Evaluating a ChatGPT-Based Hyponatremia Teaching Module for Internal Medicine Residents: A Pilot Study

Dr. Christelle El Khoury, Lebanese American University

Christelle El Khoury, MD, is a family physician with a Master's in Applied AI from the Lebanese American University (LAU). She previously worked with the Department of Family Medicine at the University of Michigan Medical School on research focused on cervical cancer screening and patient-centered exam alternatives. She has also worked as a medical expert in creating clinical prompts to train large language models for medical and educational applications. Her current interests include leveraging AI to enhance medical education and strengthen patient engagement.

Dr. Toni Sabbouh MD, FASN, University of Pennsylvania

Toni Sabbouh, MD, FASN, is an American board-certified physician in Internal Medicine and Nephrology and a fellowship-trained nephrologist from the University of Michigan Hospitals. A Fellow of the American Society of Nephrology, he is recognized for excellence in clinical care and scholarly contributions. Dr. Sabbouh currently serves as a Clinical Assistant Professor of Medicine at Penn Medicine in Philadelphia, specializing in kidney disease and advanced clinical management.

Dr. Christine Palazzolo DO, University of Pennsylvania

Christine Palazzolo, DO, is a physician who completed her Internal Medicine residency and served as Chief Resident at Pennsylvania Hospital. She will begin fellowship training in Nephrology at the University of Pennsylvania in 2026. She is actively involved in patient advocacy and women in medicine initiatives and has published with the American Society of Nephrology. Her academic interests include palliative nephrology and improving patient-centered outcomes in kidney disease.

Dr. Sola Aoun Bahous, Lebanese American University

Dr. Aoun Bahous is Professor of Medicine and Dean of the Lebanese American University School of Medicine. A nephrologist, pharmacologist, and medical educator, she has led innovations in medical curricula, including the early integration of Artificial Intelligence and Digital Health across all medical years since 2018. Her research interests include hypertension, electrolyte disorders, and medical education. She serves on the International Advisory Committee on Artificial Intelligence in Medical Education and as Deputy Chair of the International Society of Nephrology's Education Working Group. With a particular interest in pedagogy, she focuses on leveraging innovative educational tools to make complex nephrology concepts more accessible and engaging for learners.

Dr. Haidar Harmanani, Lebanese American University

Haidar Harmanani received the B.S., M.S. and Ph.D. degrees in Computer Engineering from Case Western Reserve University, Cleveland, Ohio, in 1989, 1991 and 1994, respectively. He is currently a professor of computer science and the Dean of the School of Arts and Sciences at the Lebanese American University. His research interests span high-level synthesis, design automation, parallel computing, machine learning, and applied artificial intelligence. Throughout his career, he has led and contributed to multiple institutional initiatives in academic program development, accreditation, research infrastructure, and strategic planning. Dr. Harmanani has published extensively in the areas of design automation and optimization techniques for hardware–software systems, and he has supervised numerous graduate and undergraduate research projects in computing and data-driven systems.

Evaluating a ChatGPT-Based Hyponatremia Teaching Module for Internal Medicine Residents: A Pilot Study

Abstract

Generative artificial intelligence (GenAI) is increasingly used in medical education to enhance engagement and simplify complex topics. Nephrology, often perceived as challenging by trainees, may particularly benefit from such innovations. This pilot study evaluated whether a ChatGPT-based interactive hyponatremia learning module, enhanced with Retrieval-Augmented Generation (RAG) using content from Pocket Nephrology, could support internal medicine resident education. Residents were randomized 2:1 to intervention (module + quiz) or control (quiz only, open-resource). Of 28 participants (19 intervention, 9 control), mean quiz scores were similar (7.47 ± 1.90 vs. 7.44 ± 1.94; p = 0.97). However, intervention group confidence increased from 2.67 ± 1.23 to 3.50 ± 0.80 on a 5-point scale. Usability ratings were high for clarity and appropriateness. Expert review of 14 transcripts found most outputs accurate, though occasional errors persisted despite RAG. A ChatGPT-based, RAG-enhanced module was feasible, well-received, and improved resident confidence, though it did not significantly increase short-term quiz performance. Integration of GenAI in nephrology education shows promise but requires expert oversight.

