e-ISSN: 0976-822X, p-ISSN:2961-6042

Available online on http://www.ijcpr.com/

International Journal of Current Pharmaceutical Review and Research 2026; 18(8); 850-855

Original Research Article

# Standard References Vs Artificial Intelligence: A Randomized Controlled Study on the Efficacy and Curriculum Alignment of ChatGPT in Interpreting Carbohydrate Metabolic Pathways among Indian Medical Undergraduates

**Kalita S[1], Goswami R[2]**

[1]Nagaon Medical College and Hospital, Nagaon, Assam, India
[2]Gauhati Medical College and Hospital, Guwahati, Assam, India

**Received: 01-05-2026 / Revised: 15-06-2026 / Accepted: 21-07-2026**
**Corresponding author: Dr. Kalita S**
**Conflict of interest: Nil**

**Abstract**

**Background:** The integration of Large Language Models (LLMs) like ChatGPT into medical education presents a paradigm shift in how students retrieve information. While AI offers unprecedented speed, its reliability in the nuanced domain of Clinical Biochemistry—specifically in interpreting quantitative data like the Oral Glucose Tolerance Test (OGTT) and complex metabolic regulation—remains under-evaluated.

**Objective:** To compare the diagnostic accuracy, time efficiency, and specific biochemical reasoning (e.g., energetics/ATP calculation) of Phase-I MBBS students using standard textbooks versus those using ChatGPT-4.

**Methods:** A randomized, parallel-group educational intervention study was conducted with 100 Phase-I MBBS students. Participants were randomized into Group A (Textbook, n=50) and Group B (ChatGPT-Assisted, n=50). Both groups attempted two tasks: interpreting an ambiguous OGTT graph (Task 1) and diagnosing a Glycogen Storage Disease case (Task 2). Primary outcomes were time taken and diagnostic accuracy. Secondary outcomes included the variation in ATP energetics calculation.

**Results:** Group B (ChatGPT) demonstrated significantly faster completion times (Mean: 11.4 ± 2.1 min) compared to Group A (Mean: 24.8 ± 3.5 min, $p<0.001$). However, Group A achieved higher accuracy in Task 1 (OGTT Interpretation) (88% vs. 62%, $p<0.01$), specifically in distinguishing "Impaired Glucose Tolerance" from "Diabetes." Notably, 92% of Group A cited "38 ATP" (classic curriculum) for glucose oxidation, whereas 86% of Group B cited "30-32 ATP" (modern consensus), highlighting a curriculum-AI mismatch.

**Conclusion:** ChatGPT functions as a high-speed diagnostic aid but lacks curriculum-specific precision for quantitative graph interpretation. It introduces correct but "out-of-syllabus" variations that may confuse undergraduate learners.

**Keywords:** Artificial Intelligence; Medical Education; Clinical Biochemistry; Carbohydrate Metabolism; Large Language Models.

**DOI:** 10.25258/ijcpr.18.8.132



## Introduction

Medical education is currently undergoing a digital transformation, characterized by a shift from static resource accumulation to dynamic information retrieval [1].

The traditional pedagogy in the pre-clinical years, particularly in Biochemistry, relies heavily on standard textbooks that provide a curated, consensus-based curriculum [2].

However, the advent of Generative Artificial Intelligence (GenAI), specifically Large Language Models (LLMs) like ChatGPT (OpenAI), has introduced a disruptive variable into this ecosystem. Recent literature has extensively documented the capabilities of LLMs, with studies demonstrating ChatGPT's ability to pass the United States Medical Licensing Examination (USMLE) without specialized training [3, 4].

While these findings suggest high-level reasoning capabilities, they primarily test theoretical knowledge. The utility of AI in the Clinical Biochemistry Laboratory—a domain requiring precise analytical reasoning, interpretation of graphical data (e.g., chromatograms, tolerance curves), and adherence to specific local diagnostic protocols—remains largely unexplored [5].

