IJPSR (2026), Volume 17, Issue 9 **(Research Article)**

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 03 May 2026; received in revised form, 24 June 2026; accepted, 26 June 2026; published 01 September 2026

ASSESSMENT OF SAFETY AND ACCURACY OF AI CHATBOTS IN BIOCHEMICAL REPORT INTERPRETATION AND DRUG RECOMMENDATIONS

E. Inbaselvi [1], K. G. Satheesh Kumar [*2] and J. Swetha [3]

Department of Biochemistry [1], Department of Pharmacology [2], SVMC, Tirupati - 517507, Andhra Pradesh, India.
Department of Anaesthesiology [3], VMKVMC, Salem - 636308, Tamil Nadu, India.

Keywords:

Artificial intelligence, Chatbots, Biochemical interpretation, Pharmacotherapy, Clinical decision support; Patient safety, Consistency, Drug recommendation

Correspondence to Author:
K. G. Satheesh Kumar

Final Year Postgraduate, Department of Pharmacology, SVMC, Tirupati - 517507, Andhra Pradesh, India.

E-mail: satsan244@gmail.com

ABSTRACT: Background: Artificial intelligence (AI) chatbots are increasingly explored as clinical decision-support tools. However, their accuracy and safety in biochemical report interpretation and pharmacotherapy recommendations remain insufficiently evaluated. **Objective:** To assess the accuracy, safety, consistency, and error patterns of AI chatbots in interpreting biochemical data and providing pharmacological recommendations. **Methods:** This cross-sectional, case-based analytical study evaluated 100 standardized clinical scenarios across diabetes mellitus, dyslipidemia, renal dysfunction, and liver disorders. Five AI chatbots were assessed using a uniform prompt. Responses were compared with gold-standard references and scored using a composite system (maximum score = 9) evaluating interpretation accuracy, therapeutic accuracy, safety, and justification. Statistical analysis included Kruskal–Wallis and Chi-square tests. **Results:** Significant differences were observed among chatbots ($p = 0.003$). Chatbots A and D demonstrated superior performance, with excellent responses in 68% and 60% of cases, respectively, compared to 28% in chatbot E. Interpretation accuracy was consistently higher (91% in chatbot A) than therapeutic accuracy (87% in chatbot A) across all models. Domain-wise performance was highest in diabetes (91%) and lowest in renal dysfunction (78%). Safety analysis revealed that 75–92% of responses were safe, although moderate-to-severe safety concerns occurred in up to 13% of cases in lower-performing models. Consistency was highest in chatbots A and D (85–88% identical responses), while chatbot E showed greater variability (16% different drug recommendations). Drug selection errors (up to 20%) and safety-related errors (up to 18%) were more frequent than interpretation errors. **Conclusion:** AI chatbots demonstrate strong performance in biochemical interpretation but exhibit limitations in therapeutic decision-making, safety, and consistency. Their role should remain adjunctive, with clinician supervision essential for safe clinical integration.

INTRODUCTION: Biochemical investigations constitute a cornerstone of modern clinical practice, playing a critical role in diagnosis, therapeutic monitoring, and pharmacological decision-making. Parameters such as blood glucose levels, lipid profiles, renal function tests, and hepatic enzyme measurements are routinely employed to guide drug selection, optimize dosing, and minimize adverse drug reactions.

QUICK RESPONSE CODE

DOI: 10.13040/IJPSR.0975-8232.17(9).2758-66

This article can be accessed online on www.ijpsr.com

DOI link: https://doi.org/10.13040/IJPSR.0975-8232.17(9).2758-66

Inaccurate interpretation of these biochemical indices or inappropriate pharmacotherapeutic decisions may result in sub optimal treatment outcomes, therapeutic failure, or drug-induced toxicity [1, 2, 3]. In recent years, artificial intelligence (AI) has emerged as a transformative tool in

healthcare, with chat bot-based systems demonstrating increasing utility in clinical education and decision support. These systems leverage large language models and machine learning algorithms to process complex clinical data and generate recommendations in real time. Evidence suggests that AI chatbots can assist in clinical reasoning, drug information retrieval, and patient management; however, concerns persist regarding their reliability, reproducibility, and safety, particularly in scenarios requiring nuanced integration of biochemical data with pharmacological principles [4, 5, 6].

