

 Emre Cankaya,  Ezgi Coskun Yenigun,
 Simal Koksall Cevher,  Muge Uzerk Kibar,
 Meryem Keles,  Fatih Dede

Ankara Bilkent City Hospital, Department of Nephrology,
Ankara, Türkiye

Received: 24 February, 2026

Accepted: 26 March, 2026

Published: 13 October, 2026

Corresponding Author: Emre Cankaya, Ankara Bilkent
City Hospital, Department of Nephrology, Ankara, Türkiye
Email: dremrecnky@gmail.com

CITATION

Cankaya E, Coskun Yenigun E, Koksall Cevher S. Clinical reasoning in nephrology: Comparing GPT-4 with junior fellows. *Atlantic J Med Sci Res.* 2026;6(4):436-43. DOI: 10.5455/atjmed.2026.02.016

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.



Clinical reasoning in nephrology: Comparing GPT-4 with junior fellows

Abstract

Aim: This study compared the clinical decision-making abilities of junior nephrology fellows against two Generative Pre-trained Transformer 4 (GPT-4)-based artificial intelligence (AI) models—one standard and one specifically trained with nephrology guidelines—using simulated clinical scenarios.

Materials and Methods: Eleven nephrology fellows with two years or less of experience from four academic centers participated in the study. The performance of the participants and the artificial intelligence models (TrainedGPT, a custom GPT-4 model configured with a nephrology guideline-based internal repository, and GPT-4o) was evaluated based on their written responses to nine simulated nephrology case scenarios. Responses were scored by two blinded expert nephrologists using a 5-point Likert scale across three main stages: prediagnosis, diagnosis, and treatment.

Results: Both TrainedGPT (13.2 [interquartile range (IQR): 9.9 – 13.7]) and GPT-4o (12.8 [IQR: 8.6 – 13.7]) models significantly outperformed the fellow group (8.6 [IQR: 7.0 – 10.6]) ($p = 0.001$ and $p = 0.016$, respectively). This superiority was maintained throughout all stages of prediagnosis, diagnosis, and especially treatment. There was no significant difference between the two artificial intelligence models ($p = 0.690$).

Conclusion: These pilot findings suggest that GPT-4-based AI models may serve as useful adjuncts in nephrology education and structured clinical support, outperforming junior fellows in standardized, evidence-based decision-making tasks. Given the exploratory design and small sample size, these results should be interpreted cautiously with respect to generalizability. The lack of a significant difference between TrainedGPT and GPT-4o suggests that the foundational medical knowledge of current large language models may already be robust for standard nephrology scenarios.

Keywords: Artificial intelligence; clinical decision-making; medical education; nephrology; clinical competence

INTRODUCTION

Artificial intelligence (AI), which aims to mimic human cognitive abilities, is poised to introduce revolutionary changes to the field of medicine. In 2016, the largest share of investments in AI research was directed towards healthcare applications, more than any other sector (1,2). The approval of Arterys' deep learning-based application by the U.S. Food and Drug Administration (FDA) in 2017 was considered an important step in the use of AI in medicine (3,4). Following this approval, there was a significant increase in the number of machine learning-enabled medical devices. Advanced AI methods such as neural networks and deep learning have become essential tools widely used in a variety of clinical processes, including diagnosis, treatment planning, patient

monitoring, and data management (3,5,6). Today, AI is becoming an increasingly promising tool for healthcare services (6).

In recent years, advances in deep learning have accelerated the development of large language models (LLMs) (3,7). These models, trained on large text datasets, can produce human-like texts and meaningful responses due to their ability to learn and interpret complex semantic relationships. This ability has allowed them to play an important role in healthcare, especially in medicine, by helping decision-making processes in clinical scenarios involving uncertainty (7,8). A systematic review has shown that generative LLMs have the potential to improve the quality and readability of documents such as discharge notes and clinical summaries (7).

Clinical studies clearly demonstrate the potential of AI in medical practice. A study on acute ischemic stroke management has shown that the Generative Pre-trained Transformer 4 (GPT-4) model can provide treatment recommendations on critical issues such as thrombolysis and endovascular treatment using real patient data. These recommendations were found to be highly consistent with stroke experts' assessments and actual clinical treatment decisions applied to patients (9). Furthermore, it has been shown that the contribution of ChatGPT can improve the accuracy of a decision support system predicting the risk of breast cancer recurrence (10).

The GPT-4 model trained in rheumatology achieved higher clinical judgment scores than the fellow physicians in ten clinical cases. 60% of respondents indicated that they would prefer to have access to a similar AI program to assist with patient assessments. Fellows also reported feeling more competent when working with such a system, which may reduce fatigue and burnout (3). LLMs can sometimes produce contextually irrelevant or unrealistic outputs; this phenomenon is known in the literature as "hallucination". Therefore, human supervision and careful evaluation are necessary, especially in clinical decision-making processes (7,11,12).