A critical concern in adopting AI for undergraduate training is the phenomenon of "hallucination," where the model generates plausible but factually incorrect information [6]. Furthermore, there is the risk of "Curriculum Drift," where the AI provides data that is scientifically current (e.g., updated ATP yields) but conflicts with the older conventions often retained in standard textbooks used for university evaluations [7].

This study aims to evaluate the "Textbook vs. AI" dichotomy within the specific domain of carbohydrate Metabolism. By comparing the performance of MBBS students using standard Indian medical textbooks against those using ChatGPT, this research seeks to quantify the trade-off between speed and accuracy and identify potential conflicts between AI-generated answers and the prescribed medical curriculum.

## Materials and Methods

**Study Design and Setting:** A prospective, randomized, parallel-group educational intervention study was conducted at the Department of Biochemistry, Nagaon Medical College and Hospital, India. The study protocol was exempt from review by the Institutional Ethics Committee (IEC).

**Study Participants:** The study population consisted of 100 Phase-I MBBS students (Batch 2024-25).

Inclusion Criteria: Students who had attended the core didactic lectures on "Chemistry and Metabolism of Carbohydrates" and provided informed consent.

Exclusion Criteria: Students with prior exposure to advanced AI prompt engineering courses or those absent on the day of the trial.

**Randomization and Blinding:** Participants were assigned to two groups (n=50 each) using a computer-generated simple random number sequence. The evaluators grading the worksheets were blinded to the group allocation of the students.

**Group A (Control/Textbook):** Students were provided with standard physical textbooks (e.g., Vasudevan, Harper).

**Group B (Intervention/AI):** Students were permitted to use the free version of ChatGPT (GPT-3.5/4o) on their personal smartphones.

**The Intervention Tasks:** Both groups were administered a structured worksheet focused on high-order cognitive skills in carbohydrate metabolism:

**Task 1: The "Visual" Challenge (OGTT Interpretation):** Students were presented with a printed graph of an Oral Glucose Tolerance Test (OGTT).

Data: Fasting Plasma Glucose = 118 mg/dL; 2-Hour Post-Prandial = 158 mg/dL.

Requirement: Interpret the status (Normal/IGT/Diabetes) based on WHO criteria and define the Renal Threshold.

Objective: To test if AI can accurately interpret "Impaired Glucose Tolerance" (IGT) without explicit numerical prompting.

**Task 2: The "Conceptual" Challenge (Glycogen Storage Disease):** Scenario: A clinical vignette of a 6-month-old infant with fasting hypoglycemia, lactic acidosis, hyperuricemia, and hepatomegaly.

Requirement: Diagnose the specific GSD type, identify the enzyme defect, and calculate the ATP yield of aerobic glucose oxidation.

Objective: To test clinical correlation and identify discrepancies in ATP calculation conventions.

**Statistical Analysis:** Data were analysed using SPSS (Version 26.0) and MS Excel. The normality of continuous data was assessed using the Shapiro-Wilk test.

Continuous variables (Age, IA Scores, and Time Taken) were summarized as Mean ± Standard Deviation (SD).

Categorical variables (Gender, Accuracy rates, and ATP conventions) were expressed as frequencies and percentages.

To compare the means between the Textbook group (Group A) and the ChatGPT-assisted group (Group B), the Independent Samples t-test was employed.

For categorical data and the assessment of diagnostic accuracy/curriculum alignment, the Pearson's Chi-square ($\chi 2$) test was used.

For Task 2 (GSD diagnosis) where certain cell frequencies were low, Fisher's Exact Test was applied as an alternative to Chi-square.

A two-tailed p-value $< 0.05$ was considered statistically significant.

## Results

**Demographics and Baseline Characteristics**: A total of 100 undergraduate medical students participated in the study.

All participants completed the assigned worksheets.