1 Introduction

Generative artificial intelligence (GenAI) is rapidly transforming healthcare and medical education. For graduate medical education (GME), GenAI offers unprecedented opportunities to enhance learner engagement, simulate clinical reasoning, and deliver personalized instruction at scale. Nephrology represents a particularly relevant domain for such innovations, as it is consistently cited by residents and medical students as one of the most challenging specialties to master [1, 2]. The complexity of renal physiology, combined with intricate diagnostic algorithms, creates substantial barriers to learning that traditional teaching methods may not fully address.

Among GenAI tools, ChatGPT has emerged as one of the most widely adopted platforms. Early studies demonstrate that it can answer United States Medical Licensing Examination (USMLE) style questions at a level comparable to third-year medical students [3] and achieve passing performance on the United Kingdom Medical Licensing Examination (UKMLA) [4]. While these findings suggest potential for medical education applications, existing research has predominantly focused on ChatGPT's ability to answer static, examination-style questions. There remains a critical gap in understanding whether ChatGPT can function as an interactive tutor—guiding learners through complex, stepwise clinical reasoning rather than simply providing answers.

Despite the availability of traditional teaching resources such as textbooks, lectures, and case-based discussions, many internal medicine residents struggle to apply nephrology's physiology-heavy principles in clinical settings [1, 2]. Studies indicate that limited exposure during training, coupled with perceived difficulty, contributes to declining interest in nephrology as a career subspecialty. Electrolyte disorders, particularly hyponatremia, exemplify this challenge: they require systematic diagnostic reasoning that integrates laboratory values, volume status assessment, and physiologic principles—a cognitive process that novice learners often find overwhelming. More effective and engaging teaching methods are needed to bridge this gap between theoretical knowledge and clinical application.

Retrieval-Augmented Generation (RAG) represents a promising approach to enhance ChatGPT's reliability for educational purposes. By grounding the model's responses in authoritative source material, in this case, a widely used nephrology reference text—RAG aims to reduce hallucinations and ensure content accuracy. However, whether RAG-enhanced ChatGPT can effectively teach complex medical topics through interactive dialogue, and whether such an approach improves learning outcomes, remains largely unexplored.

This pilot study evaluated whether a ChatGPT-based, RAG-enhanced interactive module on hyponatremia could support internal medicine resident education. We hypothesized that while short-term quiz performance might not differ significantly given the pilot sample size, the module would improve learner confidence, demonstrate acceptable usability, and provide a foundation for understanding GenAI's role in nephrology education.

2 Experimental Methods

2.1 Study Design and Setting

We conducted this pilot study in July 2024 at a single academic internal medicine residency program in the United States. The study employed a randomized controlled design with unequal allocation (2:1, intervention to control) to maximize exposure to the novel educational module while enabling robust usability assessment. This allocation strategy was chosen to prioritize feasibility testing and user experience data collection, which are critical outcomes for pilot educational interventions.

2.2 Participants and Recruitment

Fifty-five internal medicine residents across all postgraduate years were eligible to participate. The chief resident sent recruitment emails with pre-assigned randomized allocations over a two-week data collection period. Participation was voluntary, and no financial or academic incentives were offered. To minimize contamination between study groups, we closed recruitment one week prior to a scheduled didactic lecture on hyponatremia. This timing ensured that control participants would not be exposed to module content through informal peer discussion before completing their assessments.