AI applications are rapidly becoming widespread in the field of nephrology. Some systems developed for anemia management in haemodialysis patients have been shown to produce effective results. Especially the recommendations regarding erythropoiesis-stimulating agent (ESA) doses contribute significantly to keeping hemoglobin levels within the desired range (13,14). While these approaches help to carry out the treatment in a more balanced manner, they also have the potential to reduce the economic burden on the healthcare system. One study reported that ChatGPT-4.0 was able to determine both the need for consultation and subspecialty referral with a very high accuracy in 100 simulated nephrology cases (11).

LLMs, particularly GPT-4, have demonstrated significant success in medical licensing examinations like the United States Medical Licensing Examination (USMLE) and in specialized nephrology assessments such as the Nephrology Self-Assessment Program (NephSAP) and Kidney Self-Assessment Program (KSAP) (5,12,15-21). However, these knowledge-based exams, which sometimes reveal inconsistent performance across different subtopics, do not fully evaluate applied clinical reasoning (5,12,21). Consequently, few studies have directly compared the clinical reasoning performance of LLMs against physicians-in-training using complex, realistic case scenarios. The aim of this study is to compare the clinical decision-making abilities of nephrology fellows in the early stages of their training with those of trained and untrained AI models, using nephrology case scenarios.

MATERIALS AND METHODS

Eleven nephrology fellows with less than two years of specialty training from four academic medical centers in Ankara (Hacettepe University, Gazi University, Etlik City Hospital, and Bilkent City Hospital) voluntarily participated in the study.

A total of nine original clinical case scenarios were designed by expert nephrologists to assess nephrology knowledge and clinical reasoning skills. These scenarios were created without using actual patient data. All case scenarios and related questions were prepared in Turkish based on frequent scenarios in outpatient nephrology. The cases covered basic nephrology topics such as the diagnostic approach to hematuria and proteinuria, acute kidney injury (AKI) cases requiring and not requiring hospitalization, hyponatremia, anemia and mineral and bone diseases in chronic kidney disease (CKD), peritonitis, and renal replacement therapies in CKD. These vignettes employed explicit directives to eliminate ambiguity and were independently validated by an expert reviewer. Each case assessment was designed to measure a broad range of clinical skills. These skills included selecting initial tests for the diagnostic process, establishing preliminary and definitive diagnoses, and assessing the need for additional testing. During the treatment phase, the assessment measured the ability to formulate a treatment plan that included medication dosages, side effect management, and adherence to clinical guidelines. In addition, clinical decisions such as emergency dialysis, kidney biopsy, and selection of appropriate renal replacement therapy were evaluated.

A survey was designed to evaluate participants' knowledge. No personal or demographic information was collected; participants were advised to use a pseudonym as their ID. Once submitted, responses could not be modified. Each case consisted of six to 10 open-ended questions, and participants were instructed to provide their responses in text form. The same survey protocol was applied to both nephrology fellows and GPT-4 models, and both groups responded to all nine cases.

Two distinct AI models based on OpenAI's GPT-4 architecture were utilized in this study. The first, serving as the untrained control, was the standard GPT-4o model used without any supplementary nephrology-specific internal repository. The second, referred to as "TrainedGPT," was a custom GPT-4 model configured with a guideline-based internal knowledge repository developed specifically for this study. This repository was built from study-specific documents derived from key international guidelines (Kidney Disease: Improving Global Outcomes [KDIGO], International Society for Peritoneal Dialysis [ISPD], European Renal Best Practice [ERBP]), and tailored to address all clinical questions within the cases. We confirmed that this curated information was sufficient for all tasks. The model was explicitly instructed to rely solely on this internal repository. Any supplemental PubMed-linked material available to the model consisted of a predefined, static PubMed-derived content set compiled before study execution, rather than live internet access or real-time querying during case completion.

Two expert nephrologists (EY and EC) independently evaluated the written responses. The evaluators were blinded to the study hypothesis and the respondent's identity. The responses were scored using a 5-point Likert scale (1: Strongly Disagree, 5: Strongly Agree). The score for each question was calculated as the average of the two evaluators' ratings. In cases of significant discrepancy between the two evaluators (a score difference

of >1 point), the response was reassessed in a joint review to reach a consensus score. The responses were classified into three main domains for each question: prediagnosis, diagnosis, and treatment, and comparative analyses were performed based on the average scores for each category. While all nine cases were assessed for diagnosis and treatment, only six contained questions relevant to the prediagnosis stage. Consequently, the analysis for the prediagnosis domain was based on these six cases. These responses were then compared with the responses generated by AI models.

For each case response, question-level ratings were first averaged within each domain. The average score for each participant in each stage was then calculated separately. The total score was obtained by summing the available domain scores rather than averaging them across domains. Accordingly, the maximum possible total score was 15 for cases containing prediagnosis, diagnosis, and treatment questions, and 10 for cases without a prediagnosis component. Because all groups responded to the same case set, these totals remained directly comparable across groups. Median and interquartile ranges (IQR) were calculated for all stages for each GPT-4 model and fellow group.

