جامعة، المجلد 25 العدد 2 (2025): 123-162

An In-Depth Study of the Reliability of ChatGPT and Gemini in Addressing Complex Chemistry Questions in Instrumental Analysis

Saleh Abu-Lafi[1], and Anwar Rayan[2]

Abstract

Artificial intelligence (AI) has gained significant value in education and problem-solving across various fields. Advanced generative AI systems like ChatGPT (versions GPT-4 and GPT-4o) and Gemini AI provide notable advantages in tackling challenges, especially in chemistry. This research investigates how ChatGPT and Gemini AI help understand and solve chemistry problems, focusing on instrumental analysis for undergraduate pharmacy students. The study evaluates the problem-solving skills of both AI systems using a set of 120 multiple-choice questions (MCQs) randomly selected from exam-style queries. It also observes instances where ChatGPT, when asked the same questions twice, gave inconsistent answers - some correct and some incorrect. Gemini AI scored an overall 87.5% (105 out of 120 questions), while GPT-4 scored 84.2% (101 out of 120), and GPT-4o scored 85.3% (103 out of 120). The percentage of correct answers notably declined for calculation-based questions across all AI systems. Overall, GPT-4 and GPT-4o performed similarly in total correct answers, with the added benefit of providing logical, step-by-step explanations for their solutions. This feature is particularly useful for pharmacy students, as it can assist with calculations, improve understanding of concepts, and help overcome challenges in chemistry and instrumental analysis in pharmaceutical education.

[1] Faculty of Pharmacy, Al-Quds University, Abu-Dies, Palestine; sabulafi@staff.alquds.edu

[2] Science and Technology Department, Faculty of Science, Al-Qasemi Academic College, Baka EL-Garbiah 30100, Israel; a_rayan@qsm.ac.il,
Correspondence: a_rayan@qsm.ac.il

Therefore, AI can support chemistry and pharmacy students in performing calculations, understanding concepts, and resolving problems. However, issues such as response inconsistencies and domain-specific fine-tuning are recommended to be further addressed to maximize these systems' potential.

Keywords: artificial intelligence, GPT-4; GPT-4o, Gemini, analytical chemistry, instrumental analysis, problem solving, undergraduate pharmacy students, multiple-choice questions.

1. Introduction

Recent advancements in natural language processing (NLP) within artificial intelligence (AI) have revolutionized interactions between humans and machines [1, 2]. Two prominent AI models, ChatGPT [3] and Gemini [4], have gained considerable recognition for their capability to generate text that closely resembles human speech and effectively manages complex queries [5]. In the context of pharmacy education, students frequently encounter multiple-choice questions (MCQs) focusing on instrumental analysis, which demands a profound grasp of chemical principles and analytical techniques [6, 7]. Addressing these intricate chemistry queries is crucial in pharmacy education to ensure that future pharmacists acquire essential knowledge and skills.

Instrumental analysis, integral to chemistry, encompasses techniques such as chromatography and spectroscopy, as well as their tandem (hyphenated) combinations, which play a critical role in separating, quantifying, and identifying active pharmaceutical compounds (APIs) [8]. Chromatographic methods commonly include high-performance liquid chromatography (HPLC), ultra-performance liquid chromatography (UPLC), and hyphenated methods like liquid chromatography-mass spectrometry (LC-MS) and liquid chromatography with tandem mass spectrometry (LC-MS-MS). Additionally, significant spectroscopic detectors used in conjunction with HPLC include the photo diode array (PDA) detector and the evaporative light scattering detector (ELSD), among others. Furthermore, gas chromatography (GC) and its hyphenated counterpart gas chromatography-mass spectrometry (GC-MS) are extensively employed, utilizing detectors such as flame ionization detectors (FIDs), electron capture detectors (ECDs), and thermal conductivity detectors (TCDs). In terms of spectroscopic techniques, molecular and atomic spectroscopy methods are pivotal; these include ultraviolet-visible (UV-Vis) absorption spectroscopy, infrared (IR) spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, atomic absorption spectroscopy (AAS), and atomic emission spectroscopy (AES) [8]. These methods are crucial to ensuring the quality, safety, and efficacy of APIs in pharmaceutical products [9].

The complexity and diversity of these techniques highlight the cognitive challenges faced by pharmacy students, underscoring the potential role of AI in facilitating understanding and practice. Recently, the integration of AI into pharmaceutical analysis has become prevalent [10]. A recent review by Rafael Cardoso Rial emphasized the significance of AI in analytical chemistry, discussing its advancements, challenges, and future prospects [7]. AI in analytical chemistry revolutionizes data interpretation and optimizes processes for faster and more accurate compound separation, identification, and quantitation [7, 11]. It excels at managing large volumes of data and automating repetitive tasks, offering innovative approaches to problem analysis that surpass conventional methods [12, 13]. Yet, despite these advancements, challenges persist in the realms of model accuracy and effectiveness [7]. The success of AIs depends on having good data and finding new insights, and collaborative efforts are crucial to harnessing its capabilities responsibly and ensuring alignment with scientific integrity and sustainable progress. The fields of analytical chemistry and life sciences are ahead in using AI, possibly beyond peak expectations, which is evident in the rapid growth of publications, although this is uneven across subfields [13]. However, despite the rapid integration of AI into pharmaceutical analysis, little is known about its reliability in supporting students' learning of instrumental analysis through exam-style assessments.

This study aims to explore the reliability of advanced generative AI systems, particularly ChatGPT (versions GPT-4 and GPT-4o) and Gemini, in addressing complex multiple-choice questions (MCQs) in instrumental analysis designed for undergraduate pharmacy students. The focus is not only on their accuracy in providing correct answers but also on their potential to enhance students' conceptual understanding and support the learning process. Previous work has highlighted the importance of domain-specific fine-tuning and the limitations of large language models when applied to specialized fields such as analytical chemistry and instrumental analysis [14, 15]. By examining these AI tools in an educational context, our study contributes to the broader discussion on their reliability, trustworthiness, and implications for advancing chemistry education and research [16]. To our knowledge, this is among the first empirical investigations to systematically evaluate AI models using authentic, exam-style multiple-choice questions in pharmaceutical instrumental analysis. This novelty positions the study to provide unique insights into the educational value and reliability of generative AI tools in a highly specialized scientific domain.

2. Materials and Methods

2.1 Research content and design

This research project was carried out at Al-Quds University, Faculty of Pharmacy, in May-June 2024. It was conducted as part of a senior-year Pharmaceutical Instrumental Analysis course, which is designed to impart fundamental knowledge in pharmaceutical instrumental analysis using advanced instrumentation. The initial focus is on liquid chromatographic techniques, such as HPLC-PDA and hyphenated chromatography like LC-MS and LC-MS/MS, which are utilized for the separation, identification, and quantification of neutral, acidic, and basic drugs. Gas chromatography and hyphenated techniques such as GC-MS are also covered extensively for their applications in drug analysis. The second part of the course explores molecular and atomic spectroscopic methods. Topics include atomic absorption spectroscopy, atomic emission spectroscopy, NMR, IR and UV-Vis absorption spectroscopy, and luminescence techniques, all with applications vital to drug analysis. The course helps students to establish a strong foundation in the fundamental principles of pharmaceutical analysis and develop a comprehensive understanding of its essential components.

