Asian Journal of the Scholarship of Teaching and Learning
2026, Vol. 16, No. 1, 1-17

AJSoTL

Article

A Feasibility Study on the Use of Large Language Models in Supporting Adaptive Learning

Cheng Keat TAN[1], Qing Hao NG[2], Seh Yi Joseph TAN[3], Yin Ni NG [1]
[1] School of Applied Science, Nanyang Polytechnic, 180 Ang Mo Kio Avenue 8
[2] Lucence, 211 Henderson Rd
[3] GlaxoSmithKline Asia House, 23 Rochester Park

ABSTRACT

Large Language Models (LLMs) have gained prominence as adaptive learning tools. Although effective in assessing multiple-choice questions (MCQs), their accuracy and feedback validity remain uncertain. This feasibility study examines the accuracy, validity, and scientific substantiation of LLMs' responses to expert-generated pharmacology MCQs and assesses the automated generation of MCQs to elucidate their potential roles and limitations in adaptive learning.

Fifty MCQs, designed and validated by academic pharmacists, were used to test the accuracy of four LLMs. Each question was classified according to Bloom's Taxonomy. Direct prompting was applied to generate responses for each question. Responses were analysed for accuracy, validity of rationale, and the existence and relevance of supporting references. Chi-Square test and Fisher-Freeman-Halton Test were used to evaluate quantitative findings.

Among the four LLMs, ChatGPT-4o achieved the highest accuracy (84%), followed by Google Gemini 1.5 (Gemini) (80%), Microsoft Copilot (Copilot) (72%), and Claude Sonnet 3.5 (Claude) (68%). An answer-rationale discordance and a decline in performance with an increased cognitive complexity, stratified through Bloom's Taxonomy, were noted across the LLMs. Artificial hallucinations were observed in the study. These limitations underline the challenges of using LLMs in complex, evidence-driven disciplines like pharmacology.

LLMs, as an educational tool, may provide value to adaptive learning. However, limitations in logical reasoning, scientific support, and higher-order thinking highlight the need for cautious adoption. Continuous efforts to validate LLMs with larger, more diverse question banks are also necessary to fully investigate their potential in adaptive learning.

Keywords: Adaptive Learning, Large Language Models, Pharmacology, Competency-Based Education

Correspondence: Mr Cheng Keat TAN (tan_cheng_keat@nyp.edu.sg)

Recommended Citation:

Tan, C. K., Ng, Q. H., Tan, S. Y. J., & Ng, Y. N. (2026). A Feasibility Study on the Use of Large Language Models in Supporting Adaptive Learning. *Asian Journal of the Scholarship of Teaching and Learning*, *16*(1). 1-17.

INTRODUCTION

Large Language Models (LLMs) represent a significant breakthrough in the growing realm of artificial intelligence (AI). Mimicking human-to-human communication, LLMs such as Chat Generative Pre-Trained Transformer (ChatGPT) serve as conversational agents by analysing, processing, and summarising an enormous amount of information to generate human-like responses to any input by users (OpenAI, 2022).

The increased prominence of LLMs usage aligns with the global shift from the traditional one-size-fits-all education framework to a learner-centric, adaptive learning (Bai & Wan, 2025). Adaptive learning dynamically tailors content to individual progress, creating a customised learning experience, and improving engagement and content mastery (Bai & Wan, 2025; Pelánek, 2025). LLMs, as a cyberlearning technology, offer a personalised experience through machine learning, customising pedagogy, content, feedback, and pace to suit individual learners (Olney et al., 2022). For instance, LLMs can assess learners' responses to multiple-choice questions (MCQs), instantly provide feedback on their accuracy and support learners' metacognitive development (Tackett et al., 2018). This reinforces conceptual knowledge and supports mastery of challenging disciplines such as pharmacology, where learners often struggle with complex terminology and abstract concepts (Cain et al., 2023; Dempere et al., 2023).

MCQs are a widely accepted, reliable, and ubiquitous form of assessment to evaluate knowledge and skills in clinical education. Clinical licensing examination bodies in the United States, Taiwan and Singapore, for instance, employ MCQ-based assessments designed in alignment with Bloom's Taxonomy to assess learners' abilities across various cognitive levels, from factual recall to high-order reasoning (Cheung et al., 2023; Meo et al., 2023; Singapore Ministry of Health, 2023). LLMs have demonstrated notable success in accurately attempting MCQs. For example, ChatGPT 3.5 and ChatGPT-4 achieved 72% and 87% accuracy, respectively, on the United States Medical Licensing Examination (USMLE) (Lievin et al., 2023; Nori et al., 2023).

Previous studies have evaluated ChatGPT's accuracy in various clinical subfields. A translated set of pharmacology examination questions in the Korean Comprehensive Basic Medical Sciences Examination (CBMSE) yielded a 76% accuracy rate on ChatGPT-4o (Choi, 2023). Yet, failures such as ChatGPT 3.5's inability to pass the English edition of Taiwan's Pharmacist Licensing Examination highlight the uncertainty of its performance (Wang et al., 2025), which is crucial in its adoption in education.

The emergence of alternative platforms, such as Claude, Microsoft Copilot, and Google Gemini, expands the armamentarium of LLMs. However, their effectiveness in facilitating adaptive learning, through MCQs with accurate responses and valid feedback, remains largely unexplored. This feasibility study aims to evaluate the use of LLMs in adaptive learning by assessing the tools' accuracy in answering expert-generated MCQs, the validity of AI-generated feedback, the scientific substantiation with references and the quality of LLM-generated MCQs to reveal insights into their relevance in clinical education.

METHODS

Question Design and Question Validation of the Question Bank and Answer

This study first evaluated expert-generated pharmacology MCQs across four LLMs: ChatGPT-4o, Microsoft Copilot, Claude, and Google Gemini.

A set of 50 expert-generated MCQs was developed by the corresponding author, guided by the 'core concepts of pharmacology' (Guilding et al., 2024), established by the British Pharmacological Society. These MCQs were designed to ensure conceptual breadth and to represent the essential pharmacological principles, rather than to achieve comprehensive coverage across all sub-disciplines in this feasibility study. The MCQs were also classified according to Bloom's Taxonomy, on the scale of 'remembering', 'understanding', 'applying' and 'analysing' (Armstrong et al., 2010) (**Supplementary Table 1**).

To ensure effective benchmarking when testing the accuracy of LLMs' responses, the question set had to first be validated. Question design and classification, concepts covered, and accuracy of answers and explanations were independently reviewed by two pharmacists trained in the field of study. Any discrepancies were discussed and resolved via a consensus discussion between the corresponding author and the reviewers before finalising the expert-generated question set.