Statistical Analyses

For statistical analysis, Jamovi (v2.3) statistical software was used. Due to the non-normal distribution of the data, group comparisons were analyzed using non-parametric tests. All continuous variables were presented as median (IQR), and minimum–maximum values were additionally reported in the tables.

The Mann–Whitney U test was used for two-group comparisons, and the Kruskal–Wallis test was used for three-group comparisons;

when a significant difference was detected, Bonferroni correction was applied for pairwise comparisons. The significance level was set at $p < 0.05$ for all tests. Box-violin plots and histograms were generated for each phase to visually represent the obtained scores. These plots were used to compare intra-group variation, score distribution, and model stability.

Ethical Considerations

This study was deemed exempt from formal ethical committee review because it was based exclusively on simulated clinical scenarios and anonymous written responses and did not involve real patients, patient records, biological materials, clinical interventions, or personally identifiable participant data. Participants were evaluated solely on standardized hypothetical case responses and were advised to use pseudonymous identifiers.

RESULTS

The responses of two AI models—TrainedGPT and GPT-4o—and 11 nephrology fellows were comparatively evaluated across nine nephrology cases comprising a total of 75 questions that covered the prediagnosis, diagnosis, and treatment stages.

Analysis of the total performance scores showed that both AI models received significantly higher scores compared to the fellow group ($p < 0.001$). The median total score for TrainedGPT was 13.2 (IQR: 9.9–13.7), while that for GPT-4o was 12.8 (IQR: 8.6 – 13.7), and the median score for the fellow group was 8.6 (IQR: 7.0 – 10.6) (Table 1). The pairwise comparisons showed that both TrainedGPT ($p = 0.001$) and GPT-4o ($p = 0.016$) demonstrated significant superiority over the fellow group, while no statistically significant difference was found between the two AI models ($p = 0.690$) (Table 2).

Table 1. Comparison of TrainedGPT, GPT-4o, and human performance in simulated nephrology cases

Variables	TrainedGPT (N = 9)	GPT-4o (N = 9)	Human (N = 99)	Total (N = 117)	p value
Prediagnosis					
N-Miss	3	3	33	39	0.003
Median (IQR)	4.2 (4.0 – 4.4)	4.2 (4.0 – 4.8)	3.5 (2.7 – 4.0)	3.5 (2.9 – 4.2)	
Min–Max	3.2 – 5.0	3.8 – 5.0	1.5 – 5.0	1.5 – 5.0	
Diagnosis					
Median (IQR)	4.5 (4.5 – 5.0)	4.1 (4.0 – 4.8)	3.5 (3.0 – 4.0)	3.5 (3.0 – 4.3)	< 0.001
Min–Max	4.2 – 5.0	3.5 – 5.0	1.7 – 5.0	1.7 – 5.0	
Treatment					
Median (IQR)	4.8 (4.6 – 4.9)	4.3 (3.6 – 4.8)	3.1 (2.6 – 3.6)	3.2 (2.8 – 3.9)	< 0.001
Min–Max	3.7 – 5.0	3.0 – 5.0	1.0 – 4.5	1.0 – 5.0	
Total					
Median (IQR)	13.2 (9.9 – 13.7)	12.8 (8.6 – 13.7)	8.6 (7.0 – 10.6)	9.3 (7.4 – 11.1)	< 0.001
Min–Max	8.2 – 13.9	7.0 – 14.6	4.4 – 13.5	4.4 – 14.6	

Abbreviations: GPT-4o, standard GPT-4o model; IQR, interquartile range; N, number of case responses; N-Miss, number of missing observations in the prediagnosis domain because 3 of the 9 scenarios did not include prediagnosis questions; TrainedGPT, custom GPT-4 model configured with nephrology-specific guideline-based content.

Table 2. Comparison of TrainedGPT vs GPT-4o across nephrology domains

Variables	TrainedGPT (N = 9)	GPT-4o (N = 9)	Total (N = 18)	p value
Prediagnosis				
Median (IQR)	4.2 (4.0 – 4.4)	4.2 (4.0 – 4.8)	4.2 (4.0 – 4.6)	0.871
Min–Max	3.2 – 5.0	3.8 – 5.0	3.2 – 5.0	
Diagnosis				
Median (IQR)	4.5 (4.5 – 5.0)	4.1 (4.0 – 4.8)	4.5 (4.1 – 5.0)	0.138
Min–Max	4.2 – 5.0	3.5 – 5.0	3.5 – 5.0	
Treatment				
Median (IQR)	4.8 (4.6 – 4.9)	4.3 (3.6 – 4.8)	4.6 (3.9 – 4.9)	0.268
Min–Max	3.7 – 5.0	3.0 – 5.0	3.0 – 5.0	
Total				
Median (IQR)	13.2 (9.9 – 13.7)	12.8 (8.6 – 13.7)	13.0 (9.7 – 13.7)	0.690
Min–Max	8.2 – 13.9	7.0 – 14.6	7.0 – 14.6	
Abbreviations: GPT-4o, standard GPT-4o model; IQR, interquartile range; N, number of case responses; TrainedGPT, custom GPT-4 model configured with nephrology-specific guideline-based content.				