2.2 Construction of the MCQ Bank

The 120-MCQ bank was developed using information from university exam question pools on instrumental analysis. Initially, the questions were categorized by subject and scrutinized to ensure they aligned with the subject matter and were suitably challenging. Each question presented four options, one of which was the correct answer. We conducted a thorough evaluation of the MCQs to ensure the clarity and accuracy of the sole correct response. The language used was simple and comprehensible. Additionally, our team proofread the questions to rectify errors, typos, confusing statements, and inconsistencies.

2.3 Techniques addressed by the MCQs and data collection

Out of the randomly chosen 120 exam-style questions, 49 (40.8%) were related to chromatography, 57 (47.5%) to spectroscopy, and 14 (11.7%) to hyphenated systems. The answers of each participant were compared to an answer key; correct responses were given a score of 1 and incorrect responses a score of 0. The questions were introduced to GPT-4, GPT-4o and Gemini to assess their response. The research highlighted instances when both versions of ChatGPT, when asked the same question twice, gave a correct and an incorrect response or two different incorrect responses; this is referred to here as an “inconsistency.”

2.4 Statistical analysis

The data were carefully reviewed, and the findings were documented and analyzed. To assess differences in efficiency between the AI-based tools, we used a one-way ANOVA test, with statistical significance set at $p < 0.05$.

3. Results

This study examined the performance of GPT-4, GPT-4o, and Gemini responses to 120 multiple-choice questions (MCQs) selected randomly from exam-style questions used in a course tailored for senior-year undergraduate students in the Faculty of Pharmacy, which focuses on essential topics in pharmaceutical instrumental analysis. The curriculum encompasses foundational principles of instrumental analysis, including chromatographic and spectroscopic methods for isolating and quantifying active ingredients in drugs. Overall, there were a total of 49 MCQs (40.8%) related to chromatography, 57 (47.5%) focused on spectroscopy, and 14 (11.7%) addressing hyphenated systems (Figure 1). Each AI's answers were compared with a predetermined answer key to assess their accuracy and were evaluated and categorized as correct or incorrect. However, the study observed occurrences where ChatGPT provided inconsistent answers upon being asked a question twice. As depicted in Figure 2, the Gemini AI achieved an overall score of 87.5% (105 out of 120 questions) across all the MCQs; GPT-4 scored 84.2% (101 out of 120) and GPT-4o 85.3% (103 out of 120).

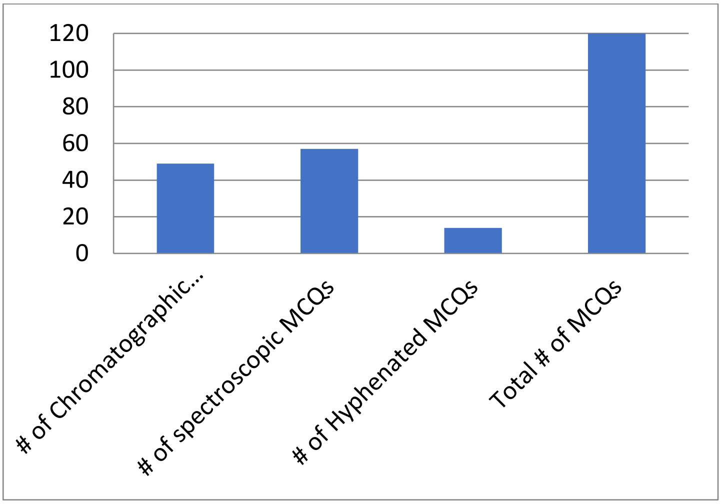


Figure 1. Distribution of the MCQs.

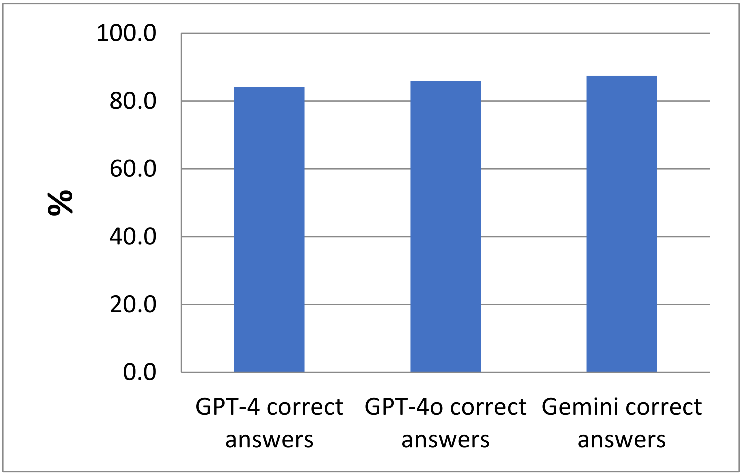


Figure 2. Percentage of correct answers of GPT-4, GPT-4o and Gemini.

Inconsistencies were observed in GPT-4o at a rate of 7.5% (9 out of 120 questions), while GPT-4 exhibited a lower inconsistency rate of 2.5% (3 out of 120 questions) (Figure 3). Tables 1-3 show instances of inconsistencies in the MCQ responses provided by GPT-4 and GPT-4o, respectively.

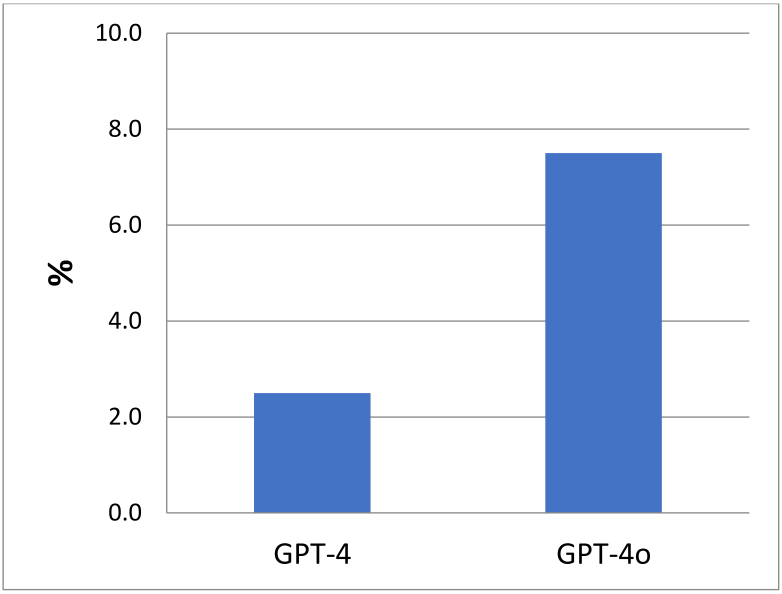


Figure 3. Inconsistency rate of GPT-4 and GPT-4o.