Large Language Models

This study analysed the accuracy of the responses generated by four LLMs, specifically, ChatGPT-4o, Microsoft Copilot, Google Gemini and Claude. The versions of the LLMs used are as follows:

- ChatGPT-4o, a successor of GPT-4 Turbo, released by OpenAI on 13 May 2024 (OpenAI, 2024).
- Microsoft Copilot, which utilised the Microsoft Prometheus model, was released on 1 October 2024 (Suleyman, 2024).
- Gemini 1.5, a successor of Gemini 1.0, released by Google DeepMind on 15 February 2024 (Pichai & Hassabis, 2024).
- Claude Sonnet 3.5, released by Anthropic on 20 June 2024, with information updated till April 2024 (Anthropic, 2024).

Prompt Inputs and Engineering

The same engineered, direct, and straightforward prompt was given to each of the 4 LLMs before querying for an answer on the application programming interface by supplying a question from the question bank (**Table 1**). The direct prompt sought to maintain consistency across all LLMs and eliminate potential biases associated with multi-step reasoning. The responses of the various LLMs ('LLM Output 2') were collated for data analysis. The chat window and memories were cleared after every question to prevent one response from influencing the next.

Table 1. Prompt engineering and inputs of the multiple-choice questions into the large language models.

User Input 1: I am working on a Multiple-Choice Question (MCQ) related to pharmacology. Each MCQ includes a question or context followed by four listed options. For the MCQ that I will provide next, I need your assistance to: Select the correct answer from the four options. Provide a scientific rationale for selecting that option. Provide a reference of scientific journals used to formulate the rationale in (2). For each reference, provide the Uniform Resource Locator (URL), title, authors, edition, page number, and other relevant details. Specify the exact paragraph, sentence, or line from the reference, which you provided in (3), where you found the relevant information. Do you understand?
LLM Output 1: (LLMs response)
User Input 2: (Input Questions and options from the Question Bank)
LLM Output 2: (LLMs response with answers, rationale, citations and references)

Generation of LLM's Output and Categorisation of Data

This study measured four key categorical variables, specifically, the accuracy of LLMs' responses on each MCQ, the validity of the rationale behind the selected option, the presence of citations to support the rationale, and the relevance of the citations to substantiate the rationale. Two authors extracted and evaluated data from four LLMs – ChatGPT, Gemini, Copilot, and Claude. The evaluated data were adjudicated using a straightforward, binary classification, as either 'Yes' or 'No'. The differences in judgment between the two authors were resolved through discussion until both authors reached a consensus.

Design and Evaluation of Automated Question Generation Using ChatGPT-4o

To evaluate the capacity of LLMs to generate adaptive learning materials, ChatGPT-4o was employed to design MCQs. ChatGPT-4o was selected for preliminary analyses because the LLM has a descriptively higher accuracy compared with other available models.

Question generation was guided by the 24 concepts outlined in the *Core Concepts of Pharmacology* (Guilding et al., 2024), which served as the reference framework for content selection. For each concept, ChatGPT-4o was prompted to generate MCQs representing four cognitive levels of Bloom's Taxonomy – 'remembering', 'understanding', 'applying' and 'analysing' (Armstrong et al., 2010). A standardised prompt was used for all items (**Supplementary Table 2**), resulting in a total of 96 questions generated.

The quality of the generated items was evaluated against three predefined criteria – accuracy, validity (Armstrong et al., 2010; Yaacoub et al., 2025), and readability (Moore et al., 2024). Each criterion was rated using a four-point rubric (**Supplementary Table 2**). Two independent domain experts served as raters to ensure the reliability of the evaluation process. In instances of scoring discrepancies, raters engaged in a structured resolution process. Initial disagreements were addressed through discussion, and consensus was achieved via deliberation.

In addition to rubric-based ratings, the domain experts were invited to provide free-text qualitative comments to capture nuanced feedback. These comments were subsequently collated and thematically summarised to complement the quantitative findings.

Hypotheses and Statistical Analysis

This study assessed the performance of four LLMs in answering pharmacology MCQs. The primary outcome was response accuracy among the LLMs. Secondary outcomes included the validity of rationales provided, provision of citations, relevance of references, as well as differences in performance across Bloom's Taxonomy levels — remembering, understanding, applying, and analysing — to determine how effectively each LLM handled questions of varying cognitive complexity.

Quantitatively, all variables were described as frequency and proportion [n (%)]. Chi-square test was applied to assess whether accuracy and citation reliability differed across the four LLMs. A post-hoc analysis was conducted to identify differences in performance between pairs of LLMs. Additionally, the Fisher-Freeman-Halton Test was used for three analyses: (1) assessing the ability of each LLM to answer questions across different cognitive levels in Bloom's Taxonomy, (2) evaluating the accuracy of each LLM in answering questions within each Bloom's Taxonomy level, and (3) evaluating the relevance of references obtained. Error analysis was also conducted for the provision of citation and reference relevance (Variables 3 and 4). Citations and references for each LLM were processed to extract unique URL links. Unique links that were non-existent and irrelevant were thematically analysed to identify the underlying causes of the errors.

For the evaluation of automated question generation, each MCQ was independently assessed by two domain experts using a four-point rubric for accuracy, validity, and readability. Scores for each criterion were collated and summarised using mean and standard deviation, stratified by the four Bloom's Taxonomy levels (Remembering, Understanding, Applying, and Analysing).

Inter-rater agreement was quantified, using Cohen's kappa, to ensure reliability of the evaluation process. In addition, thematic analysis of the qualitative comments was conducted to identify recurrent patterns and provide explanatory context to the quantitative findings. All statistical analyses were conducted on StatsKingdom (StatsKingdom, n.d.), aStatsa (aStatsa, 2026) and GraphPad (GraphPad, 2025), at a significant level of $p = 0.05$.

RESULTS

Accuracy of Option Selection and the Validity of Rationales by the four LLMs

Among the four LLMs evaluated for their accuracy in answering pharmacology MCQs, ChatGPT-4o demonstrated the highest overall accuracy at 84%. This was followed by Google Gemini at 80%, Microsoft Copilot at 72%, and Claude at 68%.

Similarly, the validity of the pharmacological rationales provided by the four LLMs was consistent, with valid rationales observed in 33 to 36 out of 50 questions, corresponding to a validity range of 66% to 72%. ChatGPT-4o achieved the highest validity at 72%, while Claude

had the lowest at 66%. Despite these differences, statistical analysis revealed no significant differences in the overall accuracy of responses and validity of the pharmacological rationales across the four LLMs (**Table 2**).

Table 2. Performance of the LLMs measured from the accuracy of the answer (by Chi-Square Test) and the scientific validity of rationale (by Fisher-Fullman-Halton Test).

