kidney and dialysis

MDPI

Article

Comparative Analysis of ChatGPT and Gemini in Addressing Questions from Chronic Kidney Disease Patients

Yasemin Bati Sutcu [1,†], Seyda Gul Ozcan [2,†], Mevlut Tamer Dincer [2], Zeynep Atli [3], Sinan Trabulus [2] and Nurhan Seyahi [2,*]

[1] Department of Internal Medicine, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, Istanbul 34098, Türkiye
[2] Department of Nephrology, Cerrahpasa Medical Faculty, Istanbul University-Cerrahpasa, Istanbul 34098, Türkiye; seydagul.ozcan@iuc.edu.tr (S.G.O.); mevluttamer.dincer@iuc.edu.tr (M.T.D.); sinan.trabulus@iuc.edu.tr (S.T.)
[3] Department of Data Science and Analytics, Sinop University, Sinop 57000, Türkiye; zynp.atlii@gmail.com
* Correspondence: nseyahi@gmail.com
† These authors contributed equally to this work.

Abstract

Background: Chronic kidney disease (CKD) is a major global health burden. Patient education is a crucial part of CKD management. Large language models (LLMs) such as ChatGPT and Gemini may help patients access medical information, but their reliability in CKD-related contexts is uncertain. **Methods**: We collected 291 questions from 100 CKD patients and selected and analyzed 123 of them across three categories: medical condition and treatment, nutrition and diet, and symptom management. Responses from ChatGPT and Gemini were assessed by two nephrology specialists using the Quality Assessment of Medical Artificial Intelligence (QAMAI) scale. **Results**: When all 123 questions were evaluated together, ChatGPT outperformed Gemini in terms of clarity and usefulness. However, when the questions were analyzed by category, Gemini demonstrated relatively stronger performance in the nutrition and symptom management domains. Accuracy and relevance were comparable between the two models. Neither consistently provided adequate citations. **Conclusion**: ChatGPT and Gemini demonstrate potential as supplementary tools for CKD patient education, with complementary strengths across different domains. Although they cannot replace clinical expertise, their supervised use could enhance information access and reduce clinician burden.

Keywords: chronic kidney disease; ChatGPT; Gemini; large language models; artificial intelligence

Academic Editor: Francesco Locatelli

Received: 27 November 2025
Revised: 21 January 2026
Accepted: 22 January 2026
Published: 3 February 2026

1. Introduction

Chronic Kidney Disease (CKD) is a major and growing global health challenge [1,2]. CKD is associated with complex systemic complications, including cardiovascular, endocrine, and hematologic disorders, which demand comprehensive management and close patient engagement [3–7]. However, previous research has shown that many patients remain inadequately informed about their condition even after diagnosis [8]. Insufficient knowledge may compromise adherence to treatment and self-management strategies. In clinical practice, time constraints during nephrology consultations, high patient volumes, complex follow-up schedules, and the sheer breadth of information requiring communication all limit opportunities for comprehensive patient education. Many patients

report that their questions remain unanswered, and studies consistently show that substantial proportions of CKD patients lack a basic understanding of their condition even after diagnosis.

At the same time, the information environment surrounding CKD has changed dramatically. Many patients search the internet or use social media platforms to obtain health information, frequently encountering fragmented, commercially driven, or conflicting messages. This uncontrolled information landscape can exacerbate anxiety, reinforce misconceptions, and make it difficult for patients to distinguish evidence-based recommendations from misleading content. For CKD—where diet, medication adherence, symptom monitoring, and timely referral to specialist care are all critical—these informational gaps and inconsistencies may translate directly into worse outcomes. Systematic reviews show that limited health literacy is common in CKD and is associated with poorer self-management and higher morbidity [8–10]. Against this background, scalable, high-quality tools for patient education have become a central priority in kidney care.

Advances in artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT and Gemini, have opened new avenues for patient education. These conversational AI systems generate human-like text responses and are increasingly used by the general public to seek health-related information [11,12]. Their accessibility and interactive design may offer an effective supplement to traditional education methods, particularly for chronic conditions that require continuous self-management. Despite this potential, questions remain regarding their accuracy, clarity, and reliability in specialized areas such as nephrology, as well as their adequacy in languages other than English.