Table 1. An inconsistency in answering a MCQ on system suitability, using GPT-4

ChatGPT 4

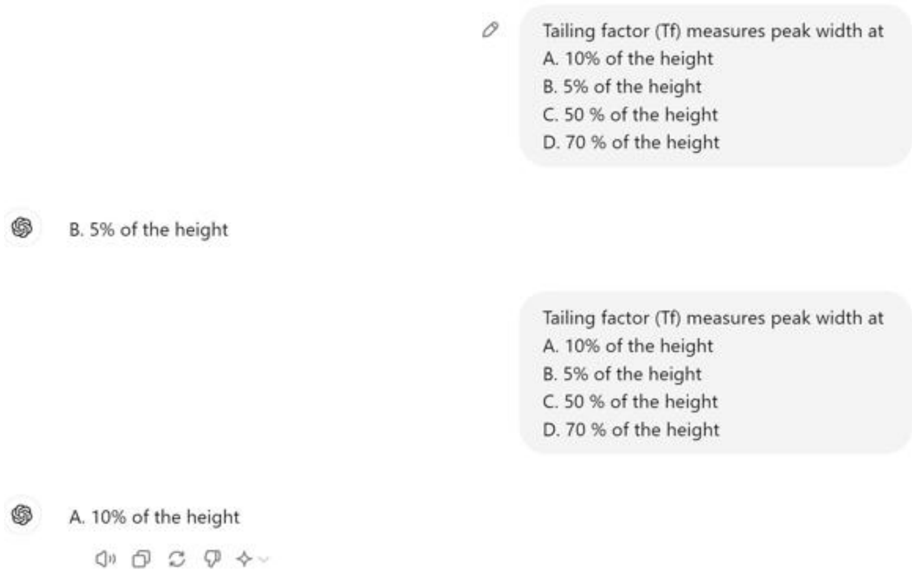


Table 2. An inconsistency in answering a MCQ on molecular weight in LC-MS, using GPT-4o

ChatGPT 4o

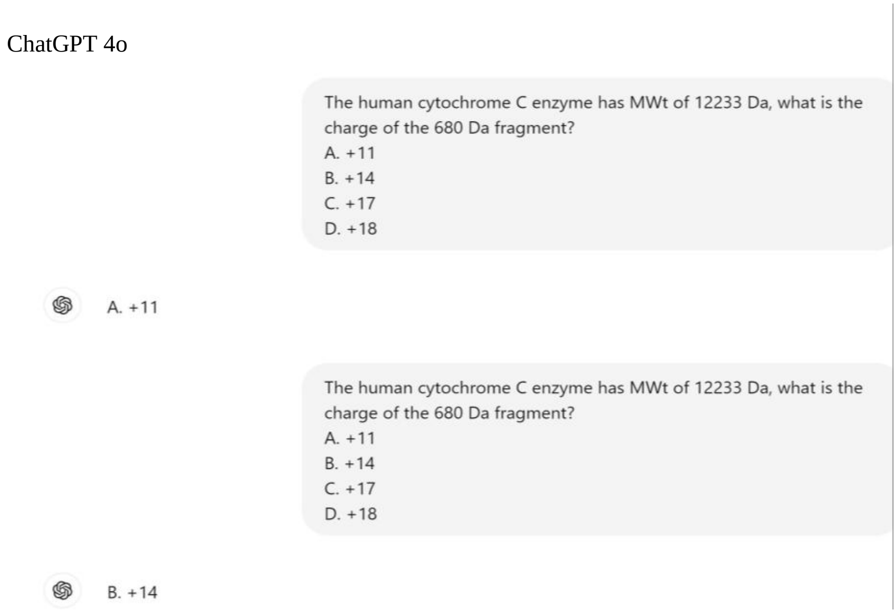
Hispidulin using (-)-ESi-LCMS shows a base peak at 299 Da. What is the molecular weight in Daltons?
A. 300
B. 298
C. 301
D. 298

A. 300

Hispidulin using (-)-ESi-LCMS shows a base peak at 299 Da. What is the molecular weight in Daltons?
A. 300
B. 298
C. 301
D. 298

B. 298

Table 3. An inconsistency in answering a MCQ on calculating the charge of a fragment in LC-MS. using GPT-4o


ChatGPT 4o

The human cytochrome C enzyme has MWt of 12233 Da, what is the charge of the 680 Da fragment?
A. +11
B. +14
C. +17
D. +18

A. +11

The human cytochrome C enzyme has MWt of 12233 Da, what is the charge of the 680 Da fragment?
A. +11
B. +14
C. +17
D. +18

B. +14

When we asked ChatGPT why ChatGPT sometimes provides different answers to the same question, the response was:

1. Context Sensitivity: ChatGPT tries to generate responses that are contextually relevant and can vary its responses based on subtle differences in how a question is asked or the context provided in the conversation. This means that even if the core question is the same, the surrounding conversation or additional details provided may influence the response.
2. Variety and Exploration: Sometimes, ChatGPT may offer different answers to explore different facets of a topic or to provide a broader range of information. This helps in enriching the conversation and offering different perspectives or insights that may be relevant to the user's query.
3. Learning and Improvement: AI models like ChatGPT are constantly learning from interactions. Providing different responses allows the model to learn

which responses are more effective or preferred by users, helping it to improve over time.

4. Randomness or Diversity: In some cases, especially when there are multiple valid answers or interpretations of a question, ChatGPT may choose to provide different responses randomly to maintain variety in the conversation. In essence, the varying responses aim to enhance the interaction by providing more tailored or diverse information, adapting to the nuances of human communication and learning from user feedback.

Categorized by type, 17.5% (21 out of 120) of the MCQs involved calculations and 82.5% were conceptual (99 out of 120) (Figure 4). The percentage of correct answers was notably lower for calculation-based questions across all AIs, with both GPT-4 and GPT-4o achieving the same score of 57.1% (12 out of 21 questions), compared to 47.6% (10 out of 21 questions) for Gemini (Figure 5).

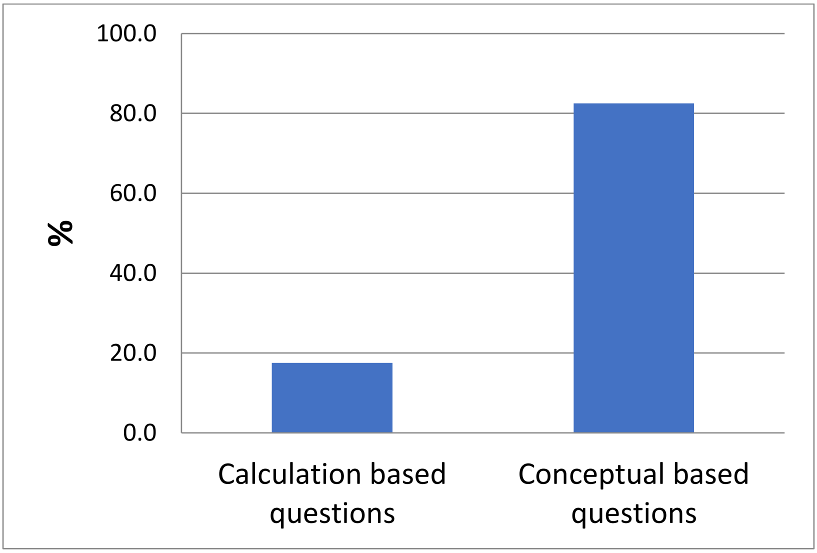


Figure 4. MCQs categorized by type.