2.3 Intervention Development and Design

The intervention consisted of a ChatGPT-based interactive learning module on hyponatremia that integrated structured prompting with Retrieval-Augmented Generation (RAG). We developed the module using content from pages 218–225 of *Pocket Nephrology* [5], a widely used nephrology reference text. The module's pedagogical framework was grounded in the stepwise diagnostic approach to hyponatremia taught in clinical nephrology: (1) assess serum osmolality to classify the disorder, (2) evaluate urine osmolality to determine renal water handling, and (3) measure urine sodium and assess volume status to identify the underlying etiology.

The module presented learners with case-based clinical vignettes followed by multiple-choice questions at each diagnostic step. After each response, ChatGPT provided immediate, personalized feedback that explained the correct reasoning pathway and addressed common misconceptions. The system adapted dynamically to learner responses, offering additional clarification or advancing to the next diagnostic step based on answer accuracy. RAG technology ensured that all generated explanations were grounded in the source text, thereby reducing the risk of hallucinations (fabricated information) and maintaining consistency with established nephrology guidelines.

We refined the module through iterative pilot testing with volunteer residents who were not included in the study sample. A board-certified nephrologist (T.S.) reviewed all module outputs to verify medical accuracy and appropriateness for the target learner level. These steps served as safeguards to improve reliability, minimize hallucinations, and ensure alignment with established nephrology teaching prior to deployment. It is to be noted that RAG reduces but does not eliminate the need for oversight. A sample system prompt, instructional flow, and example interactions are provided in **Appendix A**

2.4 Control Group

Control group participants completed only the 10-item quiz without prior access to the ChatGPT module. They were permitted to use traditional learning resources (textbooks, online references, notes) but were explicitly instructed not to use generative AI tools during the quiz. This design allowed us to compare the module's effectiveness against typical self-directed learning approaches. Control participants were not given access to the module as part of the study protocol.

2.5 Outcome Measures

Primary Outcome: The primary outcome was performance on a 10-item multiple-choice quiz assessing diagnostic reasoning in hyponatremia. Co-author T.S. (American Board of Internal Medicine–certified nephrologist) developed the quiz to align with the module's learning objectives and reflect clinically relevant scenarios encountered in nephrology practice. Questions were designed to test application of the stepwise diagnostic framework rather than rote memorization. The module was designed to target both improvement in diagnostic reasoning skills and learner confidence in managing hyponatremia, reflecting complementary educational objectives.

Secondary Outcomes. Secondary outcomes included: (1) self-reported confidence in managing nephrology-related conditions, measured before and after the intervention using a 5-point Likert scale. The confidence survey was administered immediately after completion of the module and

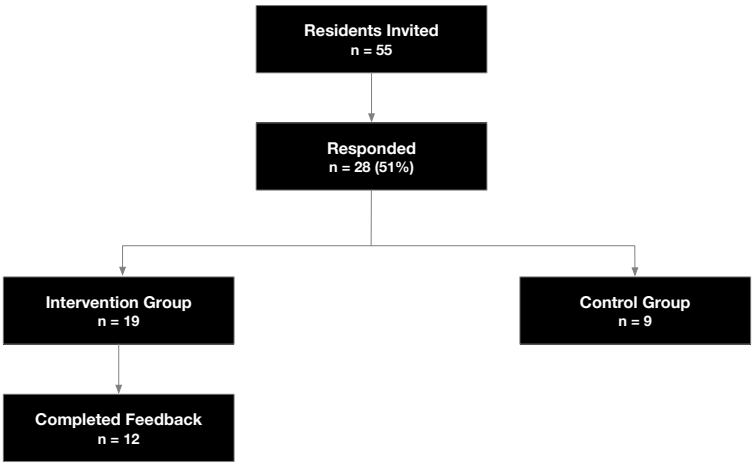

Figure 1: Study flow diagram showing participant enrollment, randomization, allocation, and completion of study activities

quiz, with participants retrospectively rating their pre-module confidence and current post-module confidence within the same instrument; (2) perceived usability of the module, assessed through Likert-scale items addressing ease of interaction, clarity of explanations, and appropriateness of content complexity; (3) overall satisfaction and likelihood of recommending the module to peers; and (4) qualitative feedback from open-ended survey questions exploring learner experiences and suggestions for improvement.

