium thiosulphate is used to investigate factors that influence reaction rate.

$$Na_2S_2O_3(aq) + 2HC\ell(aq) \rightarrow 2NaC\ell(aq) + S(s) + H_2O(\ell) + SO_2(g)$$

The concentration of $HC\ell(aq)$ used is 1 mol·dm^{-3}. The same volume of $HC\ell(aq)$ is used in each run.

The time taken for the cross on the paper under the flask to become invisible is measured.



The table below summarises the reaction conditions and results of the experiment.

RUN	VOLUME $Na_2S_2O_3(aq)$ (cm^3)	VOLUME $H_2O(\ell)$ ADDED (cm^3)	CONCENTRATION $Na_2S_2O_3(aq)$ (mol·dm^{-3})	TIME (s)
1	50	0	0,13	20,4
2	40	10	0,10	26,7
3	30	20	**P**	33,3

5.1 Define *reaction rate*. (2)

5.2 Write down the independent variable for this investigation. (1)

5.3 Calculate the value of **P** in the table. (3)

5.4 When 0,21 g of sulphur has formed in Run **1**, the cross becomes invisible.

Calculate the average reaction rate with respect to sodium thiosulphate, $Na_2S_2O_3(aq)$, in g·s^{-1}. (5)

Another investigation is performed at different temperatures.

5.5 Sketch the Maxwell-Boltzmann distribution curve for the reaction at 20 °C. Label this curve as **A**. On the same set of axis, draw the curve that will be obtained at 35 °C and label it as **B**. (4)

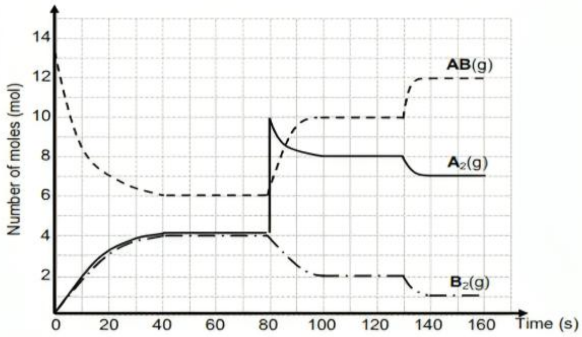
5.6 Explain the effect of temperature on reaction rate in terms of the collision theory. (4)
[19]

QUESTION 6 (Start on a new page.)

Consider the following hypothetical reaction reaching equilibrium in a 4 dm^3 closed container at 150 °C.

$$2AB(g) \rightleftharpoons A_2(g) + B_2(g)$$

The graph below shows the changes in the amounts of reactants and products over time.



6.1	Write down the meaning of the term *reversible reaction*.		(1)
6.2	State Le Chatelier's principle.		(2)
6.3	A change was made to the equilibrium mixture at t = 80 s.		
	6.3.1	Write down the change made at t = 80 s.	(1)
	6.3.2	Use Le Chatelier's principle to explain how the system reacts to this change.	(2)
6.4	Calculate the equilibrium constant, K_c, at t = 120 s.		(4)
6.5	At t = 130 s the temperature of the system is decreased to 100 °C.		
	6.5.1	Draw a potential energy diagram for this reaction.	(3)
	6.5.2	Will the equilibrium constant, K_c, at 100 °C be GREATER THAN, LESS THAN or EQUAL TO the K_c at 150 °C? Explain the answer.	(3)
6.6	The initial reaction now takes place in the presence of a catalyst at 150 °C.		
	Describe the changes that will be observed on the graph between t = 0 s and t = 60 s.		(3)
			[19]

QUESTION 7 (Start on a new page.)

To identify metal **M** in an unknown metal carbonate, $\mathbf{M}CO_3$, the following procedure is carried out:

Step 1: 0,198 g of IMPURE $\mathbf{M}CO_3$ is reacted with 25 cm^3 of 0,4 $mol{\cdot}dm^{-3}$ nitric acid, $HNO_3(aq)$.

Step 2: The EXCESS $HNO_3(aq)$ is then neutralised with 20 cm^3 of 0,15 $mol{\cdot}dm^{-3}$ barium hydroxide, $Ba(OH)_2(aq)$.

Assume that the volumes are additive.

The following reactions take place:

$$2HNO_3(aq) + \mathbf{M}CO_3(s) \rightarrow \mathbf{M}(NO_3)_2(aq) + CO_2(g) + H_2O(\ell)$$

$$2HNO_3(aq) + Ba(OH)_2(aq) \rightarrow Ba(NO_3)_2(aq) + 2H_2O(\ell)$$

7.1 Define the term *strong base*. (2)

7.2 Calculate the:

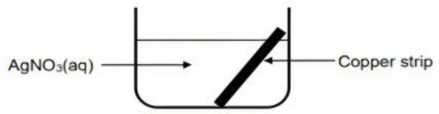
7.2.1 Number of moles of $Ba(OH)_2(aq)$ that reacted with the excess $HNO_3(aq)$ (3)

7.2.2 pH of the solution after Step 1 (5)

7.3 The percentage purity of the $\mathbf{M}CO_3(s)$ in the sample is 85%. Identify metal **M**. (8)
[18]

QUESTION 8 (Start on a new page.)

A cleaned pure copper strip, Cu(s), is placed in a beaker containing a colourless silver nitrate solution, $AgNO_3$(aq), at 25 °C, as shown below.



After a while, it is observed that the solution in the beaker becomes blue.

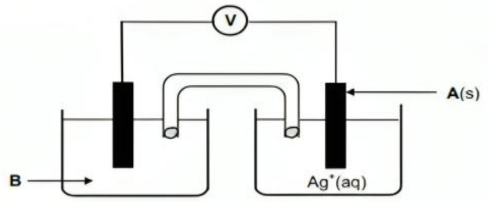
8.1 Write down:

8.1.1 ONE other OBSERVABLE change, besides the solution turning blue (1)

8.1.2 The NAME or FORMULA of the oxidising agent (1)

8.2 Explain the answer to QUESTION 8.1.1 by referring to the relative strengths of the oxidising agents or reducing agents. (3)

A galvanic cell is now set up using Cu and Ag strips as electrodes. A simplified diagram of the cell is shown below.



8.3 Write down the:

8.3.1 NAME or FORMULA of electrode **A** (1)

8.3.2 NAME or FORMULA of solution **B** (1)

8.3.3 Overall (net) balanced equation for the cell reaction (3)

8.4 The salt bridge contains potassium nitrate, KNO_3(aq).

Write down the FORMULA of the ion in the salt bridge that will move into the silver ion solution. Choose from K^+(aq) or NO_3^-(aq).

Give a reason for the answer. (2)
[12]

https://www.ejmste.com