Parameters[a]	ChatGPT 4o (n = 50)	Copilot (n = 50)	Gemini (n = 50)	Claude (n = 50)	p value[b]
Accuracy of LLM's Answer					
Correct	42 (84)	36 (72)	40 (80)	34 (68)	0.223
Incorrect	8 (16)	14 (28)	10 (20)	16 (32)	
Scientific Validity of Rationale					
Correct Answer & Valid Rationale	36 (72)	34 (68)	35 (70)	33 (66)	
Correct Answer & Invalid Rationale	6 (12)	2 (4)	5 (10)	1 (2)	0.220
Incorrect Answer & Invalid Rationale	8 (16)	14 (28)	10 (20)	16 (32)	

Note. Abbreviation: Large Language Models (LLMs)

[a] Data given as n (%).

[b] All quantitative tests were conducted as two-tailed tests, with a statistical significance of p = 0.05

Post-Hoc Analysis on the Accuracy of the Option Selection

The post-hoc pairwise comparison of MCQ accuracy between the models produced varying p-values. None of the pairwise differences reached the 0.05 threshold for significance, except for the comparison between ChatGPT-4o and Claude, which yielded a p-value of 0.061 (**Table 3**).

Table 3. A post-hoc analysis on the accuracy of MCQ option selection

	ChatGPT-4o & Copilot	ChatGPT-4o & Gemini	ChatGPT-4o & Claude	Copilot & Gemini	Copilot & Claude	Gemini & Claude
Difference in Accuracy[a]	4 (8)	2 (4)	8 (16)	4 (8)	2 (4)	6 (12)
p-value	0.148	0.603	0.061	0.349	0.663	0.171

[a] Data given as n (%).

When stratifying performance by Bloom's Taxonomy levels, all models achieved 100% accuracy on 'remembering' questions, with a general trend of decreasing accuracy as the cognitive level increased. For 'understanding' questions, ChatGPT-4o and Gemini both attained 92.3% accuracy. Gemini excelled at 'applying' with an 84.6% of accuracy, whereas ChatGPT-4o outperformed the others at 'analysing', achieving a 69.2% of accuracy rate. However, these differences did not show strong evidence of variation across the models (**Table 4**).

Table 4. Stratification of questions by the Bloom's taxonomy level and the number of correctly answered questions by the Fisher-Fullman-Halon Test.

Parameters[a]	Remember (11)	Understand (13)	Applying (13)	Analysing (13)	p (GenAI) [b, c]
ChatGPT 4o	11 (100)	12 (92.3)	10 (76.9)	9 (69.2)	0.172
Copilot	11 (100)	9 (69.2)	9 (69.2)	7 (53.8)	0.068
Gemini	11 (100)	12 (92.3)	11 (84.6)	6 (46.2)	0.006*
Claude	11 (100)	9 (69.2)	6 (46.2)	8 (61.5)	0.029*
p (Bloom) [c]	-	0.257	0.081	0.784	-

* Statistical significance with $p < 0.05$

[a] Data given as n (%).

[b] All quantitative tests were conducted as two-tailed tests, with a statistical significance of $p = 0.05$.

[c] p(GenAI) represents the statistical significance of LLMs' performance in answering questions at various Bloom's Taxonomy levels. p(Bloom) indicates the statistical significance of differences in performance among various LLMs when answering questions at the same Bloom's Taxonomy level.

Further analysis within each LLM showed differences across Bloom's Taxonomy levels for Copilot ($p = 0.068$), Gemini ($p = 0.006$), and Claude ($p = 0.029$). Copilot showed a declining trend in accuracy at higher-order thinking levels, answering 69.2% of 'applying' questions correctly. Gemini performed consistently across 'remembering', 'understanding', and 'applying' questions, but demonstrated lower performance in 'analysing' questions. Claude's performance was relatively consistent from 'understanding', 'applying' and 'analysing', except for a drop in 'applying', where only 46.2% of the questions were correctly answered (**Table 4**).

Existence of Citations and the Relevance of References

Across the four LLMs, ChatGPT-4o generated the greatest number of uniform resource locators (URL): 99 unique links were cited by the LLM to justify its option selections of the 50 MCQs. This was followed by Copilot, Claude, and Gemini, in descending order (**Supplementary Table 3**). A statistically significant difference was observed in the number of unique URLs.

Despite generating fewer URLs than ChatGPT-4o, Copilot produced the greatest number of valid citations. 91.8% of the URLs cited by Copilot were links that existed. Copilot also significantly outperformed Claude (24.7%), ChatGPT-4o (13.1%), and Google Gemini (12.5%) (**Table 5A**).

Among the non-existent links, the most frequent issue encountered for ChatGPT-4o, Copilot, and Claude was the generation of invalid URLs that led to 'Page Not Found' errors, accounting for 60.5%, 71.4%, and 61.5% of the invalid links, respectively (**Table 5A**). While Gemini did not generate any broken URLs, all URLs cited by Gemini directed users to unrelated articles and books, failing to provide the scientific citations intended to support the MCQ rationales. This was also observed in the other three LLMs, albeit at a lower frequency of 29.1%, 28.6% and 32.7%, respectively. Other issues observed include directing the authors to the general menu of a publisher and using non-medical references to support the rationale for selecting an option.

Table 5. Analysis of (A) citation existence and (B) the relevance of scientific references in supporting the selected options in 50 pharmacology MCQs.

Parameters[a]	ChatGPT 4o	Copilot	Gemini	Claude
(A) Citation Existence				
Yes	13 (13.1)	78 (91.8)	4 (12.5)	17 (24.7)
No	86 (86.9)	7 (8.2)	28 (87.5)	52 (75.3)
Error Analysis				
Page Not Found	52 (60.5)	5 (71.4)	-	32 (61.5)
Directs to a normal webpage but cited as a scientific publication.	3 (3.49)	-	-	3 (5.8)
Directs to the general menu of the publisher.	6 (6.98)	-	-	17 (32.7)
Directs to a book/article that has no relevance to the citation	25 (29.1)	2 (28.6)	28 (100)	-
(B) Reference Relevance				
Yes	3 (23.1)	38 (48.7)	1 (25)	0
No	10 (76.9)	40 (51.3)	3 (75)	17 (100)
Error Analysis				
No specific mention of the reference in verbatim.	10 (100)	36 (90)	3 (100)	10 (58.8)
Found in verbatim but have non-scientific origins	-	4 (10)	-	-
Limited by Paywall	-	-	-	7 (41.2)

[a] Data given as n (%).

Among the existing URLs, the number of relevantly cited references varied, with differences observed in the p-values. Microsoft Copilot generated 48.7% of relevant scientific references. This was a notable improvement over ChatGPT-4o, Gemini, and Claude, which achieved relevance rates of 23.1%, 25%, and 0%, respectively (**Supplementary Table 4**). The predominant reason that contributed to the inability to cite the relevant responses was due to the non-specific mention of concepts that supported the references. Additionally, Copilot and Claude occasionally cited sources verbatim from non-scientific contexts and generated references behind paywalls, limiting verification (**Table 5B**).