Expert Review: To assess content accuracy and identify potential errors, a board-certified nephrologist (T.S.) conducted an independent expert review of all ChatGPT session transcripts from intervention group participants. The reviewer evaluated outputs for medical accuracy, appropriateness of clinical reasoning, and presence of hallucinations or misleading information.

Instrument Validity. Although the quiz and survey instruments do not have published psychometric validation, their development and review by a content expert (T.S.) support face validity and content alignment for this pilot study. The quiz, feedback survey, and complete ChatGPT session transcripts are provided in Appendices 2–4.

2.6 Data Collection and Analysis

Intervention group participants accessed the ChatGPT module via a secure web-based interface with embedded RAG functionality, powered by GPT-4.0 (OpenAI). Both groups completed the quiz through an online survey platform. The intervention group additionally completed a post-module feedback survey. We summarized quiz scores using descriptive statistics (mean, standard deviation, range) and compared group differences using independent-samples t-tests with a two-tailed alpha of 0.05. We analyzed Likert-scale survey items using descriptive statistics and examined open-ended responses through thematic analysis to identify recurring themes and representative quotes. All data cleaning, statistical analyses, and figure generation were performed in Python using *Pandas*, *SciPy*, and *Matplotlib* libraries.

Given the pilot nature of this study, we did not conduct formal a priori power calculations. Sample

Cohort	N	Mean (SD)	Range
Control (quiz only)	9	7.44 (1.94)	4–9
Intervention (ChatGPT module)	19	7.47 (1.90)	3–9

Table 1: Quiz Performance by Cohort

size was determined by the number of available resident participants willing to engage with the study during the recruitment period.

2.7 Ethics Approval

This study was reviewed by the Institutional Review Board (IRB) of the University of Pennsylvania Institutional Review Board, and was determined to be exempt under 45 CFR 46.104(d)(2) (Exempt Category 2: Educational Tests, Surveys, Interviews, or Observation of Public Behavior) (Protocol #858691). Participation was voluntary. Informed consent was obtained electronically prior to participation, and all data were deidentified prior to analysis.

3 Experimental Results

3.1 Participant Flow and Characteristics

Of 55 eligible internal medicine residents invited to participate, 28 enrolled in the study, yielding a 51% participation rate (Figure 1). Following randomization, 19 participants were allocated to the intervention group (ChatGPT module + quiz) and 9 to the control group (quiz only). All 28 participants completed the quiz assessment. In the intervention group, 12 of 19 participants (63%) completed the post-module feedback survey.

3.2 Quiz Performance

Quiz performance was similar between groups, with no statistically significant difference detected (Table 1, Figure 2). The intervention group achieved a mean score of 7.47 out of 10 (SD = 1.90, range $3 - 9$), compared with 7.44 (SD = 1.94, range $4 - 9$) in the control group. An independent-samples *t-test* confirmed no significant difference between groups (t = -0.04, p = 0.97).

3.3 Learner Confidence and Perceived Usability

Among the 12 intervention participants who completed the feedback survey (63% response rate), self-reported confidence in managing nephrology-related conditions increased after using the module. Pre-module confidence averaged 2.67 on a 5-point Likert scale (SD = 1.23), increasing to 3.50 post-module (SD = 0.80), representing a gain of 0.83 points or 31% improvement. It should be noted that confidence ratings were collected retrospectively for the intervention group only, and no comparable confidence measure was obtained from the control group.

