Comparison of ChatGPT and Claude in Managing Real-Life Difficult Nephrology Cases

Keywords
nephrology, ChatGPT, Large Language Models, Claude, QAMAI

Abstract

Introduction

Artificial intelligence (AI) based large language models (LLMs) are promising tools for clinical decision support, but their reliability in specialized fields like nephrology is still uncertain. ChatGPT and Claude represent distinct AI architectures with potentially different clinical utilities. We aimed to compare the diagnostic accuracy, treatment recommendations, and overall clinical utility of these two AI models in managing real life difficult nephrology cases.

Material and methods

Twenty-two real nephrology cases from a tertiary care university hospital were presented to both models, covering disorders such as glomerulonephritis, acute kidney injury, vasculitis, and transplant complications. Each model's output was assessed for diagnostic accuracy, risk evaluation, test recommendations, and treatment planning. Three independent nephrologists evaluated the responses using the Quality Assessment of Medical Information (QAMAI) and Global Quality Score (GQS) tools. Statistical comparisons were performed using the Wilcoxon signed-rank test, with $p<0.05$ considered significant.

Results

Claude achieved higher diagnostic accuracy than ChatGPT (4.59 ± 0.41 vs. 4.36 ± 0.48; p=0.048), whereas ChatGPT scored better in clarity (4.63 ± 0.30 vs. 4.32 ± 0.29; p=0.002). No significant differences were found in relevance, completeness, usefulness, or source citation. Overall QAMAI scores were comparable between the two models (ChatGPT: 23.72 ± 1.46; Claude: 23.39 ± 1.43; p=0.371). Inter-rater reliability ranged from moderate to good, with the highest agreement observed for ChatGPT's GQS.

Conclusions

Both ChatGPT and Claude demonstrate notable potential as decision-support tools in nephrology. Claude provided slightly higher diagnostic accuracy, while ChatGPT offered greater clarity. Despite these promising results, clinical judgment remains essential when interpreting LLM-generated suggestions.

Comparison of ChatGPT and Claude in Managing Real-Life Difficult Nephrology Cases

Abstract

Background

Artificial intelligence (AI) based large language models (LLMs) are promising tools for clinical decision support, but their reliability in specialized fields like nephrology is still uncertain. ChatGPT and Claude represent distinct AI architectures with potentially different clinical utilities. We aimed to compare the diagnostic accuracy, treatment recommendations, and overall clinical utility of these two AI models in managing real life difficult nephrology cases.

Methods

Twenty-two real nephrology cases from a tertiary care university hospital were presented to both models, covering disorders such as glomerulonephritis, acute kidney injury, vasculitis, and transplant complications. Each model's output was assessed for diagnostic accuracy, risk evaluation, test recommendations, and treatment planning. Three independent nephrologists evaluated the responses using the Quality Assessment of Medical Information (QAMAI) and Global Quality Score (GQS) tools. Statistical comparisons were performed using the Wilcoxon signed-rank test, with $p<0.05$ considered significant.

Results

Claude achieved higher diagnostic accuracy than ChatGPT (4.59 ± 0.41 vs. 4.36 ± 0.48; $p=0.048$), whereas ChatGPT scored better in clarity (4.63 ± 0.30 vs. 4.32 ± 0.29; $p=0.002$). No significant differences were found in relevance, completeness, usefulness, or source citation. Overall QAMAI scores were comparable between the two models (ChatGPT: 23.72 ± 1.46; Claude: 23.39 ± 1.43; $p=0.371$). Inter-rater reliability ranged from moderate to good, with the highest agreement observed for ChatGPT's GQS.

Conclusions

Both ChatGPT and Claude demonstrate notable potential as decision-support tools in nephrology. Claude provided slightly higher diagnostic accuracy, while ChatGPT offered greater clarity. Despite these promising results, clinical judgment remains essential when interpreting LLM-generated suggestions.

Keywords: Large Language Models, ChatGPT, Claude, QAMAI, nephrology

Introduction

Artificial intelligence (AI) has rapidly advanced in healthcare, and large language models (LLMs) such as ChatGPT and Claude have shown growing potential as decision-support tools (1). Trained on extensive text-based datasets that include medical literature, these models can interpret complex clinical information and generate human-like, contextually relevant responses (1). However, their reliability in specialized fields remains uncertain (2).

Nephrology poses a particular challenge for AI systems due to its diagnostic complexity, requiring integration of clinical data, laboratory findings, imaging, and often histopathology (2). Diseases like acute kidney injury, glomerulonephritis, and transplant complications demand complex reasoning usually provided by experienced nephrologists. Although LLMs have shown encouraging results in general medical and surgical fields their performance in nephrology remains largely unexplored (2-7).