Evaluation of Automated Question Generation Using ChatGPT-4o

The evaluation of ChatGPT-4o-generated MCQs (**Supplementary Table 5**) revealed consistently high ratings across the core criteria, with mean scores of 3.93 ± 0.28 for accuracy, 3.99 ± 0.10 for validity, and 3.96 ± 0.22 for readability, on a four-point scale. Similarly, across all four cognitive levels of Bloom's Taxonomy ('remembering', 'understanding', 'applying', and 'analysing'), ratings remained consistently strong (**Table 6**).

Accuracy ratings were consistently high across Bloom's taxonomy, though a modest decline was noted as items increased in cognitive complexity. Lower-order MCQs ('remembering' and 'understanding') achieved near-perfect accuracy, while 'applying' (3.90 ± 0.37) and 'analysing' (3.90 ± 0.31) items showed slightly reduced scores, reflecting greater variability at higher levels (**Table 6**). In contrast, validity and readability remained uniformly strong across all categories, with several domains receiving perfect ratings.

Table 6: Evaluation of automated generation of MCQs using ChatGPT-4o.

	Accuracy [a]			Validity [a]			Readability [a]		
	Rater 1	Rater 2	Mean	Rater 1	Rater 2	Mean	Rater 1	Rater 2	Mean
Overall (n = 96)	3.95±0.23	3.91±0.33	3.93±0.28	3.99±0.10	3.99±0.10	3.99±0.10	3.97±0.18	3.95±0.27	3.96±0.22
Bloom's Taxonomy									
Remembering (n = 24)	3.92±0.28	3.96±0.20	3.94±0.25	4.00±0.00	4.00±0.00	4.00±0.00	4.00±0.00	4.00±0.00	4.00±0.00
Understanding (n = 24)	4.00±0.00	3.96±0.20	3.98±0.14	4.00±0.00	4.00±0.00	4.00±0.00	3.96±0.20	3.96±0.20	3.96±0.20
Applying (n = 24)	3.92±0.28	3.88±0.45	3.90±0.37	3.96±0.20	4.00±0.00	3.98±0.14	4.00±0.00	3.92±0.28	3.96±0.20
Analysing (n = 24)	3.96±0.20	3.83±0.38	3.90±0.31	4.00±0.00	3.96±0.20	3.98±0.14	3.92±0.28	3.92±0.41	3.92±0.35

[a] Data reported as mean + standard deviation (SD).

Cohen's kappa was calculated to assess inter-rater agreement; however, the coefficients substantially underestimated concordance owing to the prevalence effect, also referred to as the 'prevalence paradox' (Zec et al., 2017). In contrast, the evaluations demonstrated a high degree of consistency between raters, with a particularly strong alignment observed for validity and readability, and similarly close agreement for accuracy (**Table 6**). As a result, the kappa values did not yield meaningful estimates, and inter-rater agreement was better reflected by the raw similarity of scores.

While the numerical ratings highlight overall quality, the qualitative remarks provided a more nuanced perspective. The thematic analysis highlighted areas for refinement. Accuracy issues were primarily related to conceptual misalignment, imprecise terminology, and ambiguous references to pharmacological processes. Readability concerns centred on unclear question phrasing and distractor wording that reduced plausibility. Finally, while validity scores were numerically high, expert feedback pointed to limitations in the depth of testing and the diagnostic value of distractors, indicating that high quantitative ratings may mask subtle but important qualitative shortcomings (**Table 7**).

Table 7. Thematic analysis of expert remarks on automated generation of MCQs on ChatGPT-4o.

Criterion	Issues	Example
Accuracy	1. Misalignment of intended concept versus tested concept	Questions intended to assess the concept of drug distribution instead focused narrowly on the parameter of volume of distribution (Vd), which belongs to the subsequent concept area in the Core Concept of Pharmacology.
	2. Incorrect or Imprecise Terminology	Key pharmacological terms (e.g., elimination vs metabolism, potency vs affinity) were used imprecisely, creating conceptual inaccuracies.
	3. Ambiguous Pharmacological References.	Some items referenced pharmacological properties (e.g., CYP metabolism, drug accumulation) without sufficient clarity, leading to premature or unsupported conclusions.

Table 7. Thematic analysis of expert remarks on automated generation of MCQs on ChatGPT-4o (*continued*).

Criterion	Issues	Example
Validity	1. Depth of Testing	Some items emphasized on surface recognition (e.g.: keyword spotting) rather than deeper reasoning about pharmacological mechanisms.
	2. Distractor Quality	Many questions included only one plausible option, while others were clearly irrelevant or exaggerated, limiting the item's diagnostic value.
Readability	1. Distractor Wording	Answer options sometimes used unrealistic absolutes (e.g., 'always', 'randomly'), which weakened their plausibility.
	1. Clarity Issues	Several questions were confusing or poorly phrased, making them difficult for learners to interpret accurately.

Overall, these findings suggest that while ChatGPT-4o can reliably generate MCQs that meet acceptable quality standards, nuanced issues of conceptual precision, clarity, and distractor design persist and warrant targeted refinement.

DISCUSSION

Adaptive learning, facilitated by LLMs, allows learners to advance individually and independently address learning gaps, mastering the educational content (Bai & Wan, 2025; Pelánek, 2025). The emergence of LLMs opens new possibilities for adaptive learning. While LLMs are readily accessible, these models require careful evaluation regarding the feasibility of adoption in discipline-based clinical domains, such as pharmacology. This study systematically evaluated the accuracy of the AI-generated responses and the validity of their scientific responses towards a set of expert-generated MCQs, and the accuracy, validity and readability of MCQs generated by the LLMs. This feasibility study assessed the reliability of four widely used LLMs - ChatGPT-4o, Microsoft Copilot, Google Gemini, and Claude.

Accuracy & Validity of the LLMs

Overall, ChatGPT-4o (84%) demonstrated higher numerical accuracy than Gemini (80%), Copilot (72%) and Claude (68%). Similar trends have been observed in previous studies, where ChatGPT-4o achieved 87.5% accuracy, followed by Copilot (62.5%) and Gemini (57.5%), with a statistically significant difference noted (Semeraro et al., 2024). The variations in LLMs performance are likely attributed to the inherent differences in artificial neural networks and model training, and model optimisation approaches (Giannakopoulos et al., 2023).

Additionally, the absence of significant differences in overall performance across the LLMs in this study contrasts with previous findings, which identified ChatGPT-4.0 and Microsoft Copilot as significantly more accurate than Claude (Rossettini et al., 2024). This discrepancy is likely attributed to the relatively small sample size (50 MCQs), which may have limited statistical power. However, pairwise comparison revealed that ChatGPT-4o achieved 84% accuracy compared to Claude's 68%, which aligns with previous findings, where ChatGPT-4o (71.1%) also outperformed Claude (61.0%) in the Critical Care Assessment in pharmacy education (Rossettini et al., 2024).