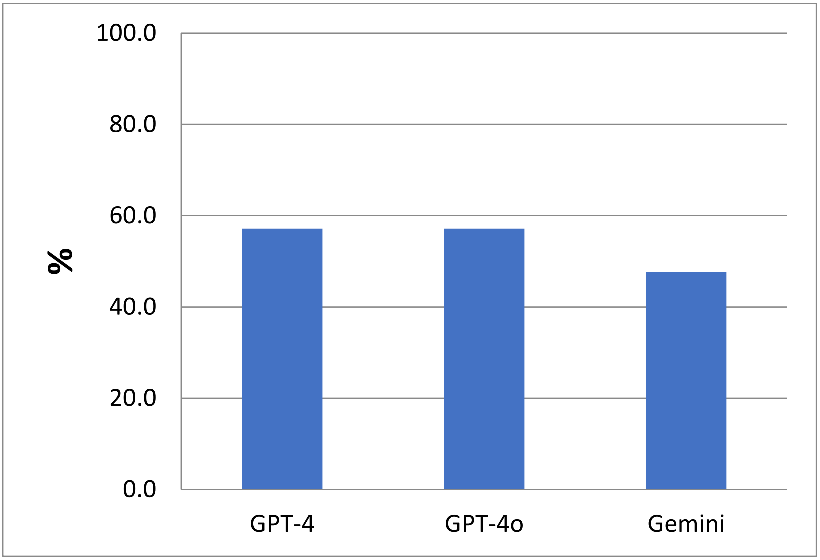


Figure 5. Percentage of calculation-based-questions' correct answers.

In contrast, for the remaining 99 conceptual MCQs, there were a total of six incorrect responses from GPT-4, 8 from GPT-4o, and 5 from Gemini AI (Figure 6).

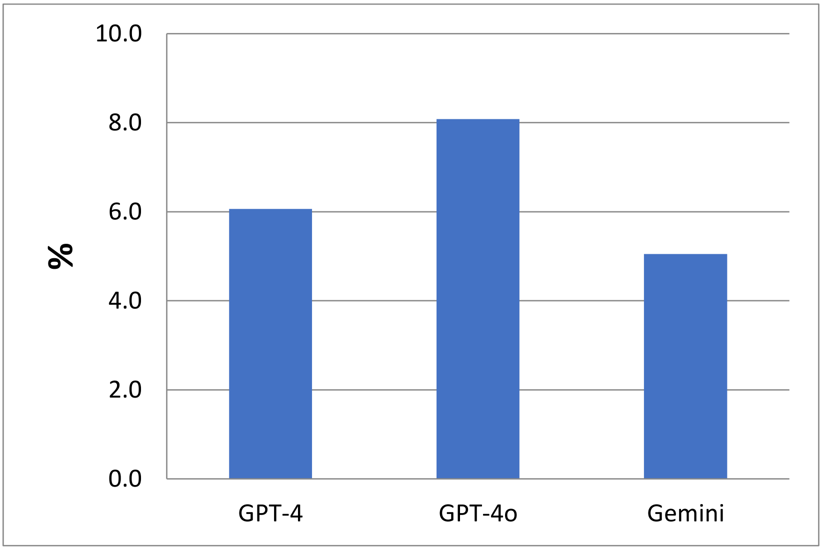


Figure 6. Percentage of conceptual incorrect answers.

Tables 4 and 5 show instances where GPT-4 and Gemini AI struggled with calculations related to chromatography parameters, such as the number of theoretical plates (N) and the height equivalent of theoretical plates (HETP). However, GPT-4 performed correctly in a subsequent attempt. For purposes of comparison, the correct answers from the answer key are provided at the end of the tables.

Tables 6 and 7 demonstrate the challenges Gemini confronted when calculating the selectivity (α) chromatographic parameter and resolution (R).

Table 4. GPT-4's and Gemini's difficulty calculating the number of theoretical plates (N)

Problem	Four histamine H_2 receptor antagonists that inhibit stomach acid production, namely (Nizatidine, Ranitidine, Famotidine and Cimetidine) were analyzed using a C18 column. The experimental setup included a column with dimensions 15cm x 4.6mm ID packed with 5µm particles. The mobile phase consisted of 90% 25 mM Potassium Phosphate (pH 3.0) and 10% methanol, flowing at a rate of 1.0 mL/min at 20°C. Detection was achieved using UV at 214nm with an injection volume of 10µL. The dead time was measured at 1.8 mins. Nizatidine retention equals 4.2 mins, and its width equals 0.412 mins. Ranitidine retention equals 6.9 mins, and its width equals 0.49 mins. Famotidine retention equals 8.1 mins, and its width at half height equals 0.25 mins. Cimetidine retention equals 8.5 mins, and its width at half height equals 0.26 mins. The number of theoretical plates (N) of Famotidine is equal to A. 5815.67 B. 16796.16 C. 518.40 D. 179.49

GPT's answer–second attempt	To calculate the number of theoretical plates (N) for Famotidine, we use the formula: $$N = \frac{5.54 \cdot t_R^2}{W_{0.5}^2}$$ where: • t_R is the retention time, • $W_{0.5}$ is the width at half height of the peak. **Given Data:** • Retention time (t_R) for Famotidine = 8.1 minutes • Width at half height ($W_{0.5}$) for Famotidine = 0.25 minutes **Calculation:** 1. **Calculate t_R^2:** $$t_R^2 = (8.1)^2 = 65.61$$ 2. **Calculate $W_{0.5}^2$:** $$W_{0.5}^2 = (0.25)^2 = 0.0625$$ 3. **Substitute into the formula:** $$N = \frac{5.54 \cdot 65.61}{0.0625}$$ 4. **Perform the calculation:** $$N = \frac{363.7794}{0.0625} \approx 5815.67$$ **Conclusion:** The number of theoretical plates (N) for Famotidine is approximately: **A. 5815.67**

GPT's answer– first attempt	B. 16796.16 - Theoretical plates can be calculated using the formula N=16(w1/2tR)2.
Gemini's answer	B. 16796.16
Key answer	Since we only know the width at half height $W_{1/2}$ rather that width at the base W, therefore, the number of theoretical plate of Famotidine should be calculated using the following formula: $N = 5.54\ (t_R/W_{1/2})^2$ $N = 5.54(8.1/0.25)^2$ N = 5815.6 The right answer is A. 5815.67

Table 5. GPT-4's and Gemini's difficulty calculating the height equivalent of theoretical plates (HETP)

Problem	The height equivalent of theoretical plates of Cimetidine (H) in cm is equal to A. 0.0828 B. 0.0287 C. 0.4270 D. 0.00253

GPT's answer– second attempt

To calculate the height equivalent of theoretical plates (H) for Cimetidine, we need to find the number of theoretical plates (N) first and then use the column length (L) to find H.

Step 1: Calculate the Number of Theoretical Plates (N)

The formula for the number of theoretical plates (N) is:

$$N = \frac{5.54 \cdot t_R^2}{W_{0.5}^2}$$

Given:

- Retention time (t_R) for Cimetidine = 8.5 minutes
- Width at half height ($W_{0.5}$) for Cimetidine = 0.26 minutes

Substitute these values into the formula:

$$N = \frac{5.54 \cdot (8.5)^2}{(0.26)^2}$$

Calculate t_R^2:

$$t_R^2 = (8.5)^2 = 72.25$$

Calculate $W_{0.5}^2$:

$$W_{0.5}^2 = (0.26)^2 = 0.0676$$

Now, substitute these values:

$$N = \frac{5.54 \cdot 72.25}{0.0676}$$

Perform the calculation:

$$N = \frac{400.235}{0.0676}$$

$$N \approx 5919.14$$

Step 2: Calculate the Height Equivalent of Theoretical Plates (H)

The formula for H is:

$$H = \frac{L}{N}$$

where L is the column length in cm (15 cm in this case).

