

The Performance of ChatGPT-3.5 versus ChatGPT-4 on CRRT Alarm Troubleshooting Questions



AKI & CRRT Conference

Mohammad S. Sheikh, MD¹; Fawad M. Qureshi, MD¹; Kianoush B. Kashani, MD^{1,2}; Charat Thongprayoon, MD¹; Iasmina Craici, MD¹; Wisit Cheungpasitporn, MD¹

¹Mayo Clinic Minnesota, Division of Nephrology and Hypertension, ²Mayo Clinic Minnesota, Division of Pulmonary and Critical Care Medicine

Abstract

This study evaluates the accuracy of ChatGPT-3.5 and ChatGPT-4 in addressing queries related to Continuous Renal Replacement Therapy (CRRT) machine alarm troubleshooting. Both models underwent two rounds of 50 alarm questions that were selected by two nephrologists in intensive care. Accuracy was determined by comparing the model responses to predetermined answer keys provided by critical care nephrologists, and consistency was determined by comparing outcomes across the two rounds.

The accuracy rate of ChatGPT-3.5 was 86% and 84% in the two rounds, while the accuracy rate of ChatGPT-4 was 90% and 94%. The agreement between the first and second rounds of ChatGPT-3.5 was 84%, with a Kappa statistic of 0.78, while the agreement of ChatGPT-4 was 92% with a Kappa statistic of 0.88.

Although ChatGPT-4 tended to provide more accurate and consistent responses than ChatGPT-3.5, there was no statistically significant difference between accuracy and agreement rate between ChatGPT-3.5 and ChatGPT-4. ChatGPT-4 had higher accuracy and consistency but did not achieve statistical significance. While these findings are encouraging, there is still potential for further development to achieve even greater reliability.

Introduction

In the critical care landscape, CRRT machines are indispensable, yet they are often accompanied by a multitude of alarms which can be both frequent and critical. The management of these alarms is paramount for patient safety and effective delivery of care.

An analysis of 35,732 alarm incidents by Broman et al in 2018 identified the ten most encountered CRRT machine alarms, which include issues such as extreme negative access pressure and full effluent bags. These alarms necessitate immediate and precise responses to prevent adverse patient outcomes.

This backdrop provides an ideal setting for evaluating the potential of ChatGPT-3.5 and ChatGPT-4 in interpreting and troubleshooting these alarms. By systematically assessing these AI models, we can better understand their utility in a clinical setting.

This study, therefore, not only evaluates the technological capabilities of AI but also addresses a tangible clinical challenge, providing insights into how AI can enhance patient care in real-world critical care scenarios.

Methods

The assessment consisted of two rounds of evaluation for ChatGPT-3.5 and ChatGPT-4, with each model addressing 50 CRRT machine alarm questions compiled and verified by two critical care nephrologists.

Accuracy was determined by comparing the model responses to a predetermined answer key, and consistency was noted by comparing outcomes across the two rounds.

Consistency was assessed by comparing outcomes across the two rounds, utilizing the Cohen's Kappa statistic to measure inter-rater reliability and control for chance agreement.