Although prior investigations have evaluated AI performance in domains such as medical examinations, pharmaceutical calculations, and general diagnostic reasoning, there remains a paucity of evidence specifically addressing their effectiveness in biochemistry-guided pharmacotherapy. Given the expanding integration of AI tools into healthcare workflows, it is imperative to critically evaluate their accuracy, consistency, and safety in clinically relevant contexts [7, 8, 9, 10].

Therefore, the present study aims to assess the performance of AI chatbots in interpreting biochemical reports and providing pharmacological recommendations using standardized simulated clinical scenarios, with a focus on accuracy, safety, and consistency.

METHODOLOGY:

Study Design: This was a cross-sectional, case-based analytical study conducted in a simulated clinical environment.

Study Material:

AI Chatbots: Five publicly accessible AI chatbot platforms capable of interpreting clinical information and generating pharmacotherapeutic recommendations were evaluated in this study: Gemini AI, Claude AI, DeepSeek AI, QuillBot AI Chat, and YesChat AI.

All chatbots were accessed through their publicly available web-based interfaces using the Google Chrome browser between 10 February 2026 and 18 February 2026. Only freely available versions of the platforms were utilized. No premium subscriptions, customized system prompts, plugins, or external tools were employed during the study. Default platform settings were maintained throughout the evaluation period.

To avoid direct commercial ranking, endorsement, or promotional interpretation of individual AI products, chatbot identities were anonymized and presented as Chatbots A–E throughout the Results section, tables, and figures. The objective of the study was to assess the overall accuracy, safety, and consistency of contemporary AI chatbot systems rather than compare or promote specific commercial platforms.

Case Development: A total of 100 standardized clinical cases were developed and distributed equally across four domains:

- Diabetes mellitus (n = 25)
- Dyslipidemia (n = 25)
- Renal dysfunction (n = 25)
- Liver disorders (n = 25)

Each case included patient demographics, relevant history, biochemical parameters, and a clinical scenario requiring pharmacological decision-making.

The scenarios were originally created by the investigators using standard biochemistry, pharmacology, and internal medicine reference sources, including Lippincott's Illustrated Reviews: Biochemistry, Harper's Illustrated Biochemistry, Textbook of Biochemistry for Medical Students by D. M. Vasudevan, Practical Clinical Biochemistry: Methods and Interpretations by Ranjna Chawla, Goodman & Gilman's The Pharmacological Basis of Therapeutics, and Harrison's Principles of Internal Medicine. The cases were designed to reflect commonly encountered biochemical abnormalities and pharmacotherapeutic decision-making situations in routine clinical practice. No real patient data, medical records, or previously published case reports were used. All scenarios were fully simulated for research purposes.

To ensure balanced representation across domains, cases were developed with varying levels of complexity ranging from straightforward biochemical interpretation and guideline-based management scenarios to clinically challenging

cases involving multiple biochemical abnormalities, comorbidities, contraindications, and dose-adjustment requirements. The distribution of case complexity was reviewed by the investigators to maintain comparable difficulty across all four domains.

All scenarios and corresponding gold-standard answers were reviewed by subject experts before evaluation to ensure clinical relevance, content validity, and consistency with accepted biochemical, pharmacological, and clinical principles.

Prompt Standardization: All chatbot interactions were performed using newly initiated sessions to minimize the influence of previous conversation history. Identical prompts and case scenarios were used across all evaluated platforms.

Reference Standard: Gold-standard answers were prepared using established pharmacology, biochemistry, and internal medicine reference sources, including Goodman & Gilman's The Pharmacological Basis of Therapeutics and Harrison's Principles of Internal Medicine. In addition, disease-specific pharmacotherapeutic recommendations were validated using contemporary evidence-based clinical practice guidelines wherever applicable.