Among commercially available LLMs, ChatGPT (OpenAI) and Claude (Anthropic) represent two of the most widely used platforms with distinct architectural characteristics. ChatGPT-4o is a multimodal model capable of processing both text and image inputs, with real-time internet access for current information retrieval (8). Claude is trained using a Constitutional AI approach, in which the model is guided by an explicit set of ethical principles during training—rather than relying primarily on direct human labeling—to reduce harmful outputs while preserving helpfulness (9).

With the growing capabilities of LLMs, evaluating their performance in real-life nephrology cases has become increasingly important. This study compares ChatGPT and Claude to determine how accurately and clearly they can assist clinicians in complex decision-making.

Materials and Methods

Study Design and Case Selection

This cross-sectional comparative study evaluated the performance of ChatGPT-4o and Claude 3.7 Sonnet on real nephrology cases. Twenty-two adult cases (>18 years) discussed at the case council of the nephrology department at a tertiary university hospital between

January and August 2025 were included. The cases represented a wide range of kidney diseases, including acute kidney injury, glomerulonephritis, vasculitis, electrolyte disorders, and post-transplant complications. Each case contained demographic information, clinical findings, laboratory results, and treatment history to ensure comprehensive evaluation.

To avoid memory bias, each case was presented in a new chat session (10). Both LLMs received identical case descriptions in English and the same instruction: "I will provide real-life difficult nephrology cases discussed at a tertiary nephrology center. Evaluate each case and answer the following: 'Should additional tests be requested? What is your treatment recommendation?' Limit your response to 500 words." All evaluations were conducted between January and June 2025. One of the cases presented to the LLMs, along with the responses of both models, is shown in Supplementary Material.

Assessment of Model Outputs

Three nephrologists served as evaluators: two professors of nephrology with over 22 and 25 years of clinical experience, respectively, and one associate professor with over seven years of experience. For the evaluation, each nephrologist received only the case presentations provided to the LLMs, without access to the documented council decisions, to prevent potential bias in scoring. Evaluators independently scored the LLM-generated outputs using two validated scoring tools:

- Global Quality Score (GQS): A five-point Likert scale assessing overall quality (1 = poor, 5 = excellent).
- Quality Assessment of Medical Information (QAMAI): A structured scale evaluating six domains—accuracy, clarity, relevance, completeness, source citation, usefulness, —each rated from 1 (strongly disagree) to 5 (strongly agree), with total scores ranging from 6 to 30 (11). Responses are classified as excellent (26–30), good (21–25), moderate (16–20), poor (11–15), or very poor (6–10) (11).

Statistical Analysis

QAMAI sub-dimension scores were calculated as the mean of ratings provided by the expert evaluators for each question. Descriptive statistics (mean, standard deviation, median,

minimum–maximum) were reported. The normality of distributions was assessed using the Shapiro–Wilk test. As the data were not normally distributed, comparisons between ChatGPT and Claude were performed using the Wilcoxon signed-rank test.

For each sub-dimension, the proportion of high scores (≥4) was presented as frequency and percentage, and between-model differences were analyzed using McNemar's test. Inter-rater reliability was evaluated with the Intraclass Correlation Coefficient (ICC) and corresponding 95% confidence intervals, applying a two-way random effects model with absolute agreement. For ICC interpretation, we followed established guidelines: values below 0.50 indicated poor reliability, 0.50-0.75 moderate reliability, 0.75-0.90 good reliability, and above 0.90 excellent reliability. Effect sizes were calculated using Cohen's d to quantify the practical significance of differences between the two models. Values were interpreted as small (d = 0.2), medium (d = 0.5), or large (d = 0.8) effects. Statistical significance was defined as $p < 0.05$. All analyses were conducted using IBM SPSS Statistics version 21.0 (IBM Corp., Chicago, IL, USA).

Results

Overall Model Performance

Both LLMs performed well across all evaluation parameters (Table 1). Claude demonstrated significantly higher accuracy than ChatGPT (4.59 ± 0.41 vs. 4.36 ± 0.48, p = 0.048). In contrast, ChatGPT provided clearer responses (4.63 ± 0.30 vs. 4.32 ± 0.29, p = 0.002). No statistically significant differences were found between the models in terms of relevance, completeness, source citation, or usefulness. GQS was also comparable (ChatGPT = 4.50 ± 0.48; Claude = 4.47 ± 0.39, p = 0.674). Similarly, total QAMAI scores did not differ significantly (ChatGPT = 23.72 ± 1.46; Claude = 23.39 ± 1.43, p = 0.371). Comparative performance of ChatGPT and Claude is presented in Figure 1.