Substitute the values:

$$H = \frac{15}{5919.14}$$

Perform the calculation:

$$H \approx 0.00253\,\text{cm}$$

Conclusion

The height equivalent of theoretical plates (H) for Cimetidine is:

D. 0.00253 cm

GPT-4's answer– first attempt	B. 0.0287 - Height Equivalent to a Theoretical Plate (HETP) = Column length/N
Gemini's answer	B. 0.0287
Key answer	First we have to calculate N for Cimetidine. The width at half height $W_{1/2}$ equals 0.26 mins. Retention time equals 8.5 mins. Therefore, the number of theoretical plate of Cimetidine equals: $N = 5.54 (t_R/W_{1/2})^2$ $N = 5.54(8.5/0.26)^2$ N = 5920 HETP = column length L/N = 15/5920 = 0.00253 The right answer is D. 0.00253

Table 6. Gemini's struggle calculating the selectivity (□) chromatographic parameter

Problem	The selectivity factor (□) between Famotidine and Cimetidine is equal to A. 1.0400 B. 1.2310 C. 1.0635 D. 0.9410

GPT-4's answer

To calculate the selectivity factor (α) between Famotidine and Cimetidine, we use the formula:

$$\alpha = \frac{k'_{\text{Cimetidine}}}{k'_{\text{Famotidine}}}$$

where k' is the capacity factor. The capacity factor is calculated using the formula:

$$k' = \frac{t_R - t_0}{t_0}$$

Given Data:

- Dead time (t_0) = 1.8 minutes
- Retention time for Famotidine (t_R) = 8.1 minutes
- Retention time for Cimetidine (t_R) = 8.5 minutes

Step 1: Calculate the capacity factor (k') for each compound

For Famotidine:

$$k'_{\text{Famotidine}} = \frac{8.1 - 1.8}{1.8} = \frac{6.3}{1.8} \approx 3.5$$

For Cimetidine:

$$k'_{\text{Cimetidine}} = \frac{8.5 - 1.8}{1.8} = \frac{6.7}{1.8} \approx 3.722$$

Step 2: Calculate the selectivity factor (α)

$$\alpha = \frac{k'_{\text{Cimetidine}}}{k'_{\text{Famotidine}}} = \frac{3.722}{3.5} \approx 1.0635$$

Conclusion

The selectivity factor (α) between Famotidine and Cimetidine is:

C. 1.0635

Gemini's answer

A. 1.0400

Key answer	The selectivity factor between Famotidine and Cimetidine = □ □ = ($t_{R\ Cimetidine}$-t_0)/($t_{R\ Famotidine}$-t_0) □ = (8.5-1.8)/(8.1-1.8) = 6.7/6.3 = 1.0635 The right answer is C. 1.0635

Table 7. Gemini's struggle with the calculation of the resolution (R) chromatographic parameter

Problem	The resolution (R) between Famotidine and Cimetidine is equal to A. 0.9255 B. 1.1230 C. 1.5240 D. 2.1470

GPT-4's answer

To calculate the resolution (R) between Famotidine and Cimetidine, we can use the resolution formula for two peaks:

$$R = \frac{2(t_{R2} - t_{R1})}{W_1 + W_2}$$

where:

- t_{R2} and t_{R1} are the retention times of the two compounds,
- W_1 and W_2 are the widths at baseline of the two peaks.

Since the widths at half height ($W_{0.5}$) are given, we need to convert them to the full peak width. For Gaussian peaks, the relationship between the width at half height and the baseline width is approximately:

$$W \approx 1.7 \times W_{0.5}$$

Given Data:

- Retention time of Famotidine (t_{R1}) = 8.1 minutes
- Retention time of Cimetidine (t_{R2}) = 8.5 minutes
- Width at half height of Famotidine ($W_{0.5,1}$) = 0.25 minutes
- Width at half height of Cimetidine ($W_{0.5,2}$) = 0.26 minutes

Step 1: Convert the widths at half height to full widths

For Famotidine:
$W_1 \approx 1.7 \times 0.25 = 0.425$

For Cimetidine:
$W_2 \approx 1.7 \times 0.26 = 0.442$

Step 2: Substitute the values into the resolution formula

$$R = \frac{2(8.5 - 8.1)}{0.425 + 0.442}$$
$$R = \frac{2 \times 0.4}{0.867}$$
$$R \approx \frac{0.8}{0.867}$$
$$R \approx 0.922$$

The closest option to this calculated resolution is:

A. 0.9255

Gemini's answer	D. 2.1470
Key answer	Since we only know the width at half height $W_{1/2}$ rather that width at the base W for both peaks, therefore, resolution should be calculated using the following formula: $R = 1.18\,(t_{R\ Cimetidine} - t_{R\ Famotidine})/(W_{1/2\ Cimetidine} + W_{1/2\ Famotidine})$ R = 1.18 (8.5-8.1)/(0.25+0.26) = 0.9255 The right answer is A. 0.9255

Table 8 presents another example of the difficulties encountered by GPT-4 and Gemini in calculating the limit of linearity (LOL). However, GPT-4 eventually provided the correct answer in a subsequent attempt.

Table 8. GPT-4's and Gemini's struggle calculating the limit of linearity (LOL).

Problem	What is the limit of linearity (LOL) if the detector's linear dynamic range spans three orders of magnitude and the limit of quantitation (LOQ) is 10 parts per million (ppm)? A. 10 ppm B. 100 ppm C. 1000 ppm D. 10000 ppm

GPT-4's answer– second attempt	To determine the limit of linearity (LOL), we need to understand the relationship between the limit of quantitation (LOQ) and the linear dynamic range of the detector. Given that the detector's linear dynamic range spans three orders of magnitude and the LOQ is 10 parts per million (ppm), we can find the LOL. **Understanding the Relationship:** The linear dynamic range spanning three orders of magnitude means the maximum concentration within this range is 1000 times the minimum concentration (LOQ). **Calculation:** If the LOQ is 10 ppm, then the LOL can be calculated as follows: $\text{LOL} = \text{LOQ} \times 10^3$ **Substituting the given LOQ:** $\text{LOL} = 10\,\text{ppm} \times 1000$ $\text{LOL} = 10000\,\text{ppm}$ Therefore, the limit of linearity (LOL) is: **D. 10000 ppm**
GPT-4's answer– first attempt	C. 1000 ppm - The linear dynamic range of a detector indicates its upper limit based on the LOQ.
Gemini's answer	C. 1000 ppm
Key answer	The linear dynamic range covers three orders of magnitude, and with an LOQ of 10 ppm, the LOL would be calculated as 1000 times 10, resulting in 10,000 ppm. The right answer is D. 10000 ppm

Tables 9 and 10 highlight how GPT-4 handled questions relating to LC-MS chromatography.

Table 9. GPT-4's estimation of the molecular weight of hispidulin

Problem	Hispidulin, as detected by (-)-ESI-LCMS, exhibits a base peak at 299 atomic mass units (Da). What is the molecular weight of Hispidulin in Da? A. 300 B. 298 C. 301 D. 298
GPT-4's answer	B. 298 - The molecular weight would be one less than the observed base peak in negative mode
Gemini's answer	A. 300
Key answer	The molecular weight of Hispidulin would be one more than the observed base peak in negative mode. Therefore 299 + 1 = 300 Da The right answer is A. 300 Da

Table 10. GPT-4's struggle with estimating the charge of human cytochrome C enzyme from a fragment

Problem	The molecular weight (M) of the human cytochrome C enzyme is 12233 Daltons. What is the charge (n) of the 680 Dalton fragment? A. +11 B. +14

	C. +17 D. +18
GPT-4's answer	A. +11 - This requires additional context for a precise calculation but assumes charge states for fragmentation.
Gemini's answer	D. +18
Key answer	For protonated human cytochrome C enzyme $[M+nH]^{n+}$ The positive charge can be calculated by applying the equation [M+n/n = 680] The charge (n) = +18 The right answer is D. + 18

Tables 11 and 12 present the challenges Gemini faced with GC and drug-solution concentration conversion questions.