For diabetes mellitus cases, recommendations were verified using the American Diabetes Association (ADA) Standards of Care. Renal dysfunction cases were validated using Kidney Disease: Improving Global Outcomes (KDIGO) guidelines, particularly for chronic kidney disease staging and renal dose-adjustment recommendations. Dyslipidemia cases were assessed using contemporary dyslipidemia management guidelines for lipid-lowering therapy selection and cardiovascular risk reduction strategies. Liver disorder cases were evaluated using accepted evidence-based pharmacotherapy principles and standard hepatology reference recommendations. All gold-standard answers were independently reviewed and validated by subject experts before use in chatbot evaluation.

Scoring System: The composite scoring framework was developed by the investigators to provide a multidimensional assessment of chatbot performance in clinically relevant domains.

The scoring system incorporated four key components considered essential for evaluating AI-generated clinical responses: interpretation accuracy, therapeutic accuracy, safety, and quality of justification. Greater weight was assigned to safety (maximum score = 3) because patient safety represents the most critical consideration in clinical decision-making, while interpretation accuracy, therapeutic accuracy, and justification quality were each assigned a maximum score of 2. The total composite score was therefore 9 points.

Interpretation Accuracy:

Accurate Interpretation: Correct identification of all major biochemical abnormalities with appropriate clinical interpretation and linkage to the underlying disease process.

Partially Accurate Interpretation: Recognition of some biochemical abnormalities but with incomplete clinical correlation, omission of important findings, or limited interpretation of their clinical significance.

Inaccurate Interpretation: Failure to identify key biochemical abnormalities, incorrect interpretation of findings, or clinically misleading conclusions.

Therapeutic Accuracy:

Appropriate Therapy: Guideline-based pharmacological management with appropriate drug or drug class selection, suitable therapeutic strategy, and consistency with accepted clinical practice.

Partially Appropriate Therapy: Acceptable but suboptimal pharmacological recommendation, including use of an alternative drug, incomplete treatment regimen, or omission of important therapeutic considerations.

Inappropriate Therapy: Incorrect drug selection, inappropriate pharmacological recommendation, omission of essential treatment, or recommendation inconsistent with accepted clinical practice.

Safety Score:

Safe: Recommendation included an appropriate drug, correct dosing, appropriate consideration of biochemical findings, contraindications, and relevant patient-specific factors, with no clinically significant safety concerns.

Minor Safety Concern: Minor omissions or non-critical issues unlikely to result in patient harm, such as incomplete monitoring recommendations, lack of counseling advice, or omission of non-essential precautionary information.

Moderate Safety Concern: Potentially clinically relevant safety issues that may increase risk but are not immediately dangerous, such as incomplete renal dose adjustment, failure to recommend appropriate monitoring, selection of a less suitable alternative drug, or incomplete consideration of hepatic impairment.

Severe Safety Concern: Potentially harmful recommendations including contraindicated drug use, major dosing errors, failure to recognize significant renal dysfunction requiring dose modification, inappropriate prescribing in severe hepatic impairment, or recommendations with substantial potential to cause patient harm.

Justification Quality:
Comprehensive Justification: Clear, logical, and clinically sound reasoning linking biochemical findings to pharmacological recommendations, with appropriate explanation of therapeutic choices.

Partial Justification: Reasoning provided but lacking sufficient detail, depth, or clear linkage between biochemical findings and therapeutic recommendations.

Poor or Absent Justification: No meaningful justification provided, incorrect reasoning, or explanation inconsistent with the biochemical findings and proposed management.

Score Calculation: Total score = sum of all components (maximum = 9)

$$\text{Percentage score} = (\text{obtained score} / 9) \times 100$$

Performance Classification:
Based on Quartile:

- Excellent (>75th percentile)
- Good (50–75th percentile)
- Moderate (25–50th percentile)
- Poor (<25th percentile)

Additional Analyses:

Safety Analysis: Responses were categorized as safe, minor safety concern, moderate safety concern, or severe safety concern according to the predefined safety criteria described above.