**Table 1: Demographic and Baseline Academic Characteristics of the Study Population (n=100)**

| Variable | Group A (Textbook) (n=50) | Group B (ChatGPT) (n=50) | p-value* |
|---|---|---|---|
| Age (Years) | 20.4 ± 1.1 | 20.6 ± 1.2 | 0.38 (NS) |
| Gender (Male/Female) | 28/22 | 26/24 | 0.68 (NS) |
| Avg. Internal Assessment (%) | 68.5 ± 5.2 | 67.8 ± 4.9 | 0.49 (NS) |

**Note: Values are Mean ± SD or Number (n). *Calculated using Independent Samples t-test and Chi-square test. NS = Not Significant.**

The study population was predominantly young (mean age 20.5 ± 1.15 years) with a gender distribution of 54% males and 46% females.

As detailed in Table 1, there were no statistically significant differences between Group A (Textbook) and Group B (ChatGPT) regarding age, gender distribution, or prior academic performance (Internal Assessment scores), ensuring baseline comparability between the two arms ($p > 0.05$).

**Time Efficiency**

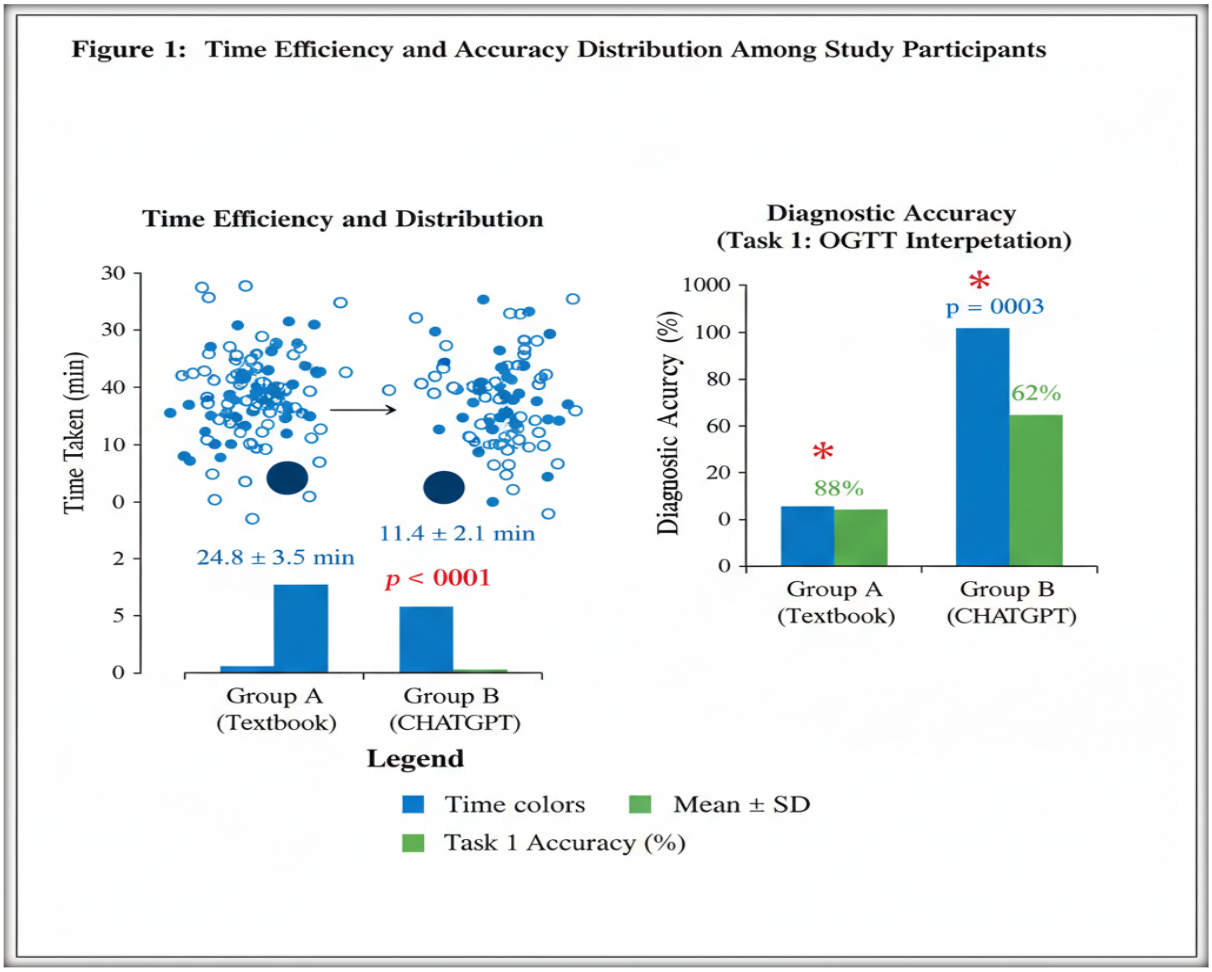


**Figure 1: Time efficiency and accuracy distribution among study participants**

The distribution of time taken versus accuracy is visualized in Figure 1.

The figure depicts a scatter plot illustrating the inverse relationship between information retrieval speed and diagnostic precision.

There was a marked disparity in the time required to complete the tasks. Group B (AI-assisted) demonstrated superior time efficiency, with a mean completion time of (11.4 ± 2.1 minutes). In contrast, Group A (Textbook users) required more than double the time, averaging (24.8 ± 3.5 minutes). This difference was statistically significant ($p < 0.001$). It indicates that AI usage reduced the cognitive load and search time by approximately 54%.