Usability ratings were consistently high across multiple dimensions (Table 2). Participants rated ease of interaction highest (mean 4.58, SD = 0.51), followed by clarity of explanations (mean 4.17, SD = 0.83). The appropriateness of content complexity received a mean rating of 2.75 (SD = 0.62)

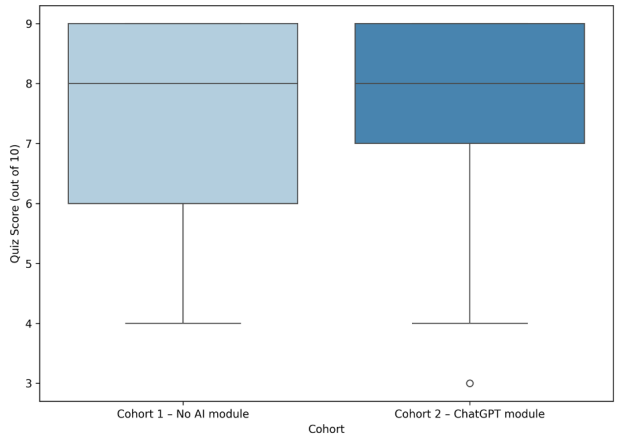

Figure 2: Distribution of Quiz Scores by Cohort

Item	Mean (SD)
Confidence before module	2.67 (1.23)
Confidence after module	3.50 (0.80)
Ease of interaction	4.58 (0.51)
Clarity of explanations	4.17 (0.83)
Complexity appropriateness	2.75 (0.62)
Would recommend	4.17 (0.72)
Likely to use AI again	4.00 (0.85)

Table 2: Summary of User Feedback on AI Module (n = 12)

on a scale where lower scores indicate content well-matched to learner level. Most participants indicated strong likelihood of recommending the module to peers (mean 4.17, SD = 0.72) and anticipated future use of similar AI-based learning tools (mean 4.00, SD = 0.85).

3.4 Qualitative Feedback

Thematic analysis of open-ended survey responses revealed three dominant themes: clarity and structure, interactivity, and convenience (Table 3). Residents consistently praised the module's systematic, step-by-step approach to diagnostic reasoning, noting that ChatGPT's structured explanations made complex material more accessible. Many emphasized the value of receiving immediate, personalized feedback and the ability to ask clarifying questions in real time. Participants also appreciated the convenience of having all learning materials consolidated in a single, user-friendly interface. Open-ended responses were organized into themes using an inductive descriptive approach. Given the small number of responses, this analysis was exploratory and not intended to achieve thematic saturation.

Suggestions for improvement included adding visual summary graphics, such as diagnostic flowcharts, and providing more detailed explanations when participants selected incorrect answers—particularly clarifying why alternative options were incorrect rather than simply confirming the correct choice.

Theme	Representative Quote
Clarity & Structure	• "I liked the step-by-step approach." • "The systematic way ChatGPT explains the subject made it easier to understand."
Interactivity	• "Being able to ask specific questions and get immediate feedback helped reinforce learning."
Convenience & Usability	• "Everything was on one screen and location. The layout was very user-friendly." • "Good summary tables and step-by-step instructions."

Table 3: Sample Themes from Qualitative Feedback

3.5 Expert Review of ChatGPT Outputs

A board-certified nephrologist (T.S.) independently reviewed 14 complete ChatGPT session transcripts from intervention participants to assess content accuracy and identify potential errors. The majority of outputs were medically accurate and clinically appropriate for the target audience. However, several noteworthy errors were identified that persisted despite the use of RAG technology. Specific errors included: (1) misinterpretation of serum osmolality values in one case, leading to incorrect initial classification; (2) reinforcement of an incorrect learner response rather than providing corrective feedback; (3) minor discrepancies in cited laboratory reference ranges compared to the source text; and (4) spontaneous generation of a new quiz question not included in the original prompt. In one complex clinical scenario involving concurrent hyperglycemia and hyponatremia, ChatGPT correctly identified the mechanism of pseudohyponatremia but subsequently misclassified the patient's overall tonicity status.

Importantly, none of the identified errors were judged to pose patient safety concerns or to recommend clinically dangerous management. The errors were primarily classificatory or pedagogical in nature.