Inter-Rater Reliability

Inter-rater reliability analysis showed moderate to good agreement across most evaluation parameters (Table 2). For ChatGPT, the highest consistency was observed for the GQS (*ICC* = 0.805, 95% CI 0.609–0.912) and accuracy (*ICC*= 0.647, 95% CI 0.273–0.843). For Claude, the strongest agreement was also for accuracy (*ICC* = 0.608, 95% CI 0.239–0.820) and usefulness (*ICC* = 0.564, 95% CI 0.165–0.798).

Agreement Among Evaluators

Evaluator agreement was high for most parameters (Table 3). For ChatGPT, 100% of raters assigned high scores (≥4) for clarity, relevance, and usefulness, with slightly lower agreement for accuracy (81.8%). Claude achieved full agreement for clarity and relevance (100%) and similarly high agreement for accuracy (95.5%), GQS (95.5%), and usefulness (95.5%).

Discussion

This study provides one of the first systematic evaluations of large language models (LLMs) in complex nephrology decision-making. Both ChatGPT and Claude performed well across multiple quality domains, producing coherent and clinically relevant responses even for challenging real-life cases.

Claude demonstrated slightly higher diagnostic accuracy, while ChatGPT provided clearer and more easily understood answers. These complementary strengths likely reflect differences in model architecture and training. In clinical practice, LLMs could help summarize complex cases, suggest differential diagnoses, and support multidisciplinary discussions, particularly in settings where nephrology expertise is limited. Despite their distinct profiles, both models achieved comparable overall quality scores, suggesting similar overall reliability.

Qualitatively, both models' recommendations were generally concordant with the multidisciplinary council decisions. However, council deliberations often incorporated additional factors—such as patient preferences, comorbidities, local resource availability, treatment tolerability, and longitudinal clinical context—that were not fully captured in the case summaries provided to the LLMs. Our diverse case selection—including membranous nephropathy (n=4), FSGS (n=4), IgA nephropathy (n=3), minimal change disease (n=3), vasculitis (n=2), and other conditions—did not reveal qualitative differences in model performance across disease categories, though formal subgroup analysis was precluded by small sample sizes per diagnosis. Despite their distinct profiles, both models achieved comparable overall quality scores, suggesting similar overall reliability.

It should be noted that both models currently process data obtained from patients and/or medical sources through the physicians in charge, who had already performed clinical

evaluations, diagnostic testing, and data synthesis. In real-world practice, patients cannot present their conditions with equivalent medical detail and coherence, underscoring that the physician's role in clinical assessment remains fundamental and precedes any AI-assisted decision support.

The most significant weakness of both models was their failure to cite appropriate references. This finding has important implications for clinical practice, as clinicians cannot verify the evidence basis of AI-generated recommendations. This may be attributable to the nature of the prompt used, which did not explicitly request citations. It should also be noted that both models offer a research option with longer response times, which was not selected in our evaluation to maintain consistency and simulate typical clinical consultation scenarios.

Future research should validate these models in real-time clinical workflows and explore secure integration with electronic health record systems. Continuous assessment of newer versions and specialty-specific fine-tuning will be essential to improve performance and maintain relevance. Developing transparent source citation mechanisms is also critical to promote evidence-based use and clinician trust.

Our findings are consistent with previous research highlighting the potential of LLMs in other medical specialties (1, 3-7, 10, 12). The high ratings for clarity, relevance, and usefulness underscore their value as supportive clinical tools. However, ensuring transparency and verifiable evidence remains an important challenge before broader adoption (3, 13, 14).

This study has several limitations. Although real clinical cases were used, the evaluations occurred in a controlled environment rather than during real-time decision-making. The tested models—ChatGPT-4o and Claude 3.7 Sonnet—will soon be replaced by newer versions, which may limit the long-term relevance of these findings. Our analysis was based on written case summaries, lacking the context of direct patient interaction and multidisciplinary input. Finally, the small sample size of 22 cases, though diverse, cannot capture the full spectrum of nephrology practice. These limitations highlight the need for continued research and validation in real-world settings, ideally with larger, disease specific cohorts.

In conclusion, both ChatGPT and Claude show promising potential as decision-support tools in nephrology, offering complementary strengths—Claude in diagnostic

precision and ChatGPT in clarity. Used responsibly, they can enhance clinical reasoning and education but should remain supportive aids rather than substitutes for expert judgment. As LLMs continue to evolve, ongoing refinement, transparency, and clinical validation will be essential for their safe and effective integration into nephrology practice.