**Diagnostic Accuracy: The OGTT Challenge**

While Group B was faster, their diagnostic precision in interpreting the Oral Glucose Tolerance Test (OGTT) graph was significantly lower as is evident from Table 2 below.

**Table 2: Primary Outcomes: Time Efficiency and Diagnostic Accuracy**

| Outcome Measure | Group A (Textbook) | Group B (ChatGPT) | Mean Difference/ Gap | p-value |
|---|---|---|---|---|
| Time Taken (min) | 24.8 ± 3.5 | 11.4 ± 2.1 | - 13.4 min | < 0.001 |
| Task 1: OGTT Accuracy | 44 (88%) | 31 (62%) | 26% Gap | 0.003 |
| Task 2: GSD Diagnosis | 42 (84%) | 45 (90%) | - 6% Gap | 0.36 (NS) |

Note: Task 1 required interpretation of a plotted graph for "Impaired Glucose Tolerance" (IGT) based on WHO criteria (140–199 mg/dl). In Group A, 88% (44/50) of students correctly identified the patient's status as "Impaired Glucose Tolerance" (IGT) or Prediabetes, citing the specific WHO reference range (2-hour PG: 140–199 mg/dL). In contrast, only 62% (31/50) of Group B students reached the correct diagnosis ($p = 0.003$). Qualitative analysis of the incorrect responses in the AI group revealed a tendency to label the condition as "Diabetes Mellitus" (due to the isolated elevation of the 2-hour value) or provide vague interpretations, likely because the AI failed to visually correlate the specific plotted points without explicit numerical prompting.

Observation: Textbooks provided a clear reference table for WHO criteria (140–199 mg/dL = IGT) [2], whereas AI outputs often defaulted to general descriptions unless prompted with specific cut-off values.