4 Results Discussion

4.1 Principal Findings

This pilot study demonstrates that a ChatGPT-based, RAG-enhanced interactive module on hyponatremia is feasible to implement in a residency training program and well-received by learners. While the intervention did not produce statistically significant improvements in short-term quiz performance compared to traditional self-study, it was associated with meaningful increases in learner confidence and high ratings for usability and educational value.

These findings suggest that generative AI may offer value in medical education through dimensions beyond immediate knowledge acquisition. Specifically, the module appeared to enhance learner engagement, provide accessible explanations of complex material, and improve self-efficacy—all important outcomes in graduate medical education that may not be captured by brief knowledge

What was the most useful part of the experience?	**What suggestions do you have to improve the educational tool?**
Everything was on one screen and location. There was not information overload which can happen. It was interactive and you could ask questions. I felt the information was useful. You could even take screenshots and add to your phone for future reference.	This was great — I think we focused mostly on diagnostics, which was fine. Future modules could focus on treatment.
Good summary tables and step-by-step instructions!	A summary graphic at the end would be great.
Being able to ask specific questions and getting a specific answer for your problem	It would be nice if the AI could provide a source or paper to link with explanations (to ensure information is valid).
The systematic way ChatGPT explains the subject	
I like the step-by-step approach	
	I would recommend more thorough answer explanations and wrong-answer analysis.
Very structured approach	More of this please.
Will appreciate feedback on questions — also appreciated the review of normal cut-offs	
The ChatGPT learning module was clear and concise. Appreciate the summary and explanation after each answer.	None.
Short and concise but great educational value	More complicated cases.

Table 4: Full Open-Ended Survey Responses from Intervention Participants

assessments.

4.2 Comparison with Existing Literature

Our results align with a growing body of evidence demonstrating that ChatGPT and similar generative AI tools can increase learner satisfaction and self-reported understanding, even when objective performance metrics remain unchanged. Moskovich and Rozani [6] similarly found that health professional students reported high satisfaction and confidence when using ChatGPT for educational purposes, despite mixed effects on objective testing. A recent meta-analysis by Wang and Fan [7] noted positive effects on learning perception and higher-order thinking skills across multiple educational contexts, supporting the role of AI as a value-enhancing complement to traditional instruction. These converging findings suggest that AI-based modules may be most appropriately positioned as supplements to, rather than replacements for, traditional teaching methods, particularly in complex domains like nephrology where learner confidence and engagement are critical.

The lack of significant difference in quiz performance merits careful interpretation. Several factors may contribute to this finding beyond the intervention's effectiveness. First, the study was underpowered to detect small to moderate effect sizes given the limited sample size (n=28). Post-hoc power analysis suggests that detecting a clinically meaningful improvement of 1.5 points (15% gain) would require approximately 90 participants per group with 80% power at $\alpha = 0.05$. Second, the control group had access to open resources during the quiz, which may have reduced observable differences. Third, a single learning session may be insufficient to produce measurable knowledge gains detectable by a 10-item assessment—multiple exposures or longer-term follow-up might reveal delayed or cumulative effects.

4.3 Persistent Limitations of RAG Technology

Despite the use of RAG to ground ChatGPT outputs in authoritative source material, expert review confirmed that errors persisted across multiple dimensions. These findings are consistent with recent literature documenting ongoing challenges with RAG-based systems in medical applications. Zhang and Zhang [8] note that hallucinations can occur during both the retrieval phase (selecting irrelevant or incorrect source passages) and the generation phase (misinterpreting or extrapolating beyond retrieved content). Gilson et al. [9] found that domain-specific RAG systems in ophthalmology still produced 18.8% minor hallucinations and 26.7% additional inaccuracies when answering consumer health questions.