Declarations

Statement of Ethics

The study was approved by the local medical ethics committee (approval no: 2025/165) and conducted according to the Declaration of Helsinki. Informed consent was taken from all patients.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Competing Interests

The authors declare no competing interests.

Funding Sources

This research received no external funding.

Authors' Contributions

SGO and NS conceptualized and designed the study. SGO, AMA, and AOP were involved in data collection. MTD, ST and NS evaluated the LLMs responses. ZA conducted the statistical analysis. SGO drafted the manuscript. NS and all authors critically reviewed and revised the manuscript for intellectual content. All authors read and approved the final version.

Acknowledgements: None declared.

Figure Legends:

Figure 1: Comparative performance of ChatGPT and Claude. Bars represent mean ± standard deviation for each evaluation parameter (Accuracy, Clarity, Relevance, Completeness, Usefulness, and GQS) on a 5-point scale (1 = poor, 5 = excellent). Asterisks indicate statistically significant differences between models (*$p < 0.05$; **$p < 0.01$).

References

1. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Network Open. 2023;6(11):e2343689-e.
2. Hu Y, Liu J, Jiang W. Large language models in nephrology: applications and challenges in chronic kidney disease management. Renal Failure. 2025;47(1):2555686.
3. Pal A, Sankarasubbu M, editors. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. Proceedings of the 6th Clinical Natural Language Processing Workshop; 2024.
4. Gomez-Cabello CA, Borna S, Pressman SM, Haider SA, Forte AJ. Large language models for intraoperative decision support in plastic surgery: a comparison between ChatGPT-4 and Gemini. Medicina. 2024;60(6):957.
5. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR medical informatics. 2024;12:e53787.
6. Ao G, Chen M, Li J, Nie H, Zhang L, Chen Z. Comparative analysis of large language models on rare disease identification. Orphanet Journal of Rare Diseases. 2025;20(1):150.
7. Li Y, Dong J, Liu D, Huang Y, Jiang Y, Chen L, et al. Systematic benchmarking of large Language models in programmed cell death-oriented gastric cancer research: a comparative analysis of DeepSeek-V3, DeepSeek-R1, and Claude 3.5. Discover Oncology. 2025;16:1227.
8. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.
9. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:221208073. 2022.
10. Zeller NP, Shah AD, Van Heest AE, Bohn DC. Assessing Accuracy of Chat Generative Pre-Trained Transformer's Responses to Common Patient Questions Regarding Congenital Upper Limb Differences. Journal of Hand Surgery Global Online. 2025;7(4):100764.
11. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. European Archives of Oto-Rhino-Laryngology. 2024;281(11):6123-31.
12. Wu X, Huang Y, He Q. Diagnostic performance of newly developed large language models for critical illness cases: A comparative study. International Journal of Medical Informatics. 2025:106088.
13. Hancı V, Ergün B, Gül Ş, Uzun Ö, Erdemir İ, Hancı FB. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Medicine. 2024;103(33):e39305.
14. Athaluri SA, Manthena SV, Kesapragada VKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus. 2023;15(4).

Table 1. Comparison of mean scores for ChatGPT and Claude across evaluation parameters.

Parameter	ChatGPT (Mean ± SD, Median)	Claude (Mean ± SD, Median)	*p*-value
Accuracy	4.36 ± 0.48 (4.33)	4.59 ± 0.41 (4.67)	0.048
Clarity	4.63 ± 0.30 (4.67)	4.32 ± 0.29 (4.33)	0.002
Relevance	4.79 ± 0.26 (5.00)	4.82 ± 0.25 (5.00)	0.642
Completeness	4.38 ± 0.36 (4.33)	4.18 ± 0.42 (4.33)	0.165
Source citation	1.00 (1.00)	1.00 (1.00)	1.000
Usefulness	4.55 ± 0.35 (4.67)	4.48 ± 0.42 (4.50)	0.544
QAMAI total score	23.72 ± 1.46 (23.83)	23.39 ± 1.43 (23.33)	0.371
GQS	4.50 ± 0.48 (4.67)	4.47 ± 0.39 (4.33)	0.674

QAMAI: Quality assessment of medical information, GQS: Global quality score

Table 2. Intraclass correlation coefficients (ICC) for both models across evaluation parameters.