A similar trend was observed when the sub-stages of the evaluation were examined. In the prediagnosis phase, the median scores of the TrainedGPT and GPT-4o models were 4.2 for both groups (IQR: 4.0 – 4.4 and 4.0 – 4.8, respectively), while the median for the fellow group was 3.5 (IQR: 2.7 – 4.0) ($p = 0.003$). TrainedGPT ($p = 0.024$) and GPT-4o ($p = 0.006$) each demonstrated significantly higher performance than fellows, but no significant difference was found between the two models ($p = 0.871$) (Table 2, Table 3, and Table 4).

In the diagnosis phase, the median scores were 4.5 (IQR: 4.5 – 5.0) for TrainedGPT, 4.1 (IQR: 4.0 – 4.8) for GPT-4o, and 3.5 (IQR: 3.0 – 4.0) for the fellow group, a difference that was statistically significant ($p < 0.001$). At this stage, both TrainedGPT and GPT-4o showed significant superiority over the fellow group separately ($p < 0.001$ for TrainedGPT, $p = 0.002$ for GPT-4o). However, no significant difference was

found between the scores of the two models ($p = 0.138$) (Table 2, Table 3, and Table 4).

The performance difference between the AI models and the fellows was most marked in the treatment stage. At this stage, the median score for TrainedGPT was 4.8 (IQR: 4.6 – 4.9), for GPT-4o it was 4.3 (IQR: 3.6 – 4.8), and for the fellow group it was 3.1 (IQR: 2.6 – 3.6) ($p < 0.001$). While both AI models achieved significantly higher scores than the fellow group ($p < 0.001$ for TrainedGPT, $p = 0.002$ for GPT-4o), there was no statistically significant difference between the two models ($p = 0.268$) (Table 2, Table 3, and Table 4).

The score distributions, as illustrated by histograms (Figure 1) and violin plots (Figure 2), revealed a wider range and greater inter-individual variability in the scores of the fellow group compared to the more concentrated distributions observed for the two AI models.

Table 3. Comparison of TrainedGPT vs human performance across nephrology domains

Variables	TrainedGPT (N = 9)	Human (N = 99)	Total (N = 108)	p value
Prediagnosis				
Median (IQR)	4.2 (4.0 – 4.4)	3.5 (2.7 – 4.0)	3.5 (2.7 – 4.0)	0.024
Min–Max	3.2 – 5.0	1.5 – 5.0	1.5 – 5.0	
Diagnosis				
Median (IQR)	4.5 (4.5 – 5.0)	3.5 (3.0 – 4.0)	3.5 (3.0 – 4.2)	< 0.001
Min–Max	4.2 – 5.0	1.7 – 5.0	1.7 – 5.0	
Treatment				
Median (IQR)	4.8 (4.6 – 4.9)	3.1 (2.6 – 3.6)	3.2 (2.7 – 3.8)	< 0.001
Min–Max	3.7 – 5.0	1.0 – 4.5	1.0 – 5.0	
Total				
Median (IQR)	13.2 (9.9 – 13.7)	8.6 (7.0 – 10.6)	9.1 (7.3 – 10.9)	0.001
Min–Max	8.2 – 13.9	4.4 – 13.5	4.4 – 13.9	
Abbreviations: GPT-4o, standard GPT-4o model; IQR, interquartile range; N, number of case responses; TrainedGPT, custom GPT-4 model configured with nephrology-specific guideline-based content.				

Table 4. Comparison of GPT-4o vs human performance across nephrology domains

Variables	GPT-4o (N = 9)	Human (N = 99)	Total (N = 108)	p value
Prediagnosis				
Median (IQR)	4.2 (4.0–4.8)	3.5 (2.7–4.0)	3.5 (2.7–4.0)	0.006
Min–Max	3.8–5.0	1.5–5.0	1.5–5.0	
Diagnosis				
Median (IQR)	4.1 (4.0–4.8)	3.5 (3.0–4.0)	3.5 (3.0–4.0)	0.002
Min–Max	3.5–5.0	1.7–5.0	1.7–5.0	
Treatment				
Median (IQR)	4.3 (3.6–4.8)	3.1 (2.6–3.6)	3.2 (2.7–3.7)	0.002
Min–Max	3.0–5.0	1.0–4.5	1.0–5.0	
Total				
Median (IQR)	12.8 (8.6–13.7)	8.6 (7.0–10.6)	8.7 (7.2–10.8)	0.016
Min–Max	7.0–14.6	4.4–13.5	4.4–14.6	