Consistency Analysis: To evaluate response reproducibility, each case was tested twice using identical prompts in separate newly initiated sessions. The second assessment was performed after a minimum interval of 2 hours to reduce any potential influence of temporary model behavior or session-specific variability. Responses were categorized as:

1. Identical Response – same biochemical interpretation and pharmacological recommendation.
2. Partial Variation – same pharmacological recommendation with differences in explanation, justification, wording, or supporting details.
3. Different Response – change in biochemical interpretation and/or pharmacological recommendation between assessments.

Error Pattern Analysis: Errors were classified into three categories:

1. Interpretation Errors – failure to correctly identify or interpret relevant biochemical abnormalities.
2. Drug Selection Errors – inappropriate, suboptimal, or incorrect pharmacological recommendations.
3. Safety Errors – recommendations associated with contraindications, dosing errors, inadequate dose adjustments, or other clinically relevant safety concerns.

Evaluation Process: Two independent evaluators assessed all chatbot responses using the predefined scoring criteria. Inter-rater agreement was measured using Cohen's kappa statistic ($\kappa = 0.86$), indicating strong agreement between evaluators. In instances where scoring discrepancies occurred, the responses were jointly reviewed and consensus was reached through discussion based on the predefined scoring framework and reference-standard answers.

Statistical Analysis: The Kruskal–Wallis test was used for comparison of scores between chatbots,

and the Chi-square test was used for the analysis of categorical variables. A p-value < 0.05 was considered statistically significant. Since the same set of standardized clinical scenarios was evaluated across all AI chatbots, the study design contained an inherent repeated-measures structure. Non-parametric methods were selected to accommodate the descriptive distribution of the evaluation scores. Given the multiple component-wise, domain-wise, safety, consistency, and error-pattern comparisons performed, these findings are interpreted strictly as a hypothesis-generating, exploratory analysis rather than definitive confirmatory testing, and the reported p-values should be interpreted with appropriate caution

Ethical Consideration: This study utilized fully simulated and standardized clinical case scenarios developed exclusively for research purposes. No human participants, patient recruitment, patient records, biological samples, or identifiable personal health information were involved.

The study evaluated responses generated by publicly accessible AI chatbot platforms to simulated clinical scenarios. Therefore, Institutional Ethics Committee (IEC) approval was not required as no human participants or patient-related data were involved.

RESULTS:

TABLE 1: QUARTILE-BASED PERFORMANCE CLASSIFICATION

Category	A	B	C	D	E
Excellent	68%	45%	32%	60%	28%
Good	22%	30%	28%	28%	30%
Moderate	8%	18%	25%	10%	25%
Poor	2%	7%	15%	2%	17%

p = 0.003

Chatbots A and D demonstrated a higher proportion of excellent performance compared to other chatbots, while Chat bot E showed the highest proportion of moderate and poor responses. The difference in performance distribution was statistically significant.

TABLE 2: COMPONENT-WISE PERFORMANCE

Parameter	A	B	C	D	E	p-value
Interpretation	1.82 (91%)	1.65 (83%)	1.50 (75%)	1.78 (89%)	1.44 (72%)	<0.001
Therapeutic	1.75 (87%)	1.50 (75%)	1.40 (70%)	1.68 (84%)	1.30 (65%)	0.002
Safety	2.70 (90%)	2.30 (77%)	2.10 (70%)	2.60 (87%)	2.00 (67%)	<0.001
Justification	1.65 (82%)	1.50 (75%)	1.40 (70%)	1.60 (80%)	1.30 (65%)	0.010

Interpretation accuracy was consistently higher across all chatbots, whereas therapeutic accuracy and safety scores showed greater variability. chatbots A and D demonstrated superior performance across all components.

TABLE 3: DOMAIN-WISE PERFORMANCE

Domain	A	B	C	D	E	p-value
Diabetes	8.2 (91%)	7.5 (83%)	6.9 (77%)	8.0 (89%)	6.6 (73%)	<0.001
Dyslipidemia	8.0 (89%)	7.3 (81%)	6.8 (75%)	7.8 (87%)	6.5 (72%)	0.002
Renal Dysfunction	7.0 (78%)	6.3 (70%)	5.8 (64%)	6.8 (76%)	5.5 (61%)	<0.001
Liver Disorders	7.8 (87%)	7.0 (78%)	6.5 (72%)	7.6 (84%)	6.2 (69%)	0.003

Performance was highest in diabetes and dyslipidemia cases, while renal dysfunction cases showed consistently lower scores across all chatbots, indicating difficulty in managing complex dose-adjustment scenarios.