Focusing specifically on ChatGPT-4o, the overall accuracy in this study exceeded the 76% reported for both the UK Medical Licensing Assessment (Lai et al., 2023) and the Korean CBMSE (Choi, 2023), the 72.5% in the Japanese National License Examination for Pharmacists (Sato & Ogasawara, 2024) and 61.1% in USMLE (Gilson et al., 2023).

Among the questions correctly answered by the LLMs, the correct alignment between the chosen answer and the valid rationale exceeded the empirically reported rate of 19% (Choi, 2023). However, the models occasionally produced incorrect justifications. For example, some responses included miscalculations of a foundational mathematical question in pharmacology. This discrepancy indicates that LLMs may have difficulties with logical reasoning, as they lack the ability to self-correct factual or computational errors. These reasoning inconsistencies raise broader concerns about LLMs' reliability in adaptive learning. Moreover, in this study, most results fell below the 95% confidence limit (Bharatha et al., 2024; Choi, 2023), adding to the concerns about their reliability as standalone learning tools in adaptive learning (Cain et al., 2023; Choi, 2023).

Secondary Analysis Stratified by Bloom's Taxonomy and LLM

This study found no significant difference in ChatGPT's performance across different Bloom's Taxonomy levels, aligning with findings from previous research (Bharatha et al., 2024). This outcome may be attributed to ongoing model training and improvements, which have enhanced the logical flow and reasoning capabilities of LLMs. However, this was not observed in Copilot, which showed a close-to-significant difference, or in Gemini and Claude, where significant differences were found. These results align with existing limitations of LLMs. As they tend to struggle with higher-order cognitive tasks (Choi, 2023; Meo et al., 2023), it may potentially limit their role in case-based scenarios in pharmacology education. Notably, this study is among the first to evaluate the accuracy of MCQs stratified by Bloom's Taxonomy across Copilot, Gemini, and Claude. Future research using larger and more diverse question banks may provide deeper insights into this aspect of LLMs' performance.

Citation Existence and Link Relevance

A comparison of ChatGPT-4o's findings revealed that 86.9% of its citations were either irrelevant or non-existent. Among these, 61% of URLs were non-existent, a frequency comparable to the rate of 69.7% reported in previous literature. The presence of irrelevant citations, often redirecting users to the general pages of publishers and/or unrelated articles, was also observed (Choi, 2023). The challenge in synchronising facts with reliable scientific support was consistently noted across other LLMs. Gemini, which generated the fewest citations, failed entirely in the domain of reference relevance, while Copilot, although producing incorrect and irrelevant citations less frequently, still exhibited issues.

These findings highlight the presence of 'artificial hallucination', generally defined as the formulation of impeccably logical responses without a scientific grounding (Abd-Alrazaq et al., 2023; Alkaissi et al., 2023), usually associated with uncomprehensive training data, limited real-world understanding and the limitations of algorithm design (Mu & He, 2024). This was observed in a higher frequency of accurate answers and valid explanations, but an equally high frequency of irrelevant citations and references. In a complex field that demands accurate decision-making and precise scientific backing, the phenomenon poses a risk of misinformation and an apocryphal impact on scientific integrity. In particular, a lack of a

profound understanding of the subject limits the ability to discern information during early clinical and educational years (Mu & He, 2024).

ChatGPT-4o-Generation of Multiple-Choice Questions

The evaluation of ChatGPT-4o–generated MCQs demonstrated high baseline quality, with consistently strong ratings for accuracy, validity, and readability. Performance across Bloom's Taxonomy was similarly robust, though a modest decline in accuracy was observed at higher-order levels ('applying' and 'analysing'), reflecting greater variability with item complexity. These findings align with published literature, which suggests that LLMs can generate questions broadly consistent with Bloom's levels but require improvement in distinguishing between cognitive domains and ensuring closer alignment with human standards (Maity et al., 2024). This limitation is particularly evident in higher-level items, where AI-generated MCQs are more likely to include multiple correct answers, highlighting the difficulty of maintaining precision in complex question design (Doughty et al., 2024).

Beyond numerical scores, thematic analysis further revealed issues of conceptual precision, phrasing clarity, distractor plausibility, and diagnostic depth. These are also concerns reported by Law et al. (2025), who found higher rates of factual inaccuracies, irrelevance, and inappropriate difficulty in AI-generated questions compared with human-authored ones, underscoring the need for expert oversight in item refinement.

Implication in Adaptive Learning in Clinical Education

This feasibility study examines the potential of LLMs as an adaptive learning technology. LLMs personalise learning progression – by providing immediate feedback and enhancing metacognition – to enhance connectivism and efficiently guide learners towards the mastery of competencies in clinical sciences (Tacettin et al., 2021).

In our exploratory context, our findings suggest that LLMs demonstrate accuracy in lower-order cognitive tasks, making them well-suited for reinforcing foundational knowledge through knowledge checks. However, their inconsistencies in higher-order reasoning tasks and adaptive functions, as demonstrated in our findings, underscore this distinction by highlighting how conceptual precision and diagnostic depth heighten the risk of deploying these learning artefacts without expert review. This potentially limits the adoption of LLM-based adaptive learning in applied fields (Pelánek, 2025); the utilisation of LLMs in these higher-order reasoning tasks still requires domain expert oversight (Law et al., 2025; Pelánek, 2025).

LLMs can serve as a supplementary tool within the adaptive learning models. One application is in flipped classroom pedagogy, where AI-generated MCQs support pre-class assessments, enabling students to identify knowledge gaps before faculty-led discussions (Lopez-Villanueva et al., 2024). Similarly, LLMs can provide learning insights for faculty to analyse, actively address misconceptions and deliver targeted lessons (du Plooy et al., 2024; Rincon-Flores et al., 2024). This approach promotes personalised learning pathways while ensuring critical reasoning skills are reinforced by educators.

However, the prevalence of artificial hallucination underscores the need to incorporate AI literacy training in clinical curricula. Strengthened AI literacy may enhance the benefits of LLM-based adaptive learning (Yaseen et al., 2025). Educators should integrate AI literacy programs into clinical education to enhance learners' critical thinking skills, enabling them to effectively

formulate queries, critically assess AI-generated responses, and counter misinformation through evidence-based practices (Tacettin, 2021). Structured training ensures that LLM-driven adaptive learning remains scholastic and aligned with current clinical practice (Alkaissi et al., 2023; Cain et al., 2023; Mu & He, 2024).