Table 11. Gemini's struggle estimating the total run time in GC

Problem	The temperature gradient of the gas chromatograph (GC) began at 150°C and increased steadily to 250°C at a rate of 20°C per minute. After reaching 250°C, it was maintained at this temperature for an additional 3 minutes. What is the total run time in minutes? A. 3 B. 10 C. 7 D. 8

GPT-4's answer	To determine the total run time of the gas chromatograph (GC), we need to account for both the time taken to increase the temperature from 150°C to 250°C and the time it was maintained at 250°C. **Step 1: Calculate the time to increase the temperature** The temperature gradient is given as 20°C per minute. The initial temperature is 150°C, and the final temperature is 250°C. The total temperature change is: $250^\circ C - 150^\circ C = 100^\circ C$ At a rate of 20°C per minute, the time taken to increase the temperature by 100°C is: $\frac{100^\circ C}{20^\circ C/\text{minute}} = 5 \text{ minutes}$ **Step 2: Calculate the time maintained at 250°C** The temperature is maintained at 250°C for an additional 3 minutes. **Step 3: Calculate the total run time** The total run time is the sum of the time to increase the temperature and the time maintained at 250°C: $5 \text{ minutes} + 3 \text{ minutes} = 8 \text{ minutes}$ Therefore, the total run time in minutes is: D. 8
Gemini's answer	C. 7
Key answer	To find the total runtime in minutes, we need to calculate the time taken for the temperature to increase from 150°C to 250°C and add the 3 minutes of holding time. First, calculate the time taken to reach 250°C: Time = Temperature change/ Ramp = 250°C -150°C /20°C per minute = 5 mins Total run time = 5 mins + holding time = 8 mins The right answer is D. 8 mins

Table 12. Gemini’s struggle with concentration units’ conversion of drugs

Problem	Percentage weight in volume (%w/v) is typically utilized to denote the concentration of active pharmaceutical ingredients (APIs) in liquid formulations like injections, infusion and eyedrops. What is (%w/v) of a solution with a concentration of 100□g/ml? A. 0.1% B. 0.01% C. 1% D. 10%
GPT-4’s answer	To convert the concentration of a solution from μg/mL to percentage weight/volume (%w/v), we need to understand the relationship between these units. Percentage weight/volume (%w/v) is defined as grams of solute per 100 mL of solution. **Given:** • Concentration = 100 μg/mL First, convert μg to g: $100\,\mu\mathrm{g/mL} = 100 \times 10^{-6}\,\mathrm{g/mL}$ $100\,\mu\mathrm{g/mL} = 0.0001\,\mathrm{g/mL}$ **Converting to %w/v:** Since %w/v is defined as grams per 100 mL of solution, we need to convert 0.0001 g/mL to a per 1(mL basis: $0.0001\,\mathrm{g/mL} \times 100\,\mathrm{mL} = 0.01\,\mathrm{g}/100\,\mathrm{mL}$ Therefore, the concentration in %w/v is: $0.01\,\%\mathrm{w/v}$ Thus, the %w/v of a solution with a concentration of 100 μg/mL is: **B. 0.01%**

Gemini's answer	D. 10%
Key Answer	The concentration of the API equals to 100 µg/ml Covert micrograms to grams: 100 µg = 100×10^{-6} grams = 0.0001 grams % (w/v) = (0.0001 g/1 ml) x 100 = 0.01% The right answer is B. 0.01%

In Table 13, we observe GPT-4's difficulty handling the use of ion-pair chromatography for polar basic drugs.

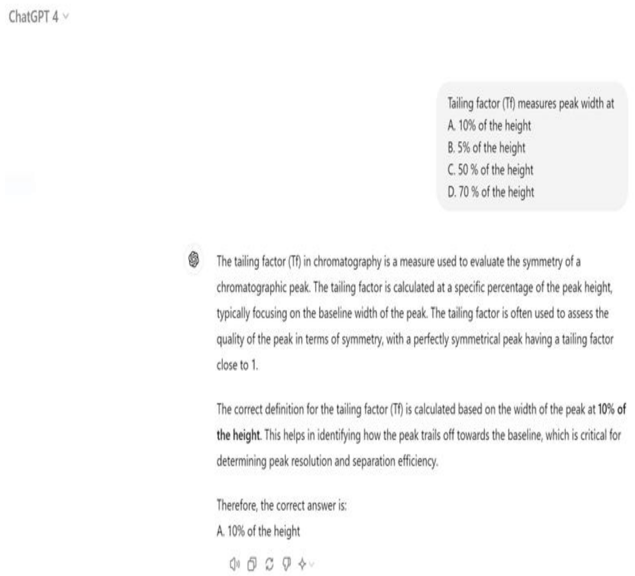
Table 13. GPT-4's difficulty with ion-pair chromatography of a polar basic drug

Problem	If a highly polar basic drug elutes with the void peak using reversed phase chromatography on C18 column, then to enhance its retention: A. Employ heptanesulfonic acid ion pair. B. Utilize tertbutyl quaternary ammonium salt. C. Depends on the specific chemical structure. D. Use less organic solvent percentage in the mobile phase
GPT-4's answer	B. Utilize tertbutyl quaternary ammonium salt.
Gemini's answer	A. Use heptanesulfonic acid ion pair
Key answer	The use of Heptanesulfonic acid ion pair reagents are typically used in reversed-phase chromatography when dealing with highly polar basic compounds that have poor retention on the stationary phase. They are

	particularly effective for enhancing the retention of basic polar analytes by forming ion pairs with them, thereby improving their chromatographic separation and peak shape.

Lastly, in Table 14, GPT-4 and GPT-4o had difficulty answering a question about the tailing factor (Tf) of compounds.

Table 14. GPT-4's and GPT-4o's difficulty explaining the tailing factor of peaks

Problem	Tailing factor (Tf) measures peak width at A. 10% of the height B. 5% of the height C. 50 % of the height D. 70 % of the height
GPT-4's answer	

GPT-4o's answer	ChatGPT 4o The tailing factor (Tf) in chromatography is a measure of the asymmetry of a peak and is used to describe how much a peak deviates from the ideal Gaussian shape. The tailing factor is calculated based on the width of the peak at a specific percentage of its height. The tailing factor is typically measured at **10% of the peak height**. Therefore, the correct answer is: A. 10% of the height
Gemini's answer	D. 10%
Key answer	The tailing factor is calculated by drawing a perpendicular line from the peak center to the baseline. Subsequently, the peak width and the front half-width are measured at 5% of the peak height.