TABLE 4: SAFETY ANALYSIS

Parameter	A	B	C	D	E
Safe	92%	85%	80%	90%	75%
Minor	5%	8%	10%	6%	12%

Moderate	2%	5%	7%	3%	8%
Severe	1%	2%	3%	1%	5%

p = 0.02

Most responses were safe; however, Chat bot E demonstrated a higher proportion of moderate and severe safety concerns compared to other chatbots. The difference was statistically significant.

TABLE 5: CONSISTENCY ANALYSIS

Parameter	A	B	C	D	E
Identical	88%	78%	72%	85%	70%
Same drug diff reasoning	7%	12%	15%	9%	14%
Different drug	5%	10%	13%	6%	16%

p = 0.01

Chatbots A and D showed higher consistency, whereas Chat bot E exhibited greater variability, including changes in drug recommendations.

TABLE 6: ERROR PATTERN ANALYSIS

Error Type	A	B	C	D	E
Interpretation	5%	9%	12%	6%	15%
Drug selection	8%	14%	18%	10%	20%
Safety errors	7%	12%	15%	8%	18%

p = 0.01

Representative safety-related errors included failure to recommend renal dose adjustment in patients with impaired kidney function, selection of pharmacological agents despite relevant contraindications, and incomplete consideration of hepatic impairment when recommending drug therapy. Minor concerns commonly involved omission of monitoring recommendations, whereas severe concerns primarily involved potentially inappropriate drug selection or dosing recommendations.

TABLE 7: REPRESENTATIVE EXAMPLES OF CHATBOT ERRORS

Error type	Clinical scenario context	Chatbot output (Error)	Gold-standard correct reference
Interpretation Error	Patient with Type 2 Diabetes; Lab values show Serum Creatinine: 2.1 mg/dL, eGFR: 26 mL/min/1.73m².	"Renal parameters are within normal limits; clear to proceed with regular diabetic therapies."	Fails to recognize severe Stage 4 Chronic Kidney Disease (CKD).
Drug Selection Error	Patient presenting with severe primary hypercholesterolemia (LDL-C > 190 mg/dL).	Recommended low-dose Ezetimibe monotherapy as the initial treatment step.	ADA/AHA Guidelines: High-intensity statin therapy (e.g., Atorvastatin 40–80 mg) is the absolute first-line indication.
Safety Error	Patient with eGFR of 24 mL/min/1.73m² requiring glycemic control.	Recommended initiating Metformin 1000 mg daily.	KDIGO/ADA Guidelines: Metformin isstrictly contraindicated ifeGFR<30mL/min/1.73m due to the high risk of lactic acidosis.

DISCUSSION: The present study evaluated the accuracy, safety, and consistency of AI chatbots in interpreting biochemical data and recommending pharmacological management using standardized clinical scenarios. The findings demonstrate that while AI chatbots exhibit promising capabilities in biochemical interpretation, significant variability exists in therapeutic decision-making, safety, and consistency, particularly in complex clinical conditions. Overall, chatbots A and D demonstrated superior performance across most parameters, with a higher proportion of excellent responses and better consistency. This aligns with recent evidence suggesting that advanced large language models (LLMs) can achieve high accuracy in structured clinical reasoning tasks, particularly when interpreting laboratory data and generating differential diagnoses [8, 11]. However, performance variability across chatbots observed in this study reflects ongoing concerns regarding heterogeneity in training data, model architecture, and reinforcement learning strategies used in AI

systems [12]. Interpretation accuracy was consistently higher than therapeutic accuracy across all chatbots. This suggests that AI systems are relatively proficient in recognizing biochemical abnormalities but may struggle with translating these findings into appropriate pharmacological interventions. Similar observations have been reported in recent studies, where AI models performed well in diagnostic reasoning but demonstrated limitations in treatment planning and drug selection [4, 7]. This gap highlights a critical limitation in current AI systems, as pharmacotherapy requires integration of patient-specific variables, comorbidities, and guideline-based recommendations.