Additionally, the 'human-in-the-loop' strategy may offer a pragmatic way to address these challenges by leveraging performance analytics and learner feedback to iteratively refine LLM-generated items (Tarun et al., 2025). This approach establishes a feedback loop between AI-generated MCQs, expert validation and learner performance, enabling adaptive learning systems to scale efficiently while evolving in response to learners' needs (Gligorea et al., 2023; Tarun et al., 2025).

While LLMs remain auxiliary tools, the hitherto advances in adaptive learning necessitate continuous evaluation and integration. Embracing AI advancements will enhance competency mastery, ensuring its effective contribution to clinical education (Bharatha et al., 2024; Cain et al., 2023).

Limitations & Future Directions

This study serves as a first step in exploring LLMs for adaptive learning. While the set of expert-generated MCQs was mapped to the core pharmacology concepts, expanding the quantity and quality of the question set – to wider subfields and greater cognitive complexity – will provide a robust and generalizable evaluation of AI's capabilities in education. Additionally, incorporating learner engagement will provide valuable insights into how students interact with LLMs and AI-generated MCQs, offering a real-world perspective on AI's role in adaptive learning and critical thinking development. Longitudinal studies examining LLMs' impact on knowledge retention, critical reasoning, and adaptive learning over time will further clarify their effectiveness as educational tools. As AI continues to evolve, structured faculty oversight and AI literacy training will be essential in equipping future pharmacists with critical appraisal skills, reinforcing AI's role as a complementary aid rather than a replacement in clinical education.

CONCLUSIONS

This study highlights the potential of LLMs as adaptive learning tools while underscoring the need for cautious integration and faculty oversight. While LLMs demonstrated high accuracy in lower-order cognitive tasks, their inconsistencies in higher-order reasoning, rationale validity, and citation reliability limit their independent use in case-based and clinical decision-making education. To ensure responsible adoption, AI literacy, critical thinking, and evidence evaluation must be incorporated into clinical education. Future research should focus on expanding question diversity, engaging student learners, and validating LLM-assisted learning models to enhance their role in adaptive learning and clinical education.

ABOUT THE AUTHORS

Cheng Keat is a Lecturer in the Diploma in Pharmaceutical Science at the School of Applied Science, Nanyang Polytechnic, Singapore. Trained as a pharmacist, he teaches pharmacotherapy, with a focus on the application of medicines in disease management for higher education learners. His research interests centre on AI-enhanced pedagogies, active learning strategies, and the design of learning experiences in higher education.

Qing Hao is currently a Medical Data Associate at Lucence, where he interprets cancer-related genetic data, integrates scientific evidence, and delivers accurate clinical reports to support data-driven oncology care. Lucence is a Singapore-based biotechnology company advancing early cancer detection and personalized treatment through blood- and tissue-based genomic testing, using advanced sequencing technologies to improve clinical outcomes. Qing Hao holds a degree in Biological Sciences from Nanyang Technological University, where he also participated in research on the multiphase separation of TDP-43.

Joseph Tan, BSc (Hons) Pharmacy, is an Account Manager at GlaxoSmithKline, Singapore. He graduated with Distinction from the National University of Singapore and practised as a registered pharmacist in Tan Tock Seng Hospital, Singapore. His research interests include healthcare education, evidence-based practice, and approaches to enhancing learning and decision-making in clinical and academic settings.

Annie Ng is a Senior Specialist in Teaching and Learning and a Senior Lecturer at the School of Applied Science, Nanyang Polytechnic, Singapore. Her scholarly work focuses on the scholarship of teaching and learning, with particular interest in authentic assessment, technology-supported pedagogy, and evidence-based curriculum design in applied and health sciences education. She adopts mixed-methods research approaches to examine student learning, self-directed learning, and professional readiness.

REFERENCES

Abd Alrazaq, A., AlSaad, R., Alhuwail, D., et al. (2023). Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Medical Education, 9,* Article e48291. https://doi.org/10.2196/48291

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), Article e35179. https://doi.org/10.7759/cureus.35179

Anthropic. (2024, June 21). *Claude 3.5 Sonnet*. https://www.anthropic.com/news/claude-3-5-sonnet

aStatsa. (2016). Online statistics calculators. *aStatsa*. https://astatsa.com/

Armstrong, P. (2010). *Bloom's taxonomy*. Vanderbilt University Center for Teaching. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/

Bharatha, A., Ojeh, N., Fazle Rabbi, A. M., et al. (2024). Comparing the performance of ChatGPT 4 and medical students on MCQs at varied levels of Bloom's taxonomy. *Advances in Medical Education and Practice*, *15*, 393–400. https://doi.org/10.2147/AMEP.S457408

Cain, J., Malcom, D. R., & Aungst, T. D. (2023). The role of artificial intelligence in the future of pharmacy education. *American Journal of Pharmaceutical Education*, *87*(10), Article 100135. https://doi.org/10.1016/j.ajpe.2023.100135

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., et al. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study. *PLOS ONE*, *18*(8), Article e0290691. https://doi.org/10.1371/journal.pone.0290691

Choi, W. (2023). Assessment of the capacity of ChatGPT as a self learning tool in medical pharmacology. *BMC Medical Education*, *23*(1), Article 864. https://doi.org/10.1186/s12909-023-04832-x

Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education*, *8*, Article 1206936. https://doi.org/10.3389/feduc.2023.1206936

Doughty, J., Wan, Z., Bompelli, A., et al. (2024). A comparative study of AI generated (GPT 4) and human crafted MCQs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference (ACE 2024)*. ACM. https://doi.org/10.1145/3636243.3636256

du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon*, *10*(21), Article e39630. https://doi.org/10.1016/j.heliyon.2024.e39630

Giannakopoulos, K., Kavadella, A., Salim, A. A., Stamatopoulos, V., & Kaklamanos, E. G. (2023). Evaluation of the performance of generative AI large language models in dentistry. *Journal of Medical Internet Research, 25,* Article e51580. https://doi.org/10.2196/51580

Gilson, A., Safranek, C. W., Huang, T., et al. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education, 9,* Article e45312. https://doi.org/10.2196/45312

Gligorea, I., Cioca, M., Oancea, R., et al. (2023). Adaptive learning using artificial intelligence in e learning: A literature review. *Education Sciences, 13*(12), Article 1216. https://doi.org/10.3390/educsci13121216

GraphPad. (2025). *Quantify interrater agreement with kappa.* GraphPad QuickCalcs. https://www.graphpad.com/quickcalcs/kappa1/

Guilding, C., White, P. J., Cunningham, M., et al. (2024). Defining and unpacking the core concepts of pharmacology. *British Journal of Pharmacology, 181*(3), 375–392. https://doi.org/10.1111/bph.16222

Lai, U. H., Wu, K. S., Hsu, T. Y., & Kan, J. K. C. (2023). Evaluating the performance of ChatGPT 4 on the United Kingdom Medical Licensing Assessment. *Frontiers in Medicine, 10,* Article 1240915. https://doi.org/10.3389/fmed.2023.1240915