IQR: interquartile range, N: number of case responses, GPT-4o: standard GPT-4o model

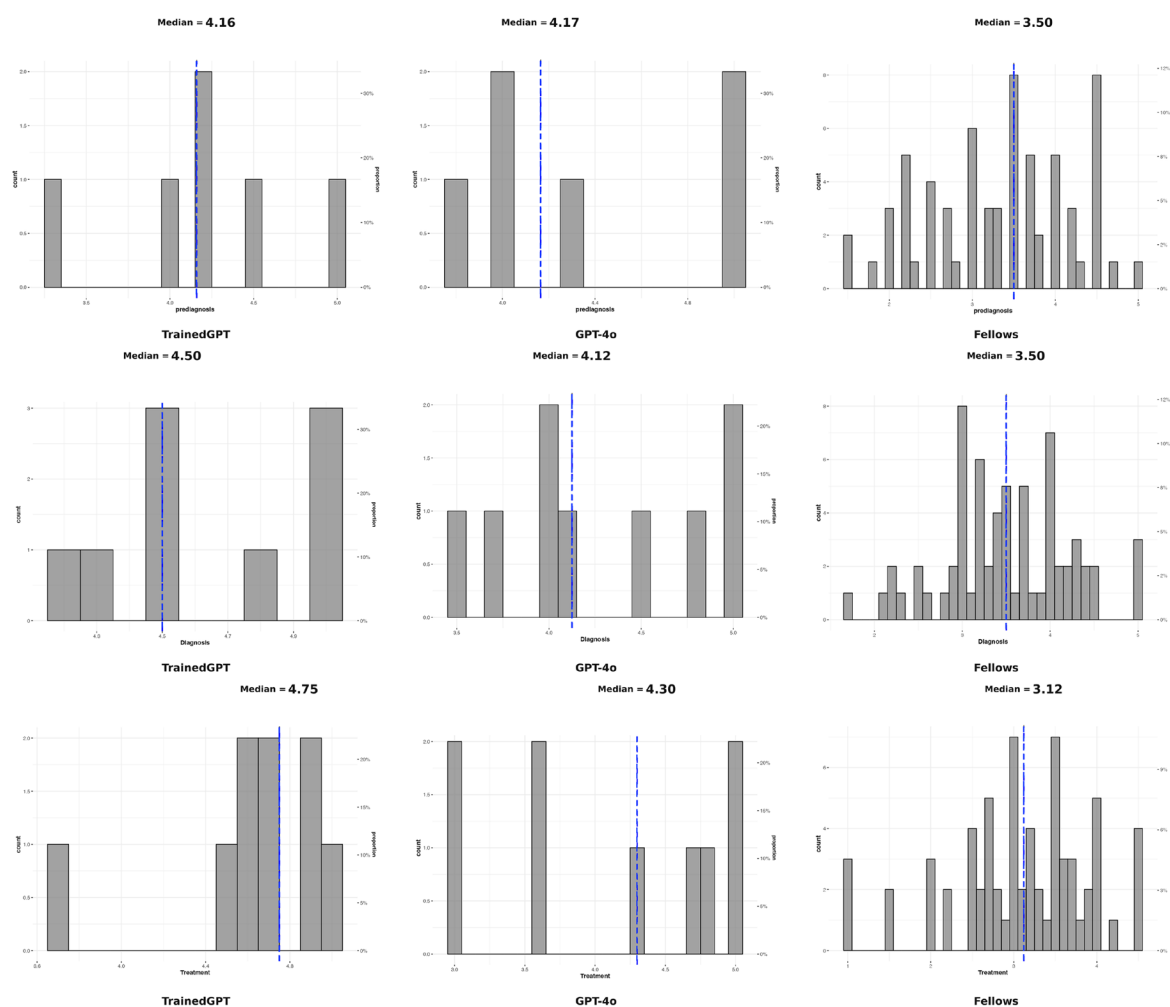


Figure 1. Histogram of case-level clinical reasoning scores by respondent group and clinical stage. The figure displays score distributions for TrainedGPT, GPT-4o, and junior nephrology fellows in the prediagnosis, diagnosis, and treatment domains. In each histogram, the vertical dashed line indicates the median score for the corresponding respondent group, consistent with the non-parametric analyses used in the study. Sample sizes were $n = 6$ per AI model and $n = 66$ for fellows in the prediagnosis domain, and $n = 9$ per AI model and $n = 99$ for fellows in the diagnosis and treatment domains. Abbreviations: GPT-4o, standard GPT-4o model; TrainedGPT, custom GPT-4 model configured with nephrology guideline-based content.