4. Discussion

ChatGPT and Gemini have sparked significant interest among students, academics, and researchers for their ability to provide prompt and articulate responses to queries across various fields of chemistry [14]. They are valued as tools for enhancing scientific knowledge, generating essays, and offering explanations. However, ongoing academic discussion revolves around the validity of their knowledge and intelligence [14, 15]. This study remains focused on assessing the effectiveness of GPT-4, GPT-4o, and Gemini for answering MCQs on tests and how well they can handle the material and support students learning basic analytical chemistry, particularly in instrumental analysis education for pharmacy students in their senior year. The study indicates that both AIs performed well on conceptual exam questions but encountered challenges with calculation-based assessments (Figure 4). On the whole, there is limited research on their performance in instrumental analysis education, especially in regard to calculation problems [7].

On most MCQs, both AIs encountered no difficulties with conceptual material but had problems handling calculations (Figures 2, 5). Tables 1-3 present errors made by GPT-4 and GPT-4o, showing incorrect answers. In some cases, when presented with the same question twice, they responded inconsistently, initially giving an incorrect response and then a correct response on the second attempt, or vice versa (Figure 3). Table 1 shows the inconsistency in GPT-4's determination of the tailing factor (Tf), a measure of peak shape symmetry. The tailing factor is calculated by drawing a vertical line from the peak's apex to the baseline, measuring the full peak width and the front half-width at 5% of the peak height (above the baseline), and dividing the full peak width by twice the front half-width. Similar inconsistencies are observed in Tables 2 and 3 with GPT-4o, particularly in calculating molecular weight and charge in the LC-MS domain. For instance, GPT-4o provided varying incorrect answers for the charge of a fragment (+11 and then +14, when the correct answer was +18). According to ChatGPT, these inconsistencies are attributable to context sensitivity, whereby ChatGPT aims to generate responses that are contextually relevant. This sensitivity implies that slight differences in how a question is asked or additional context provided in a conversation may influence the response, even if the core question remains the same.

Tables 4 and 5 show errors made by both Gemini and GPT-4 in their calculations of the number of theoretical plates and the height equivalent of theoretical plates (HETP). On the second attempt, GPT-4 provided the accurate answer. Gemini, on the other hand, provides answers with less explanation of the process used to arrive at them. The number of theoretical plates (*N*) can be calculated using the following formulas: $N = 5.545\ (t_R\ /W_{1/2})^2$ or $N = 16\ (t_R\ /W)^2$, where $W_{1/2}$ represents the peak width at half of the peak height, *W* represents the peak width at the base, and t_R represents the retention time of the eluted peak. The *N* value signifies the efficiency of the column. Generally, the quantity of plates depends on the column length; specifically, a longer column tends to yield a larger number of plates (Table 1).

Efficiency can also be expressed in terms of plate height (*H*) or the height equivalent of a theoretical plate (HETP), calculated as $HETP = L/N$, where *L* stands for the column length. When presented with questions like those in Tables 4 and 5, asking for calculations of *N* and *H*, both GPT-4 and Gemini were initially unsure which equation to apply. The questions explicitly provided the peak width at half of the peak height rather than *W* at the base, leading to incorrect responses. However, in the second attempt, GPT-4 came up with the correct answer (Table 5).

Tables 6 and 7 demonstrate Gemini's difficulties in calculating chromatographic parameters such as selectivity (α) and resolution (*R*), whereas

GPT-4 provided correct responses. The selectivity or selectivity factor (□) refers to the chromatographic system's capability to differentiate between neighboring components. It is typically quantified by comparing the retention factors (k') of two specific peaks (Table 6). The primary parameter of significance in HPLC is resolution (Table 7). A resolution value of 1.5 or higher between adjacent peaks ensures that sample components are separated enough to enable accurate measurement of each peak's area and height. Resolution is typically evaluated using three equations, with the most commonly used being: $R = 2\ (t_{RB} – t_{RA})\ /\ (W_A + W_B)$, where $t_{RB} > t_{RA}$.

Another approach to calculating resolution is $R = 1.18\ (t_{RB} - t_{RA})\ /\ (W_{1/2A} + W_{1/2B})$, where the given width is at half the height. The third fundamental equation for resolution depends on three key factors: selectivity (the separation factor), efficiency (N), and retention (the capacity factor) where $R = (k'/k'+1)$(□-1/1)$(\sqrt{N}/4)$. GPT-4 correctly provided the R value, despite its inability to calculate N and H (Table 4 and 5) using the width at half height.

Another example is when GPT-4 and Gemini were tasked with calculating the limit of linearity (LOL), based on a limit of quantitation (LOQ) of 10 ppm, given that the detector's linear dynamic range spans three orders of magnitude. Both provided an incorrect answer of 1000 ppm. It appears that their focus on the three orders of magnitude led them to mistakenly suggest 1000 ppm instead of the correct value of 10,000 ppm (Table 8).

We noticed that GPT-4 struggled with estimating the molecular weight of the hispidulin compound using the negative ESI mode in LC-MS, based on the m/z value provided (Table 9). Similar difficulties were noted when it attempted to estimate the charge of the human cytochrome C enzyme from specific fragments (Table 10). This underscores the need for further training in calculation-based tasks for both AIs. It also demonstrates that both models performed better on questions that required recall-based answers rather than those demanding critical thinking or problem-solving skills [7, 14, 15].

Tables 11 and 12 describe instances when Gemini struggled with calculating the total run time of a GC experiment and converting units to express drug concentration of solutions, respectively. In contrast, GPT provided a detailed response with a step-by-step explanation to reach the correct answer.

Table 13 demonstrates an example of GPT-4's difficulty in determining the appropriate ion pair reagent for use in the mobile phase to increase retention of very polar basic drug, while Table 14 highlights the challenges faced by both GPT-4 and GPT-4o in defining the tailing factor.

Overall, although both models showed comparable performance on multiple-choice questions (MCQs), GPT-4 and GPT-4o distinguished themselves by providing relatively precise answers and comprehensive explanations, as illustrated by the data in the preceding tables. This indicates that these AI models are particularly effective in not only delivering correct responses but also in elucidating their reasoning, which enhances the learning process. Therefore, ChatGPT proves to be a valuable asset in pharmacy education. Its strengths lie in its capacity to assist students with understanding complex pharmaceutical concepts, performing intricate calculations, and solving challenging problems. By offering detailed explanations and clarifications, ChatGPT helps bridge gaps in knowledge and supports students in mastering key aspects of pharmaceutical science. This makes it an excellent tool for enhancing educational outcomes and improving comprehension in the field.

5. Conclusion

Artificial intelligence, represented here by GPT-4, GPT-4o, and Gemini, has proven to be useful for students and academics in addressing fundamental conceptual questions in analytical chemistry, especially within the realm of instrumental analysis in undergraduate pharmacy classes. The AIs in this study excelled at rapidly and accurately answering most of the MCQs. However, they had trouble performing complex calculations and handling unexpected scenarios that require creativity, adaptability, and human judgment. Humans still demonstrate superior proficiency with respect to critical thinking, problem-solving, and decision-making tasks. While AI systems perform well when conducting straightforward calculations and yield accurate results, they often struggle with complex calculations, which occasionally lead to erroneous output. Overall, our findings indicate that GPT-4, GPT-4o, and Gemini hold promise for supporting pharmacy students and faculty members in their educational endeavors. Further research across various disciplines within the pharmaceutical sciences is encouraged to refine these insights and propose more definitive enhancements.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

Baum, Z. J.; Yu, X.; Ayala, P. Y.; Zhao, Y.; Watkins, S. P.; Zhou, Q., Artificial intelligence in chemistry: current trends and future directions. *Journal of Chemical Information and Modeling* 2021, 61, (7), 3197-3212.