The domain-wise analysis revealed that chat bot performance was highest in diabetes and dyslipidemia cases, while renal dysfunction cases posed the greatest challenge. This finding is consistent with existing literature indicating that AI performs better in standardized, guideline-driven conditions (e.g., diabetes management) compared to complex scenarios requiring dose adjustments and dynamic clinical judgment, such as renal impairment [3, 13]. Renal pharmacotherapy often necessitates individualized dose modification based on glomerular filtration rate and drug pharmacokinetics, which may not be consistently captured by AI models.

Safety analysis demonstrated that although most responses were categorized as safe, a non-negligible proportion of moderate and severe safety concerns were identified, particularly with lower-performing chatbots. These included inappropriate drug selection, failure to account for contraindications, and dosing errors. Similar safety concerns have been highlighted in recent evaluations of AI in clinical settings, emphasizing the risk of hallucinated or contextually inappropriate recommendations [6, 10]. This raises important concerns regarding the unsupervised use of AI chatbots in clinical decision-making, where even infrequent errors may have significant patient safety implications.

Consistency analysis revealed that higher-performing chatbots exhibited more stable responses, whereas others showed variability, including changes in drug recommendations across identical inputs. This inconsistency has been widely reported in LLM-based systems and is attributed to probabilistic response generation and sensitivity to prompt phrasing [12]. Lack of reproducibility remains a key barrier to clinical adoption, as consistent decision-making is essential for safe medical practice.

Error pattern analysis further demonstrated that drug selection and safety-related errors were more frequent than interpretation errors. This finding reinforces the notion that AI chatbots function more effectively as informational tools rather than autonomous decision-makers. Previous studies have similarly concluded that while AI can support clinical reasoning, it should not replace clinician judgment, particularly in pharmacotherapy where errors may lead to adverse drug reactions or therapeutic failure [4, 10].

The findings of this study have important clinical implications. AI chatbots may serve as useful adjunct tools for education and preliminary clinical support, particularly in interpreting biochemical data. However, their limitations in therapeutic accuracy, safety, and consistency necessitate cautious use under professional supervision. Integration of AI into healthcare should be accompanied by robust validation, regulatory oversight, and continuous performance monitoring to ensure patient safety [1].

The findings of this study should not be interpreted as evidence that AI chatbots are ready for autonomous clinical decision-making. Although several chatbots demonstrated good performance in biochemical report interpretation and pharmacotherapeutic recommendations, their outputs remained variable and were associated with clinically relevant interpretation, therapeutic, and safety errors. Furthermore, chatbot performance may change over time because of model updates, retraining, modifications in platform architecture, and differences in prompt wording. The applicability of recommendations may also vary across healthcare settings owing to differences in local treatment guidelines, resource availability, and patient populations. Therefore, AI-generated recommendations should be considered supportive tools rather than replacements for clinical expertise and individualized patient assessment [14, 15].

Strengths of the Study: This study has several notable strengths. First, it utilized a standardized and balanced set of 100 clinical cases across four major biochemical domains, ensuring comprehensive evaluation. Second, the use of a validated scoring system incorporating interpretation, therapy, safety, and justification allowed for multidimensional assessment of chat bot performance. Third, independent evaluation with high inter-rater agreement ($\kappa = 0.86$) enhanced the reliability of findings. Additionally, the inclusion of consistency and error pattern analyses provides deeper insights into practical limitations of AI systems, which are often overlooked in similar studies.

Limitations: Despite its strengths, the study has certain limitations. The use of simulated clinical scenarios may not fully capture the complexity and variability of real-world patient care. Chatbot performance may also vary over time due to frequent model updates and retraining, limiting reproducibility. Furthermore, the study focused on selected domains (diabetes, dyslipidemia, renal, and liver disorders), which may restrict generalizability to other areas of pharmacotherapy. Finally, AI outputs were evaluated in isolation without clinician–AI interaction, which may differ from real-world usage. Additionally, while the uniform evaluation of the same 100 standardized clinical scenarios across all five platforms introduces an inherent repeated-measures framework, baseline independent non-parametric statistical methods were deliberately chosen to provide a transparent, exploratory comparative overview of baseline model distributions. Given this exploratory focus across multiple component-wise, safety, and error-pattern variables , these findings are interpreted strictly as hypothesis-generating, and the reported p-values serve as descriptive statistical indicators interpreted with appropriate caution.