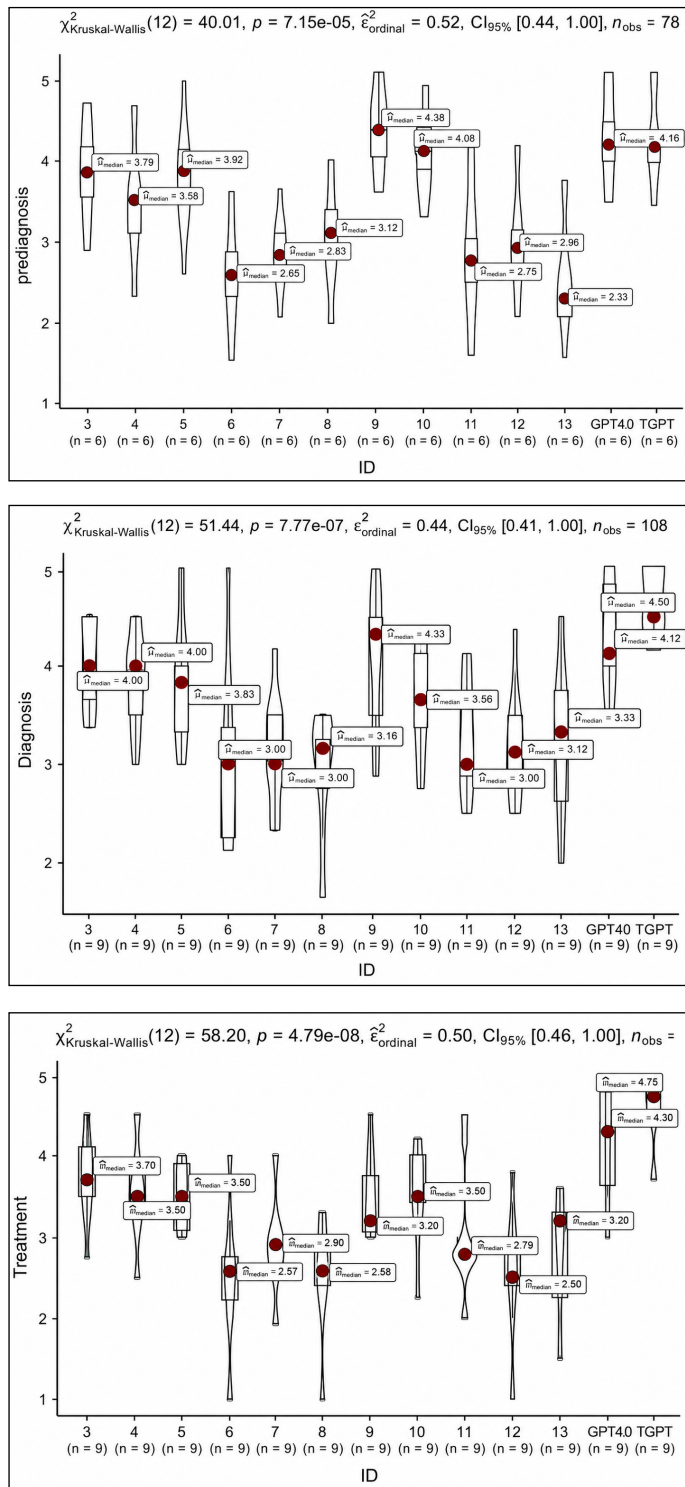


Figure 2. Violin plots of case-level clinical reasoning scores across individual fellows and AI models. The figure shows score distributions for each junior nephrology fellow (participant identifier) and for the two AI models in the prediagnosis, diagnosis, and treatment domains. Red dots denote median scores, and Kruskal-Wallis test results are displayed above each panel. Prediagnosis was assessed in 6 cases, whereas diagnosis and treatment were assessed in all 9 cases. Abbreviations: GPT-4o, standard GPT-4o model; ID, participant identifier; TrainedGPT, custom GPT-4 model configured with nephrology guideline-based content.

DISCUSSION

Our primary finding is that both TrainedGPT and GPT-4o demonstrated superior performance and greater consistency than junior nephrology fellows across all stages of clinical reasoning in simulated nephrology cases. Furthermore, TrainedGPT, which was configured using guideline-derived documents, showed no statistically significant advantage over GPT-4o.

Our study demonstrated that the case-based clinical reasoning performance of GPT-4-based models was significantly higher than that of junior nephrology fellows. This finding aligns with recent evidence from other specialties comparing AI with physicians-in-training. For instance, a pilot study in rheumatology demonstrated that a guideline-trained GPT-4 model outperformed junior fellows in clinical reasoning and management of inflammatory arthritis cases, achieving higher Revised-IDEA (R-IDEA) scores and demonstrating greater consistency (3). Furthermore, a multicenter study evaluating Canadian emergency medicine in-training examinations reported that GPT-4 surpassed the performance of residents at all postgraduate levels for two consecutive years, providing additional evidence of its advantages in clinical reasoning tasks (22).