Particularly concerning for educational applications, Wong et al. [10] and Pham and Vo [11] caution that even factually accurate but decontextualized outputs may mislead learners by omitting important caveats, oversimplifying complex concepts, or failing to acknowledge areas of clinical uncertainty. In our study, while no outputs posed direct safety risks, the occasional reinforcement of incorrect learner responses represents a meaningful pedagogical failure that could entrench misconceptions.

These findings underscore a critical point for medical educators: RAG technology reduces but does not eliminate the need for expert oversight. Current AI systems, even when augmented with curated source material, cannot yet be deployed as autonomous tutors in medical education without faculty validation and monitoring.

4.4 Limitations

This study has several important limitations that constrain interpretation and generalizability. The modest sample size (n=28) and 51% participation rate limited statistical power to detect performance differences and may have introduced selection bias if more motivated or tech-savvy residents chose to participate. The 63% survey response rate among intervention participants raises the possibility of response bias in usability and satisfaction metrics. Confidence was measured only in the intervention group using a retrospective survey, limiting direct comparison with the control group and warranting cautious interpretation of the confidence-related findings.

The inclusion of residents across all postgraduate years likely introduced heterogeneity in baseline nephrology knowledge and learning needs, potentially obscuring intervention effects. The open-resource design of the control group, while reflecting realistic self-study conditions, may have diluted measurable differences by allowing participants to access high-quality reference materials during the quiz.

Findings may not generalize beyond this single academic institution, the specific topic of hyponatremia, or the particular implementation of RAG technology used in this study. Different nephrology topics, alternative AI architectures, or varied pedagogical approaches might yield different results. Additionally, users without ChatGPT Plus subscriptions face usage limits that could restrict engagement with similar modules, limiting scalability.

Finally, this study assessed only immediate outcomes. We did not measure knowledge retention at later timepoints, transfer of learning to clinical practice, or long-term effects on diagnostic reasoning skills—outcomes that may be more relevant to the ultimate educational value of AI-enhanced learning.

4.5 Implications for Medical Education

Despite these limitations, this pilot study contributes meaningful evidence to support the integration of generative AI as an educational adjunct in graduate medical education. For residency program directors and medical educators, several practical implications emerge.

First, AI-based modules like the one tested here can provide dynamic, interactive platforms for teaching complex clinical reasoning in a learner-centered manner. The high usability ratings and positive qualitative feedback suggest that residents find such tools engaging and accessible, which may be particularly valuable for traditionally challenging content areas like nephrology.

Second, the improvement in learner confidence, even without corresponding gains in quiz performance, suggests that AI modules may help address affective barriers to learning—reducing intimidation, increasing self-efficacy, and promoting willingness to engage with difficult material. These outcomes have intrinsic value in medical education and may facilitate future learning even if not immediately measurable.

Third, the persistent errors identified during expert review highlight that current AI systems require faculty oversight and quality assurance. Educators should not view these tools as autonomous teaching replacements but rather as supervised supplements that extend teaching capacity and provide additional practice opportunities.

4.6 Future Improvements

Future studies should address this pilot's limitations through larger, multi-institutional samples adequately powered to detect meaningful performance differences. Longitudinal designs with delayed post-testing would clarify whether AI-enhanced learning produces durable knowledge gains or merely temporary confidence boosts. Comparative effectiveness research examining different AI architectures, RAG implementations, and pedagogical approaches could identify optimal design features. Additionally, research is needed on: (1) the optimal balance between AI-guided and traditional instruction; (2) which specific topics or learning objectives are best suited to AI-enhanced modules; (3) methods for real-time detection and correction of AI errors during educational sessions; and (4) whether repeated exposure to AI tutoring systems improves clinical reasoning skills observable in authentic practice settings.