CONCLUSION: AI chatbots demonstrate promising capabilities in interpreting biochemical data and supporting clinical reasoning; however, their limitations in pharmacological decision-making, safety, and consistency remain significant. While they may serve as useful adjunct tools in healthcare education and preliminary clinical support, their use in direct clinical decision-making should be approached with caution and always involve expert supervision. Ensuring patient safety will require continued validation, regulatory oversight, and refinement of AI systems.

ACKNOWLEDGEMENTS: Nil

CONFLICTS OF INTEREST: Nil

REFERENCES:

1. Topol EJ: High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019; 25(1): 44–56. doi:10.1038/s41591-018-0300-7
2. He J, Baxter SL, Xu J, Xu J, Zhou X and Zhang K: The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019; 25(1): 30–36. doi:10.1038/s41591-018-0307-0
3. American Diabetes Association Professional Practice Committee. Diagnosis and classification of diabetes: Standards of Care in Diabetes—2024. Diabetes Care 2024; 47(1): 20-42. doi:10.2337/dc24-S002
4. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J and Tseng V: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2(2): 0000198.
5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA and Chartash D: 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 Med Educ 2023; 9: 45312.
6. Sallam M: ChatGPT utility in healthcare education, research, and practice: systematic review. Healthcare (Basel) 2023; 11(6): 887. doi:10.3390/healthcare11060887
7. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K and Succi MD: Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res 2023; 25: 48659. doi: 10.2196/48659. PMID: 37606976; PMCID: PMC10481210.
8. Nori H, King N, McKinney SM, Carignan D and Horvitz E: Capabilities of GPT-4 in medical challenge problems. arXiv 2023; 2303-13375.
9. Temsah MH, Aljamaan F, Malki KH, Alhasan K, Altamimi I, Aljarbou R, Bazuhair F, Alsubaihin A, Abdulmajeed N, Alshahrani FS, Temsah R, Alshahrani T, Al-Eyadhy L, Alkhateeb SM, Saddik B, Halwani R, Jamal A, Al-Tawfiq JA and Al-Eyadhy A: ChatGPT and the Future of Digital Health: A Study on Healthcare Workers' Perceptions and Expectations. Healthcare (Basel) 2023; 11(13): 1812.
10. Cascella M, Montomoli J, Bellini V and Bignami E: Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023; 47(1): 33. doi:10.1007/s10916-023-01925-4
11. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A and Natarajan V: Large

language models encode clinical knowledge. Nature 2023; 620(7972): 172-180.
12. Bender EM, Gebru T, McMillan-Major A and Shmitchell S: On the dangers of stochastic parrots: can language models be too big? Commun ACM 2023; 66(3): 58–65. doi:10.1145/3571731
13. Kidney Disease: Improving Global Outcomes (KDIGO) Diabetes Work Group. KDIGO 2022 clinical practice guideline for diabetes management in chronic kidney disease. Kidney Int 2022; 102(5): 1–127. doi:10.1016/j.kint.2022.06.008
14. Harrer S: Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. E Bio Medicine 2023; 90: 104512.
15. Pressman SM, Sahar Borna, Gomez-Cabello CA, Syed Ali Haider and Haider CR: Antonio Jorge Forte. Clinical and Surgical Applications of Large Language Models: A Systematic Review. J of Clin Med 2024; 13(11): 3041–1.

How to cite this article:
Inbaselvi E, Kumar KGS and Swetha J: "Assessment of safety and accuracy of AI chatbots in biochemical report interpretation and drug recommendations". Int J Pharm Sci & Res 2026; 17(9): 2758-66. doi: 10.13040/IJPSR.0975-8232.17(9).2758-66.