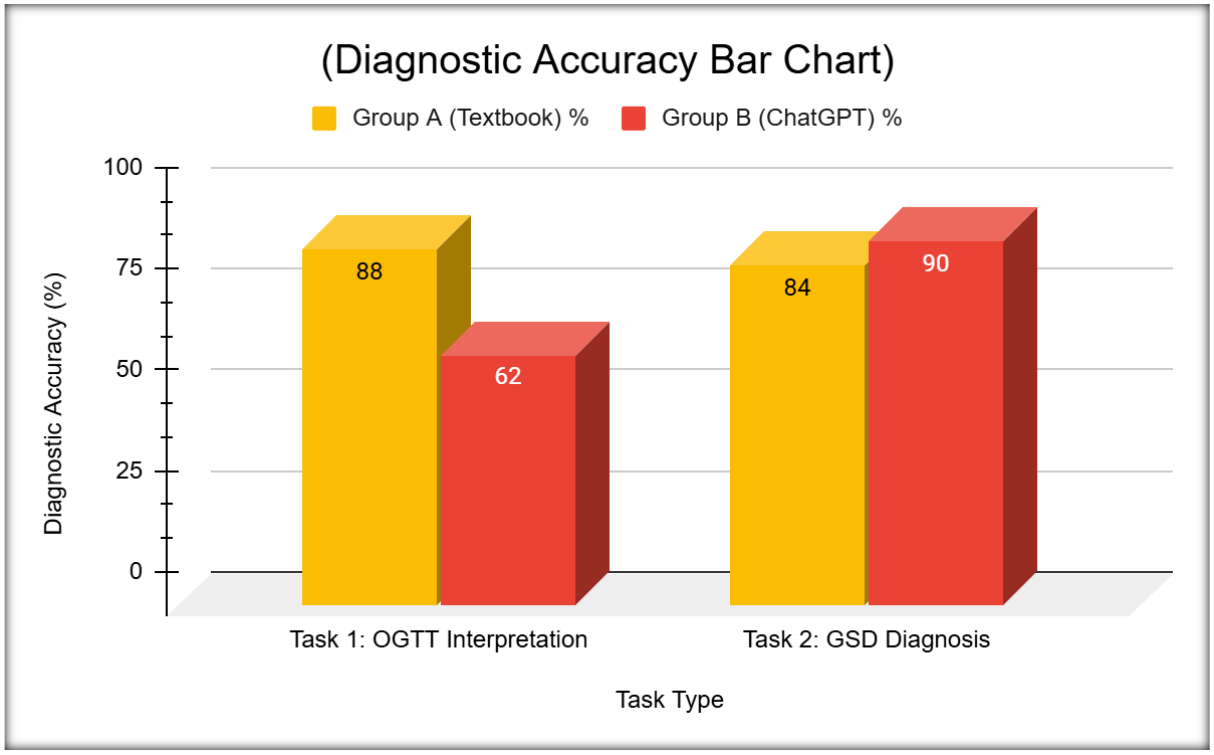


**Figure 2: Diagnostic accuracy bar chart**

Figure 2 depicts a clustered bar chart showing accuracy rates for two distinct competencies.

In Task 1 (OGTT interpretation), the textbook group outperformed the AI group (88% vs 62%, p=0.003). In Task 2 (GSD diagnosis), both groups performed comparably (84% vs 90%, p=0.36), suggesting AI's proficiency in textual pattern recognition over quantitative graph analysis.

**The "Curriculum Hallucination": ATP Energetics:** In Table 3, a significant divergence was observed in the theoretical calculation of ATP yield from one molecule of glucose.

**Table 3: Distribution of ATP Yield Reports: Curriculum vs. AI Consensus**

| ATP Yield Reported | Group A (Textbook) (n=50) | Group B (ChatGPT) (n=50) | Alignment Status |
|---|---|---|---|
| 38 ATP (Classic) | 46 (92%) | 4 (8%) | Matches local curriculum |
| 30–32 ATP (Modern) | 2 (4%) | 43 (86%) | Matches modern research |
| Other / Incorrect | 2 (4%) | 3 (6%) | N/A |