However, the integration of AI into clinical practice is complex, with its impact appearing to be task-dependent. A randomized clinical vignette study focused on diagnostic reasoning found that providing physicians access to GPT-4 did not significantly improve their performance compared to using traditional resources (23). Interestingly, in that study, GPT-4 alone achieved higher diagnostic accuracy than both AI-assisted and unassisted physician groups. This finding suggests that effective human-AI collaboration may require structured support systems or specialized prompting techniques. In contrast, a separate randomized controlled trial examining treatment planning demonstrated that GPT-4 assistance significantly improved physicians' management reasoning performance (mean improvement of 6.5%; $p < 0.001$), suggesting that AI may be particularly beneficial in treatment-oriented contexts (24). Our findings in nephrology contribute to this nuanced discussion, confirming that GPT-4's capabilities in integrated case-based clinical tasks exceed those of early-stage physicians. Collectively, these results underscore that achieving optimal clinical utility of AI will depend on developing structured strategies for human-AI collaboration. In practice, such models may be integrated into fellowship training or attending-supervised case conferences, where fellows first formulate an independent management plan and then compare it with an AI-generated, guideline-aligned response during structured debriefing.

Another finding of this study is the absence of a statistically significant performance difference between the TrainedGPT and GPT-4o ($p = 0.690$). This situation should not be interpreted as a failure of specific training but rather as evidence of the high baseline capabilities of advanced LLMs such as GPT-4. Models like GPT-4 have been pre-trained on extensive text data, enabling

them to acquire a broad medical knowledge base and strong clinical reasoning capabilities. It is plausible that prominent guidelines were already part of its knowledge base, making the explicit training largely redundant. This highlights the inherent "black box" nature of LLMs; we can hypothesize about how they process information, but it is impossible to precisely trace how new data is weighed against pre-existing knowledge. Nevertheless, the trained model did show a narrower score distribution, suggesting that the training, while not boosting the overall score, did promote greater response consistency.

The demonstrated proficiency of LLMs in complex clinical reasoning, as shown in our study, positions them as powerful tools that can expand the current ecosystem of AI in nephrology (5,25). By moving beyond multiple-choice questions to complex case scenarios, our study provides an evaluation that is closer to the real-world application of LLMs. AI-supported clinical decision support systems (CDSS) can increase the effectiveness of alerts by using clinician feedback, which can potentially reduce alert density while increasing acceptance rates (26). For example, the implementation of AI-generated recommendations resulted in an average 12.3% reduction in alert triggers and a 16.9% increase in acceptance rates across various workflows (26). Additionally, the ability of ChatGPT 4.0 to assist in the triage and specialist referral of nephrology patients has been evaluated (11).

Although several between-group differences reached statistical significance, the absolute differences on a 5-point Likert scale should be interpreted cautiously with respect to their real-world clinical impact. Accordingly, these findings are best understood as evidence of comparative performance under standardized simulated conditions rather than a direct estimate of bedside effectiveness.

The findings of this study must be considered within the context of several key limitations. First, our methodology relied on idealized case scenarios, which lack the complexity, incomplete data, and patient-specific variables of real-world clinical encounters. Second, the small cohort of 11 junior fellows limits the generalizability of our results, positioning this work as an exploratory pilot study. Third, our statistical analysis did not account for intra-individual correlations across cases, and the absence of a senior expert group restricts our conclusions to a comparison with trainees only. Furthermore, this study is subject to the inherent limitations of all LLMs. These include the risk of generating "hallucinations"—plausible but incorrect information (7,27,28)—and potential algorithmic biases. Finally, by evaluating only written responses, our design could not assess core physician competencies such as empathy, patient communication, and real-time ethical deliberation (7,11,29).

Acknowledging these limitations is crucial, as they frame the appropriate role for AI in medicine. It is precisely within the complex, unscripted domains of nuanced communication and ethical judgment—areas this study could not measure—that the human physician remains irreplaceable. Therefore, our findings should be interpreted as highlighting AI's potential as a powerful

supportive tool for education and clinical practice, rather than as a replacement for human clinicians. To safely harness this potential, future research must move beyond performance metrics to rigorously address the critical challenges of safety, data privacy, and the ethical and practical feasibility of integrating these sophisticated models into live clinical workflows.

CONCLUSION

This study demonstrated that GPT-4-based AI models outperformed junior nephrology fellows in both performance and consistency within selected simulated nephrology scenarios. The absence of a significant difference between TrainedGPT and GPT-4o suggests that the foundational knowledge base of current LLMs may already be robust for standard nephrology tasks. These findings support the use of such models as supervised adjuncts in nephrology education and structured clinical reasoning rather than autonomous decision-makers. However, broader clinical integration will require real-world validation, assessment of patient-centered outcomes, and careful attention to safety, bias, privacy, and accountability.

Conflict of Interests

The authors declare that there is no conflict of interest in the study.

Financial Disclosure

The authors declare that there is no financial support.