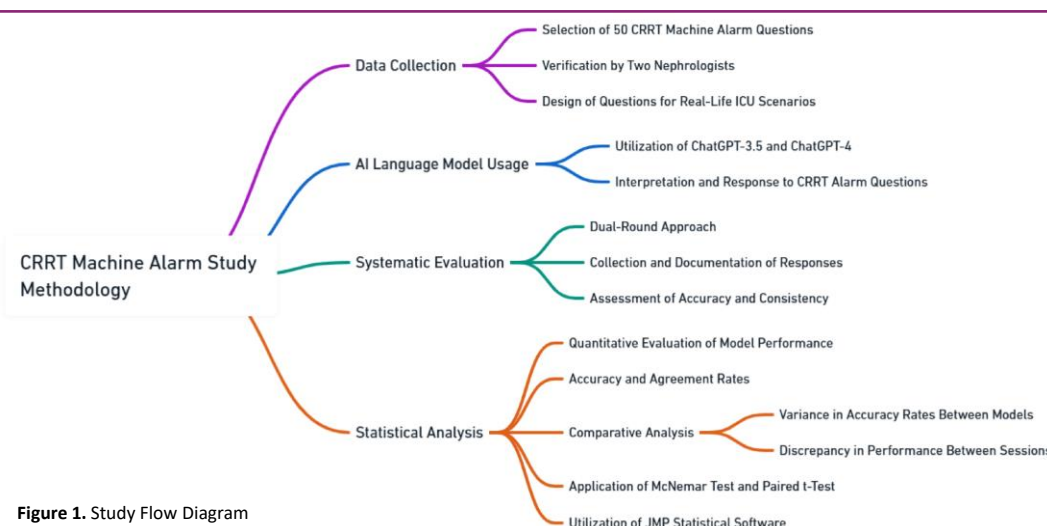


Figure 1. Study Flow Diagram

Results

- ChatGPT-3.5 scored 86.0% (43/50) on the first run and 84.0% (42/50) on the second run.
- ChatGPT-4 scored 90.0% (45/50) on the first run and 92% (46/50) on the second run.
- Agreement between ChatGPT-3.5 first and second runs was 84.0% (kappa = 0.759), with the same response in 42 of 50 questions, of which 39 were correct and 3 were incorrect.
- Agreement between ChatGPT-4 first and second runs was 92.0% (kappa = 0.889), with the same response in 46 of 50 questions, of which 44 were correct and 2 were incorrect.
- Furthermore, when assessing open-ended questions and narrative responses, both ChatGPT models produced answers that aligned with the multiple-choice answer key without leading to potentially leading to potentially harmful recommendations.

	ChatGPT-3.5	ChatGPT-4	p-Value
Accuracy—1st run	43/50 (86)	45/50 (90%)	0.63
Accuracy—2nd run	42/50 (84)	47/50 (94%)	0.18
Average accuracy	85%	92%	0.09
Agreement	42/50 (84%)	46/50 (92%)	0.34
Kappa statistics	0.78	0.88	-

Accuracy—1st run and 2nd run: represents the proportion of questions correctly answered by ChatGPT versions 3.5 and 4 during the first and second rounds of testing, respectively. Average accuracy: calculated as the mean accuracy across both runs for each ChatGPT version.

Table 1. The accuracy and agreement of ChatGPT-3.5 and ChatGPT-4 on CRRT alarm questions

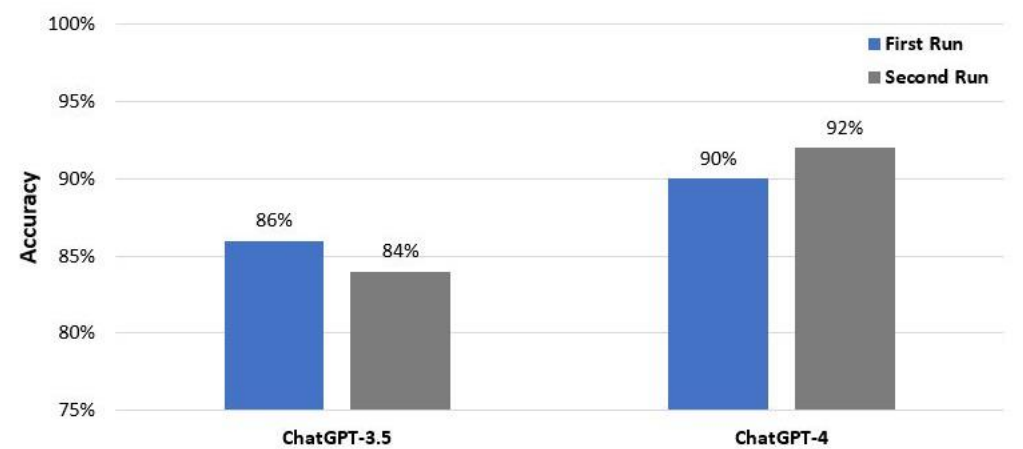



Table 2. ChatGPT 3.5 and ChatGPT-4 performance on CRRT alarm questions



Most occurring alarms Name	Ten most occurring alarms Type and action	Number and frequency, n/%
Extreme negative access pressure not self cleared	Medium Stop all pumps	4,152/11.62
Effluent bag full	Low Stop fluid pumps	3,548/10.03
Extreme negative access pressure self cleared	Low Stop all pumps	3,260/9.12
Preblood bag empty	Low Stop fluid pumps	3,079/8.62
Effluent scale open	Low Stop fluid pumps	2,235/6.25
Dialysis bag empty	Low Stop fluid pumps	1,987/5.56
Preblood scale open	Low Stop fluid pumps	1,817/5.09
Dialysate scale open	Low Stop fluid pumps	1,249/3.50
Check access	Medium None	1,183/3.31
Liquid level sensor low*	Info None	1,158/3.24

Blood Purif 2018;46:220-227
DOI: 10.1159/000489213

Broman/Bell/Joannes-Boyou/Ronco

Figure 2. Prismaflex and PrisMax CRRT Systems Table 3. Top Ten CRRT Machine Alarms Among 35,732 Evaluated Incidents

Conclusion

Within CRRT machine alarms and troubleshooting, ChatGPT-4 outperformed ChatGPT-3.5 in accuracy and consistency. These findings underscore the advancements in AI capabilities. However, there is still potential for further development to achieve even greater reliability. This advancement is essential for ensuring the highest patient care and safety standards in managing CRRT machine-related issues.



THE 29TH INTERNATIONAL CONFERENCE ON
ADVANCES IN CRITICAL CARE NEPHROLOGY

AKI & CRRT 2024

MARCH 12-15, 2024 SAN DIEGO, CALIFORNIA