Parameter	ChatGPT (ICC, 95% CI)	Claude (ICC, 95% CI)
Accuracy	0.647 (0.273–0.843)	0.608 (0.239–0.820)
Clarity	0.319 (–0.211–0.671)	–0.016 (–0.272–0.318)
Relevance	0.278 (–0.445–0.675)	0.097 (–0.663–0.573)
Completeness	0.205 (–0.657–0.649)	0.300 (–0.388–0.683)
Usefulness	0.300 (–0.273–0.666)	0.564 (0.165–0.798)
GQS	0.805 (0.609–0.912)	0.498 (–0.013–0.768)
QAMAI total	0.579 (0.165–0.809)	0.423 (0.025–0.719)

QAMAI: Quality assessment of medical information, GQS: Global quality score

Table 3. Percentage of evaluators reporting high scores (≥4) for each parameter.

Parameter	ChatGPT n (%)	Claude n (%)	*p*-value
Accuracy	18 (81.8)	21 (95.5)	0.375
Clarity	22 (100)	22 (100)	NA
Relevance	22 (100)	22 (100)	NA
Completeness	20 (90.9)	19 (86.4)	1.000
Source citation	0 (0)	0 (0)	NA
Usefulness	22 (100)	21 (95.5)	1.000
GQS	20 (90.9)	21 (95.5)	1.000

GQS: Global quality score

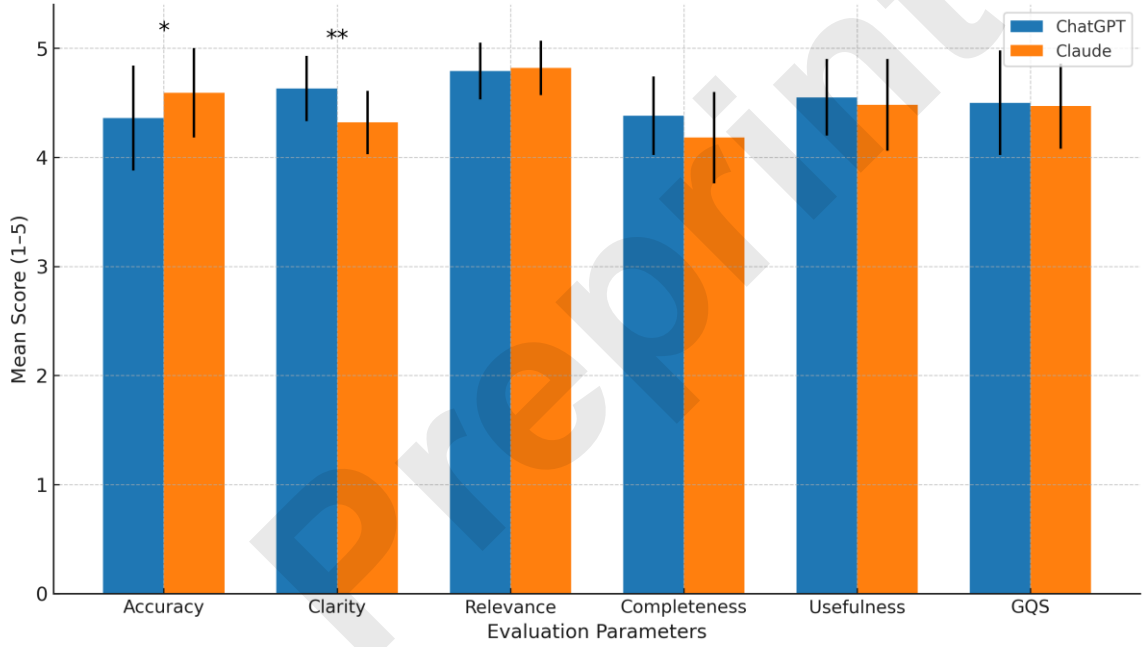

Figure 1: Comparative performances of ChatGPT and Claude.

Comparison of ChatGPT and Claude in Managing Real-Life Difficult Nephrology Cases

Cases: Membranous Nephropathy (4), FSGS (4), IgA Nephropathy (3), MCD (3), Vasculitis (2), C3GN/MPGN (2), TMA (1), Transplant (1), IgG4-RD (1), Other (1)

LLM = Large Language Model; QAMAI = Quality Assessment of Medical AI; GQS = Global Quality Score; FSGS = Focal Segmental Glomerulosclerosis; MCD = Minimal Change Disease; C3GN = C3 Glomerulopathy; MPGN = Membranoproliferative Glomerulonephritis; TMA = Thrombotic Microangiopathy; IgG4-RD = IgG4-Related Disease