Ethical Approval

This study was deemed exempt from formal ethical committee review because it used simulated clinical scenarios and anonymous participant responses and did not involve real patient data or personally identifiable information.

Informed Consent

Not applicable. The study involved only hypothetical clinical scenarios and anonymous/pseudonymous responses, without any real patient data, clinical intervention, or personally identifiable information.

REFERENCES

1. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. *J Fam Med Prim Care*. 2019;8:2328-31.
2. CB Insights Research. Healthcare remains the hottest AI category for deals. 2017. <https://www.cbinsights.com/research/artificial-intelligence-healthcare-startups-investors/> access date 09.07.2025.
3. Kayacan Erdoğan E, Babaoğlu H. Clinical reasoning and knowledge assessment of rheumatology residents compared to AI models: a pilot study. *J Clin Med*. 2024;13:7405.
4. Crotti N. Startup Arterys wins FDA clearance for AI-assisted cardiac imaging system. 2017. <https://medcitynews.com/2017/01/arterys-fda-ai/> access date 09.07.2025.
5. Miao J, Thongprayoon C, Suppadungsuk S, et al. Innovating personalized nephrology care: exploring the potential utilization of ChatGPT. *J Pers Med*. 2023;13:1681.
6. Badrouchi S, Bacha MM, Hedri H, et al. Toward generalizing the use of artificial intelligence in nephrology and kidney transplantation. *J Nephrol*. 2023;36:1087-100.

7. Du X, Zhou Z, Wang Y, et al. Performance and improvement strategies for adapting generative large language models for electronic health record applications: a systematic review. *Int J Med Inform.* 2026;205:106091.
8. Sexton DJ, Judge C. Assessments of generative artificial intelligence as clinical decision support ought to be incorporated into randomized controlled trials of electronic alerts for acute kidney injury. *Mayo Clin Proc Digit Health.* 2024;2:606-10.
9. Shmilovitch AH, Katson M, Cohen-Shelly M, et al. GPT-4 as a clinical decision support tool in ischemic stroke management: evaluation study. *JMIR AI.* 2025;4:e60391.
10. Lin TH, Chung HY, Jian MJ, et al. An advanced machine learning model for a web-based artificial intelligence-based clinical decision support system application: model development and validation study. *J Med Internet Res.* 2024;26:e56022.
11. Koirala P, Thongprayoon C, Miao J, et al. Evaluating AI performance in nephrology triage and subspecialty referrals. *Sci Rep.* 2025;15:3455.
12. Miao J, Thongprayoon C, Garcia Valencia OA, et al. Performance of ChatGPT on nephrology test questions. *Clin J Am Soc Nephrol.* 2024;19:35-43.
13. Barbieri C, Molina M, Ponce P, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int.* 2016;90:422-9.
14. Lim LM, Lin MY, Hsu C, et al. Computer-assisted prescription of erythropoiesis-stimulating agents in patients undergoing maintenance hemodialysis: a randomized control trial for artificial intelligence model selection. *JAMIA Open.* 2025;8:ooaf020.
15. Liu M, Okuhara T, Chang XY, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J Med Internet Res.* 2024;26:e60807.
16. 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:e0000198.
17. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep.* 2023;13:16492.
18. Knoedler L, Alfertshofer M, Knoedler S, et al. Pure wisdom or Potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 style questions: quantitative analysis. *JMIR Med Educ.* 2024;10:e51148.
19. Gilson A, Safranek CW, Huang T, et al. 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:e45312.
20. Wu S, Koo M, Blum L, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI.* 2024;1. doi: 10.1056/AIdbp2300092.
21. Miao J, Thongprayoon C, Cheungpasitporn W. Assessing the accuracy of ChatGPT on core questions in glomerular disease. *Kidney Int Rep.* 2023;8:1657-9.
22. Almehairi N, Clark G, Davis S. How well does GPT-4 perform on an emergency medicine board exam? A comparative assessment. *CJEM.* 2025;27:969-73.
23. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open.* 2024;7:e2440969.
24. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med.* 2025;31:1233-8.
25. Lin SY, Jiang CC, Law KM, et al. Evaluation of generative AI assistance in clinical nephrology: assessing GPT-4, GPT-4o, Gemini 1.0 Ultra, and PaLM 2 in patient interaction and renal biopsy interpretation. *Digit Health.* 2025;11:20552076251342067.
26. Liu S, McCoy AB, Peterson JF, et al. Leveraging explainable artificial intelligence to optimize clinical decision support. *J Am Med Inform Assoc.* 2024;31:968-74.
27. Suppadungsuk S, Thongprayoon C, Krisanapan P, et al. Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. *J Clin Med.* 2023;12:5550.
28. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus.* 2023;15:e35179.
29. Loftus TJ, Shickel B, Ozrazgat-Baslanti T, et al. Artificial intelligence-enabled decision support in nephrology. *Nat Rev Nephrol.* 2022;18:452-65.