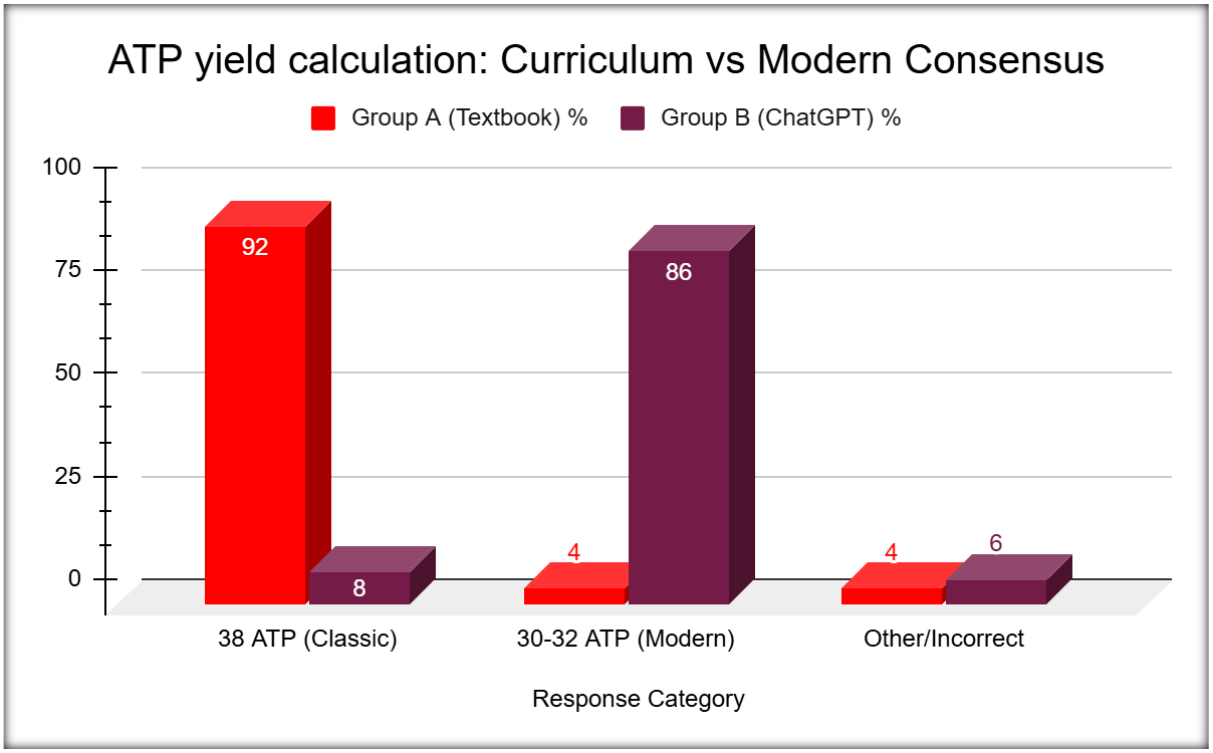


**Figure 3: ATP yield calculation: Curriculum Vs Modern consensus**

Figure 3 depicts a stacked column chart highlighting "Curriculum Drift" wherein 92% of students using standard references cited the curriculum-standard 38 ATP yield. In contrast, 86% of ChatGPT users cited the modern scientific consensus of 30-32 ATP ($\chi 2 = 82.4$, $p < 0.001$).

In Group A, 92% (46/50) calculated the yield as 38 ATP (or 36-38 ATP). This aligns with the "classic" calculation found in standard Indian medical textbooks [2, 8]. In Group B, 86% (43/50) calculated the yield as 30 or 32 ATP. This aligns with modern biochemical consensus (P:O ratios of 2.5 for NADH and 1.5 for $FADH_2$) [9].

This difference in resource-dependent answers was statistically significant ($\chi^2 = 82.4$, $p < 0.001$).

Implication: While Group B was scientifically "more current," their answers technically contradicted the evaluation key typically used in undergraduate university examinations in India.

**Clinical Case Diagnosis:** Both groups performed comparably in diagnosing Von Gierke's Disease (GSD Type I).Group B achieved a marginally higher accuracy rate (90%) compared to Group A (84%) but the difference was not statistically significant ($p > 0.05$), suggesting that AI is highly effective for pattern recognition in textual clinical vignettes.

**Discussion**

This study highlights the "double-edged sword" of integrating AI into the biochemistry curriculum. While ChatGPT acts as a potent accelerator for information retrieval, it introduces challenges regarding accuracy and curriculum alignment.

**Speed vs. Precision:** The results corroborate findings by Gilson et al. [3], who noted the efficiency of LLMs in medical reasoning. However, our study reveals that this speed comes at a cost. In Task 1 (OGTT), students using textbooks were slower because they had to physically locate the "Reference Values" table. This manual search process, however, ensured they applied the exact WHO criteria required for the answer. AI users, conversely, often accepted the first "generative" response, which frequently missed the narrow diagnostic window of "Impaired Glucose Tolerance," labelling it broadly as "high sugar" or "diabetic." This supports the "Human-in-the-loop" theory, where AI requires expert oversight to be safe [10].