Law, A. K., So, J., Lui, C. T., et al. (2025). AI versus human generated MCQs in medical education. *BMC Medical Education, 25*(1), Article 208. https://doi.org/10.1186/s12909-025-06796-6

Liévin, V., Hother, C. E., & Winther, O. (2023). Can large language models reason about medical questions? *arXiv.* https://doi.org/10.48550/arXiv.2207.08143

López Villanueva, D., Santiago, R., & Palau, R. (2024). *Flipped learning and artificial intelligence. Electronics, 13*(17), Article 3424. https://doi.org/10.3390/electronics13173424

Maity, S., Deroy, A., & Sarkar, S. (2024). How effective is GPT 4 Turbo in generating school level questions? *arXiv*. https://doi.org/10.48550/arXiv.2406.15211

Meo, S. A., Al Masri, A. A., Alotaibi, M., Meo, M. Z. S., & Meo, M. O. S. (2023). ChatGPT knowledge evaluation in medical sciences. *Healthcare, 11*(14), Article 2046. https://doi.org/10.3390/healthcare11142046

Microsoft. (2024, October 1). *An AI companion for everyone*. https://blogs.microsoft.com/blog/2024/10/01/an-ai-companion-for-everyone/

Moore, S., Costello, E., Nguyen, H. A., & Stamper, J. (2024). An automatic question usability evaluation toolkit. *arXiv.* https://doi.org/10.48550/arXiv.2405.20529

Mu, Y., & He, D. (2024). The potential applications and challenges of ChatGPT in medicine. *International Journal of General Medicine, 17,* 817–826. https://doi.org/10.2147/IJGM.S456659

Nori, H., King, N., McKinney, S. M., et al. (2023). Capabilities of GPT 4 on medical challenge problems. *arXiv*. https://doi.org/10.48550/arXiv.2303.13375

Olney, A. M., Gilbert, S. B., & Rivers, K. (2022). Preface to the special issue on creating and improving adaptive learning: Smart authoring tools and processes. *International Journal of Artificial Intelligence in Education, 32*, 1–3. https://doi.org/10.1007/s40593-021-00277-9

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

OpenAI. (2024, May 13). Introducing GPT 4o. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

Pelánek, R. (2025). Adaptive learning is hard: Challenges, nuances, and trade-offs in modeling. *International Journal of Artificial Intelligence in Education, 35*, 304–329. https://doi.org/10.1007/s40593-024-00400-6

Pichai, S., & Hassabis, D. (2024, February 15). Our next generation model: Gemini 1.5. *Google.* https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

Rincon-Flores, E. G., Castano, L., Guerrero Solis, S. L., & García-Valcárcel Muñoz-Repiso, A. (2024). Improving the learning-teaching process through adaptive learning strategy. *Smart Learning Environments, 11*(1), Article 27. https://doi.org/10.1186/s40561-024-00314-9

Rossettini, G., Bargeri, S., Cook, C., Guida, S., Palese, A., Rodeghiero, L., Pillastrini, P., Turolla, A., Castellini, G., & Gianola, S. (2025) Accuracy of ChatGPT-3.5, ChatGPT-4o, Copilot, Gemini, Claude, and Perplexity in advising on lumbosacral radicular pain against clinical practice guidelines: Cross-sectional study. *Frontiers in Digital Health, 7*, Article 1574287. https://doi: 10.3389/fdgth.2025.1574287.

Sato, H., & Ogasawara, K. (2024). ChatGPT (GPT-4) passed the Japanese National License Examination for Pharmacists in 2022, answering all items including those with diagrams: A descriptive study. *Journal of Educational Evaluation for Health Professions, 21*, Article 4. https://doi.org/10.3352/jeehp.2024.21.4

Semeraro, F., Gamberini, L., Carmona, F., & Monsieurs, K. G. (2024). Clinical questions on advanced life support answered by artificial intelligence. A comparison between ChatGPT, Google Bard and Microsoft Copilot. *Resuscitation, 195*, Article 110114. https://doi.org/10.1016/j.resuscitation.2024.110114

Singapore Ministry of Health. (2023). *Guide to competency assessment exam 2023.* https://www.healthprofessionals.gov.sg/docs/librariesprovider3/standards-exams/guide-to-competency-assessment-exam-2023.pdf

Statistics Kingdom. (n.d.). Statistics calculator. *Statistics Kingdom*. https://www.statskingdom.com/

Suleyman, M. (2024, October 1). *An AI companion for everyone*. Official Microsoft Blog. https://blogs.microsoft.com/blog/2024/10/01/an-ai-companion-for-everyone/

Tacettin, C. M. (2021). Competency-based education: Theory and practice. *Psycho Educational Research Reviews, 10*(3), 67–95. https://doi.org/10.52963/PERR_Biruni_V10.N3.06

Tackett, S., Raymond, M., Desai, R., et al. (2018). Crowdsourcing for assessment items to support adaptive learning. *Medical Teacher, 40*(7), 838–841. https://doi.org/10.1080/0142159X.2018.1490704

Tarun, B., Du, H., Kannan, D., & Gehringer, E. F. (2025). Human in the loop systems for adaptive learning. *arXiv*. https://doi.org/10.48550/arXiv.2508.11062

Wang, Y. M., Shen, H. W., Chen, T. J., Chiang, S. C., & Lin, T. G. (2025). Performance of ChatGPT-3.5 and ChatGPT-4 in the Taiwan National Pharmacist Licensing Examination. *JMIR Medical Education, 11,* Article e56850. https://doi.org/10.2196/56850

Yaacoub, A., Da Rugna, J., & Assaghir, Z. (2025). Assessing AI-generated questions' alignment with cognitive frameworks in educational assessment. *International Journal of Computer Theory and Engineering, 17*(3), 114–125. https://doi.org/10.7763/ijcte.2025.v17.1374

Yaseen, H., Mohammad, A. S., Ashal, N., Abusaimeh, H., Ali, A., & Sharabati, A.-A. A. (2025). The Impact of adaptive learning technologies, personalized feedback, and interactive AI tools on student engagement: The moderating role of digital literacy. *Sustainability, 17*(3), Article 1133. https://doi.org/10.3390/su17031133

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: The paradox of Cohen's kappa. *The Open Nursing Journal, 11*, 211–218. https://doi.org/10.2174/1874434601711010211. ■

SUPPLEMENTARY INFORMATION

Supplementary Table 1. Example of questions in the question bank, stratified based on Bloom's taxonomy.