Daher, W.; Diab, H.; Rayan, A., Artificial intelligence generative tools and conceptual knowledge in problem solving in chemistry. *Information* 2023, 14, (7), 409.

Daher, W.; Diab, H.; Rayan, A., Generative Artificial Intelligence in Chemistry Problem Solving across Versions and Languages. 2024.

Govender, R., My AI students: Evaluating the proficiency of three AI chatbots in completeness and accuracy. *Contemporary Educational Technology* 2024, 16, ep509.

Habeeba, S., Use of Artificial Intelligence in Drug Discovery and its Application in Drug Development. *Asian Journal of Research in Chemistry* 2023, 83-90.

Imran, M.; Almusharraf, N., Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart Learning Environments* 2024, 11, (1), 22.

Joshi, P. B., Navigating with chemometrics and machine learning in chemistry. *Artificial Intelligence Review* 2023, 56, (9), 9089-9114.

Mohamed, Y. A.; Kannan, A.; Bashir, M.; Mohamed, A. H.; Adiel, M. A. E.; Elsadig, M. A., The Impact of Artificial Intelligence on Language Translation: A Review. *IEEE Access* 2024, 12, 25553-25579.

Obaigbena, A.; Lottu, O. A.; Ugwuanyi, E. D.; Jacks, B. S.; Sodiya, E. O.; Daraojimba, O. D., AI and human-robot interaction: A review of recent advances and challenges. *GSC Advanced Research and Reviews* 2024, 18, (2), 321-330.

Rane, N.; Choudhary, S.; Rane, J., Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *Performance, Architecture, Capabilities, and Implementation* (February 13, 2024) 2024.

Ray, P. P., ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 2023, 3, 121-154.

Rial, R. C., AI in analytical chemistry: Advancements, challenges, and future directions. *Talanta* 2024, 125949.

Rocha, W. F.; Nogueira, R.; Fidelis, C. H.; Poppi, R., Application of spectroscopic techniques in the quality control of pharmaceuticals: An overview. In 2013; pp 1-28.

Rochet, J.-C., Pharmaceutical analysis: a textbook for pharmacy students and pharmaceutical chemists. *American Journal of Pharmaceutical Education* 2006, 70, (2), 3.

Singh, Y. R.; Shah, D. B.; Maheshwari, D. G.; Shah, J. S.; Shah, S., Advances in AI-Driven retention prediction for different chromatographic techniques: unraveling the complexity. *Critical Reviews in Analytical Chemistry* 2023, 1-11.

Swami, K., Emerging Trends in the Application of Artificial Intelligence and Machine Learning for Analytical Chemistry: Enhancing Precision and Automation. *International Scientific Journal of Engineering and Management* 2024, 03, 1-9.

Tran, D.; Pascazio, L.; Akroyd, J.; Mosbach, S.; Kraft, M., Leveraging Text-to-Text Pretrained Language Models for Question Answering in Chemistry. *ACS Omega* 2024, 9.

Zhou, X.; Zhang, S.; Agarwal, M.; Akroyd, J.; Mosbach, S.; Kraft, M., Marie and BERT─A Knowledge Graph Embedding Based Question Answering System for Chemistry. *ACS Omega* 2023, 8.

دراسة معمّقة لموثوقيّة تشات جي بي تي وجيميني في الإجابة على الأسئلة الكيميائيّة المعقّدة في التحليل الآلي

صالح أبو لافي و أنورريّان

الملخّص

تعتبر أنظمة الذكاء الاصطناعي المتقدمة مثل تشات جي بي تي (لا سيما الإصدارات الأخيرة GPT-4 و GPT-4o) وجيميني Gemini أدوات قيمة في العملية التعليمية. يفحص هذا البحث أداء ChatGPT و Gemini في فهم وحل مسائل في موضوع الكيمياء، مع التركيز على الكيمياء التحليلية والتحليل الآلي لمواد ضمن مساق لطلاب الصيدلة في مرحلة البكالوريوس. تهدف الدراسة إلى تقييم قدرات حل المسائل بواسطة كلتا الأداتين (تشات جي بي تي وجيميني) باستخدام مجموعة من 120 سؤالاً متعدد الخيارات (MCQs) تم اختيارها عشوائيًا من أسئلة الامتحانات المقدمة للطلاب الجامعيين في السنوات الأخيرة. شملت هذه الأسئلة تقنيات مختلفة: 49 سؤالًا (40.8%) عن الكروماتوغرافيا، و57 سؤالًا (47.5%) عن الطيفية، و14 سؤالًا (11.7%) عن الأنظمة المقترنة. تم تقييم استجابات كل أداة من أدوات الذكاء الاصطناعي مقابل مفتاح إجابة تم اعداده من قبل متخصصين في المجال، بحيث تم منح الإجابات الصحيحة درجة 1 والخاطئة 0.

سجلت الدراسة حالات عدم الاتساق حيث قدم تشات جي بي تي، عند سؤاله نفس الأسئلة مرتين، إجابات متناقضة، بعضها صحيح وبعضها خاطئ. حقق Gemini درجة إجمالية قدرها 87.5% (105 من أصل 120 سؤال) عبر جميع الأسئلة، بينما سجل GPT-4 نسبة 84.2% (101 من أصل 120) وسجل GPT-4o نسبة 85.3% (103 من أصل 120). لوحظ عدم الاتساق لـ GPT-4o بمعدل 7.5% (9 من أصل 120 سؤال)، بينما عرض GPT-4 معدل عدم اتساق أقل يبلغ 2.5% (3 من أصل 120 سؤال).

عند تصنيف الأسئلة حسب النوع (أسئلة حسابات وأسئلة مفاهيمية)، كان لدينا 17.5% (21 من أصل 120 سؤال) أسئلة حسابات، والباقي (99 سؤالًا) كانت أسئلة مفاهيمية. عند تحليل النتائج حسب النوع، انخفضت نسبة الإجابات الصحيحة بشكل ملحوظ للأسئلة التي تعتمد على الحسابات عبر جميع أدوات الذكاء الاصطناعي، حيث حقق كل من GPT-4 و GPT-4o نفس الدرجة وهي 57.1% (12 من أصل 21 سؤال)، مقارنة بـ 52.4% (10 من أصل 21 سؤال) لـ Gemini. بشكل عام، قدم-GPT

4 و GPT-4o أداءً مماثلًا من حيث إجمالي الإجابات الصحيحة، مع فائدة إضافية تتمثل في تقديم تفسيرات منطقية خطوة بخطوة لحلولهم. تعتبر هذه القدرة مفيدة بشكل خاص لطلاب الصيدلة؛ وتساعد الطلاب على فهم المواضيع والمصطلحات بشكل أفضل، وتعالج بعض التحديات المرتبطة في تعليم موضوع الكيمياء التحليلية والتحليل الآلي.

في مساق التحليل الآلي في تعليم الكيمياء التحليلية وتحاليل الأدوية لطلاب قسم الصيدلة، يمكن للذكاء الاصطناعي مساعدة الطلاب وتعزيز فهمهم للمفاهيم والمصطلحات. نوصي بمعالجة قضايا محددة، مثل عدم الاتساق في الاستجابات والحاجة إلى تنقيح محدد للمجال، لتحسين أداء هذه الأنظمة بشكل أكبر وتعزيز إمكانيات استخدامها في العملية التعليمية.