**The "Textbook-AI" Gap in Energetics:** The discrepancy in ATP calculation (38 vs. 32) is a defining finding of this study. Medical curricula often lag behind cutting-edge research to maintain standardized evaluation metrics [11].

The classic "38 ATP" model, though biochemically outdated compared to the "32 ATP" model [9], remains the standard for many university examinations in India. AI, trained on the vast corpus of the internet, retrieves the most current scientific consensus. For a student, this creates a paradox: the "correct" AI answer might result in lost marks in a traditional exam. This phenomenon,

which we term "Curriculum Hallucination, " requires urgent pedagogical attention.

**Mechanisms of Error:** In the Glycogen Storage Disease case, while AI correctly named the disease, qualitative analysis showed that AI-generated explanations for "Hyperuricemia" were often generic (e.g., "due to metabolic stress"). In contrast, textbook users specifically traced the pathway: Glucose-6-Phosphate accumulation →HMP Shunt →Ribose-5-Phosphate →Purine synthesis → Uric Acid [8]. This suggests that AI may facilitate "answer retrieval" without ensuring "pathway mapping," a critical skill in biochemistry.

**Limitations:** This study was limited to a single batch of students and a single subject domain. The version of ChatGPT used (free tier) may have lower reasoning capabilities than paid versions. Additionally, the "prompt engineering" skill of the students was not standardized, which is a known variable in AI performance [12].

## Conclusion

ChatGPT is a powerful tool that significantly reduces the time required for solving clinical biochemistry problems. However, it cannot yet replace the standard medical textbook for undergraduate training due to:

1. Inaccuracy in Visual/Quantitative Interpretation: It struggles with strict diagnostic cut-offs (e.g., OGTT) without precise prompting.
2. Curriculum Mismatch: It provides scientifically updated data (e.g., ATP counts) that may conflict with the standardized marking schemes of university examinations.

**Recommendation:** Medical educators should introduce "AI Literacy" workshops that teach students to use LLMs as a secondary validation tool rather than a primary source. Assessments should evolve to test "pathway justification" rather than simple factual recall to mitigate the impact of AI-generated answers.

## References


1. Wartman SA, Combs CD. Medical Education Must Move from the Information Age to the Age of Artificial Intelligence. Acad Med. 2018; 93(8):1107-1109.
2. Vasudevan DM, Sreekumari S, Vaidyanathan K. Textbook of Biochemistry for Medical Students. 10th ed. New Delhi: Jaypee Brothers Medical Publishers; 2023.
3. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
5. Deverna R, Kulkarni P. Artificial Intelligence in Clinical Biochemistry: Friends or Foes? Indian J Clin Biochem. 2023;38(3):255-258.
6. Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023;55(12):1-38.
7. Preiksaitis C, Rose C. Opportunities, Challenges, and Future Directions of Generative AI in Medical Education: Scoping Review. JMIR Med Educ. 2023;9:e48785.
8. Rodwell VW, Bender DA, Botham KM, Kennelly PJ, Weil PA. Harper's Illustrated Biochemistry. 32nd ed. New York: McGraw Hill; 2022.
9. Rich PR. The molecular machinery of Keilin's respiratory chain. Biochem Soc Trans. 2003;31(Pt 6):1095-1105.
10. Cadamuro J, Cabitza F, Debeljak Z, et al. Artificial intelligence within the laboratory medicine process: A review on the current status and future perspectives. Clin Chem Lab Med. 2021;59(5):861-868.
11. Banerjee A. Artificial Intelligence in Medical Education: The "Indian" Perspective. Med J Armed Forces India. 2023;79(4):365-367.
12. Meskó B. The impact of Generative AI on the future of medical education. BMJ. 2023; 381:1025.