Bloom's Taxonomy	Example of Questions
Remembering	What is the primary site in the gastrointestinal tract where most oral drug absorption occurs? A. Stomach B. Small intestine C. Large intestine D. Esophagus
Understanding	Which of the following **BEST** describes how a competitive antagonist affects an agonist's dose-response curve? A. Shifts the curve to the left without changing the maximum response. B. Shifts the curve to the right without changing the maximum response. C. Decreases the slope of the curve without altering the maximum response. D. Reduces the maximum response without affecting the curve's position.
Applying	Memantine is a drug used to slow down the progression of dementia. Mechanistically, the drug is a non-competitive antagonist to the NMDA receptor. By preventing glutamate, the endogenous ligand, from binding onto the receptor, it helps to reduce neuronal damage to slow the progression of dementia. Which of the following statements are **TRUE** regarding the relationship between memantine and glutamate? A. Both memantine and glutamate bind to the same ligand-binding site of NMDA receptor. B. Memantine reduces the physiological effect of the NMDA receptor to its constitutive receptor activity. C. Memantine binds to the alternative site of the NMDA receptor to prevent activation by glutamate. D. Increase in glutamate synthesis will overcome memantine's antagonism.
Analysing	Lithium is a drug used in the management of bipolar disorder. The following shows the pharmacokinetic profile of lithium • Absorption: Rapid and complete • Bioavailability: 80 – 100 • Protein Binding: 0 (not known to bind to carrier protein) • Metabolism: Not metabolized. • Excretion: Urine (Primary Route) with 80 of filtered lithium reabsorbed via the proximal convoluted tubules of the nephrons. Despite the favorable pharmacokinetic profile, lithium is a drug with narrow therapeutic index with a high risk of lithium toxicity as the concentration of lithium increases in the plasma. Based on the information provided, which of the following statement(s) about Lithium is **FALSE**? A. There is a need to reduce the dose of lithium in elderly patients with a reduced creatinine clearance to mitigate the risk of lithium toxicity. B. There is a need to discontinue the lithium in patients with acute kidney failure to mitigate the risk of lithium toxicity. C. An increase in albumin concentration will lower the fraction unbound of lithium, lowering the amount of drug available for drug elimination. D. When a patient who is receiving lithium therapy is consistently dehydrated, he/she is at higher risk of developing lithium toxicity.

Supplementary Table 2. Standardised Prompts and Four-Point Rubrics for 'Accuracy', 'Validity' and 'Readability' for the Designing and Evaluation of Automated Question Generation using ChatGPT-4o.

Input / Output	**Prompts / Responses**	
User Input 1	You are a pharmacology learner, learning basic pharmacology concepts for the first time. For the concept of ***[insert Core Concept of Pharmacology]***, generate educational content at the ***[insert Bloom's level]*** level of Bloom's Taxonomy. Please follow the structure below: Question: One multiple-choice question (MCQ) appropriate to this Bloom's level, with 1 correct answer and 3 plausible distractors. Explanation: A short, concise explanation of the answers to the question in bullet points. No analogies, no storytelling, no disclaimers.	
LLM Output 1	*(LLM's response)*	
***Rubrics for Evaluation of Automated Question Generation using ChatGPT-4o*.**		
Dimension	**Definition**	**Descriptors**
Accuracy (Content Fidelity)	The degree to which the content is factually correct, scientifically sound, and pharmacologically precise.	1 = Major factual errors/unsafe 2 = Several inaccuracies/omissions 3 = Mostly correct, minor issues 4 = Fully correct & precise
Validity (Bloom's Alignment / Construct Representation)	The degree to which the output matches the intended Bloom's cognitive level (Remember, Understand, Apply, Analyze).	1 = Misaligned 2 = Weak alignment 3 = Mostly valid, minor drift 4 = Strong validity, fully aligned
Readability (Clarity & Learner Appropriateness)	The output is clear, concise, free of ambiguity/technical flaws, and phrased at a level appropriate for the target learner.	1 = Very poor – confusing, jargon-heavy 2 = Weak – awkward or ambiguous 3 = Good – generally clear, minor issues 4 = Excellent – very clear, concise, well-structured

Supplementary Table 3. Statistical Analysis of Citation Existence Among URLs Generated by Four Large Language Models by Chi-Square Test

Parameters[a]	ChatGPT 4o (n = 99)	Copilot (n = 85)	Gemini (n = 32)	Claude (n = 69)	p value[b]
Citation Existence					
Exist	13 (13.1)	78 (91.8)	4 (12.5)	17 (24.6)	< 0.001*
Non-Exist	86 (86.9)	7 (8.2)	28 (87.5)	52 75.4)	

* Statistical significance with $p < 0.05$

[a] Data given as n (%).

[b] All quantitative tests were conducted as two-tailed tests, with a statistical significance of $p = 0.05$.

Supplementary Table 4: Statistical Analysis of Reference Relevance Among URLs Generated by Four Large Language Models by Fisher-Fullman-Halton Test

Parameters[a]	ChatGPT 4o (n = 13)	Copilot (n = 78)	Gemini (n = 4)	Claude (n = 17)	p value[b]
Reference Relevance					
Relevant	3 (23.1)	38 (48.7)	1 (25)	0 (0)	0.0002*
Irrelevant	10 (76.9)	40 (51.3)	3 (75)	17 (100)	

* Statistical significance with $p < 0.05$

[a] Data given as n (%).

[b] All quantitative tests were conducted as two-tailed tests, with a statistical significance of $p = 0.05$.

Supplementary Table 5: Examples of Automated Multiple-Choice Question Generation by ChatGPT-4o, Stratified According to Bloom's Taxonomy

Core Concepts: Drug-Target Interactions describes the different ways a drug interacts with a target to produce a biological effect.	
Bloom's Taxonomy	**Question**
Remembering	**What is the term for the specific molecule that a drug binds to produce its effect?** A. Substrate B. Hormone C. Drug target D. Antibody **Correct Answer**: C. Drug target
Understanding	**Which statement best describes the nature of drug-target interactions?** A. Drugs change their structure to match any target B. Drugs interact randomly with all cell components C. Drugs bind selectively to specific targets to exert effects D. Drugs always permanently bind to their targets **Correct Answer**: C. Drugs bind selectively to specific targets to exert effects
Applying	**A drug binds reversibly to a receptor and reduces the effect of the body's natural ligand without triggering a response itself. Which type of interaction is this an example of?** A. Irreversible agonist binding B. Competitive antagonist binding C. Allosteric activation D. Non-specific protein binding **Correct Answer**: B. Competitive antagonist binding
Analysing	**Two drugs bind to the same receptor: Drug X activates the receptor, while Drug Y binds to the same site but prevents activation. What does this indicate about Drug Y's interaction compared to Drug X?** A. Drug Y is a non-competitive agonist B. Drug Y is a full agonist with delayed onset C. Drug Y is a competitive antagonist D. Drug Y enhances Drug X's receptor activity **Correct Answer**: C. Drug Y is a competitive antagonist