5 Conclusion

This paper examined the use of a ChatGPT-based, retrieval-augmented generation (RAG) interactive module on hyponatremia. The findings indicate that the approach is feasible and well received by residents and is associated with improved learner confidence, although it did not significantly improve short-term quiz performance. High usability ratings and positive qualitative feedback suggest that generative AI has potential as a supplemental educational tool in nephrology and other complex medical specialties. However, the errors despite RAG implementation underscores the limitations of current AI systems and the need for expert oversight. As generative AI continues to evolve, its supervised integration into residency curricula may help modernize teaching, enhance learner engagement, and extend educational capacity in resource-constrained settings.

Appendix A: Prompt used in the intervention module.

Hyponatremia Case-Based Learning Module

You are a clinical nephrologist and educator designing a resident-level, case-based learning module on hyponatremia, using a stepwise diagnostic reasoning approach modeled on MKSAP-style vignettes. Use **Pocket Medicine** (not Pocket Pearls) as the source of clinical logic and knowledge.

Module Structure

– Generate all cases as brief clinical scenarios.

– Provide multiple-choice questions for each case.

– Allow learners to answer one question at a time.

Step 1: Serum Osmolality - Classify Hyponatremia

– Teach how serum osmolality differentiates true hypotonic vs hypertonic and isotonic hyponatremia.

– Include normal serum osmolality values as a reference.

– Present 2 clinical cases.

– Learners classify hyponatremia; provide explanations or hints after each response.

Step 2: Urine Osmolality - Assess ADH Activity

– Apply only to hypotonic hyponatremia cases.

– Introduce urine osmolality to evaluate ADH activity.

– Present 2 cases:

  * Low urine osmolality (e.g., polydipsia, beer potomania, tea and toast).

  * High urine osmolality (do not provide final diagnosis).

– Provide full explanation or targeted hint as feedback.

Step 3: Urine Sodium - Assess Volume Status (When ADH is Inappropriately Elevated)

– Present 2 cases with full labs.

– Use clinical signs and symptoms (not volume labels) to cue hypovolemia vs SIADH.

– One case with low urine sodium, one case with high urine sodium.

– Do not comment on ADH appropriateness when evaluating urine sodium.

– Key point: SIADH is a diagnosis of exclusion; adrenal insufficiency and hypothyroidism need to be ruled out first.

– Learners apply stepwise reasoning; provide feedback after each response.

Step 4: Summary Table

– Generate a full table summarizing the diagnostic framework:

  * Serum osmolality
  * Urine osmolality
  * Urine sodium
  * All related diagnoses

Tone: Practical, senior resident/attending-level.

References

[1] S. G. Tangye, C. S. Ma, R. Brink, and E. K. Deenick. The good, the bad and the ugly—TFH cells in human health and disease. *Nature Reviews Immunology*, 13(6):412–426, 2013.

[2] D. Nair et al. Perceptions of nephrology among medical students and internal medicine residents: A national survey among institutions with nephrology exposure. *BMC Nephrology*, 20(1):146, 2019.

[3] T. H. Kung et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, 2023.

[4] O. Casals-Farré et al. Assessing ChatGPT 4.0's capabilities in the united kingdom medical licensing examination (UKMLA): A robust categorical analysis. *Scientific Reports*, 15(1):13031, 2025.

[5] W. Ahn and J. Radhakrishnan. *Pocket Nephrology*. Lippincott Williams & Wilkins, 2019.

[6] L. Moskovich and V. Rozani. Health profession students' perceptions of ChatGPT in healthcare and education: Insights from a mixed-methods study. *BMC Medical Education*, 25(1):98, 2025.

[7] J. Wang and W. Fan. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1):1–21, 2025.

[8] W. Zhang and J. Zhang. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5):856, 2025.

[9] A. Gilson et al. Enhancing large language models with domain-specific retrieval-augmented generation: A case study on long-form consumer health question answering in ophthalmology. arXiv:2409.13902, 2024.

[10] L. Wong et al. Retrieval-augmented systems can be dangerous medical communicators. arXiv preprint arXiv:2502.14898, 2025.

[11] D. K. Pham and B. Q. Vo. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. arXiv preprint arXiv:2408.13808, 2024.