The emergence of LLMs has also created new challenges. These systems may produce plausible but incorrect statements ("hallucinations"), omit important caveats, or fail to provide verifiable citations [13,14]. Their performance can vary by task, specialty, language, and model version. Moreover, although AI has been proposed as a tool to improve health literacy in kidney care, it remains unclear whether current models truly meet the informational needs of patients, especially in non-English contexts and in populations with limited health literacy [10,15].

In response to these concerns, several instruments have been developed to systematically evaluate AI-generated medical content. The Quality Assessment of Medical Artificial Intelligence (QAMAI) tool is among the first validated frameworks specifically designed to assess the quality of information generated by AI platforms, including multiple dimensions such as accuracy, clarity, relevance, completeness, source citation, and usefulness [16]. Applying such a standardized tool to patient-centered questions can help move beyond anecdotal impressions and provide structured evidence on model performance.

Despite rapidly growing interest in AI for nephrology, there is still limited evidence regarding LLM performance for CKD patient education, particularly in languages other than English [17–20]. Turkish, spoken by a large CKD population, has fewer high-quality patient education resources and is under-represented in AI training data compared with English. Research has documented systematic performance degradation for queries in languages other than English, with studies comparing Turkish versus English performance [21]. This linguistic performance gap is particularly relevant for patient populations where English is not the primary language, as the majority of health literacy barriers already disproportionately affect non-English speaking communities. Understanding how LLMs perform in Turkish and how their strengths differ across types of patient questions is, therefore, highly relevant for clinical practice in our setting and in other non-English-speaking regions.

This study evaluates the responses of ChatGPT and Gemini to questions commonly asked by CKD patients. By analyzing their performance, we aim to determine their potential role as supplementary educational tools. Particular attention is given to the suitability

of these models in a Turkish-language context, where locally relevant patient education resources are limited.

2. Methods

2.1. Study Design and Participants

We conducted an observational, cross-sectional study between January and July 2025 in the nephrology outpatient clinic of a tertiary university hospital. The study included 100 adult patients (aged ≥18 years) with CKD who were under regular follow-up and provided informed consent. Patients who were unable to write were excluded.

This study was approved by the Istanbul University-Cerrahpasa, Cerrahpasa Medical Faculty ethics committee (approval no: E-74555795-050.04-1274848; approval date: 12 March 2025) and conducted according to the Declaration of Helsinki.

2.2. Data Collection

We asked participants to write down up to three questions they most wanted to ask regarding their disease, treatment, follow-up, and daily life. Not all patients provided three questions, resulting in a total of 291 questions. After excluding those unrelated to nephrology, 123 questions remained eligible for analysis (Supplementary Table S1). Questions were grouped into three categories:

1. Medical condition and treatment (n = 73);
2. Nutrition and diet (n = 35);
3. Symptom management (n = 15).

Although several questions overlapped in content, all were retained to allow assessment of consistency in model responses. The study workflow is shown in Figure 1.

2.3. Large Language Model Evaluation

We posed each question to two LLMs (ChatGPT interface [OpenAI], powered by the GPT-4o LLM, and the Gemini interface [Google], powered by the Gemini 1.5 model) [22,23]. The evaluated versions of ChatGPT and Gemini correspond to the publicly available and widely used models at the time of data collection; we selected these versions to reflect real-world patient access rather than the most recent experimental releases. We submitted all queries in identical wording and order without user interaction and only recorded the first responses. All responses were independently evaluated by two nephrologists using the QAMAI tool. QAMAI is a validated scoring tool specifically designed to assess the quality of health information generated by medical AI systems [16]. It provides a structured framework that enables experts to move beyond subjective impressions and evaluate AI-generated content in a consistent, reproducible manner.

QAMAI evaluates six subdimensions, accuracy, clarity, relevance, completeness, citation of sources, and usefulness, with total scores ranging from 6 to 30 [16]. Accuracy reflects the degree to which the information aligns with current evidence and clinical guidelines; clarity captures how understandable and well-structured the explanation is; relevance assesses how directly the response addresses the user's question; completeness indicates whether essential aspects of the topic are covered; citation of sources evaluates the transparency and reliability of the referenced evidence; and usefulness measures the practical value of the response for decision-making and self-management. Responses are classified as excellent (26–30), good (21–25), moderate (16–20), poor (11–15), or very poor (6–10).

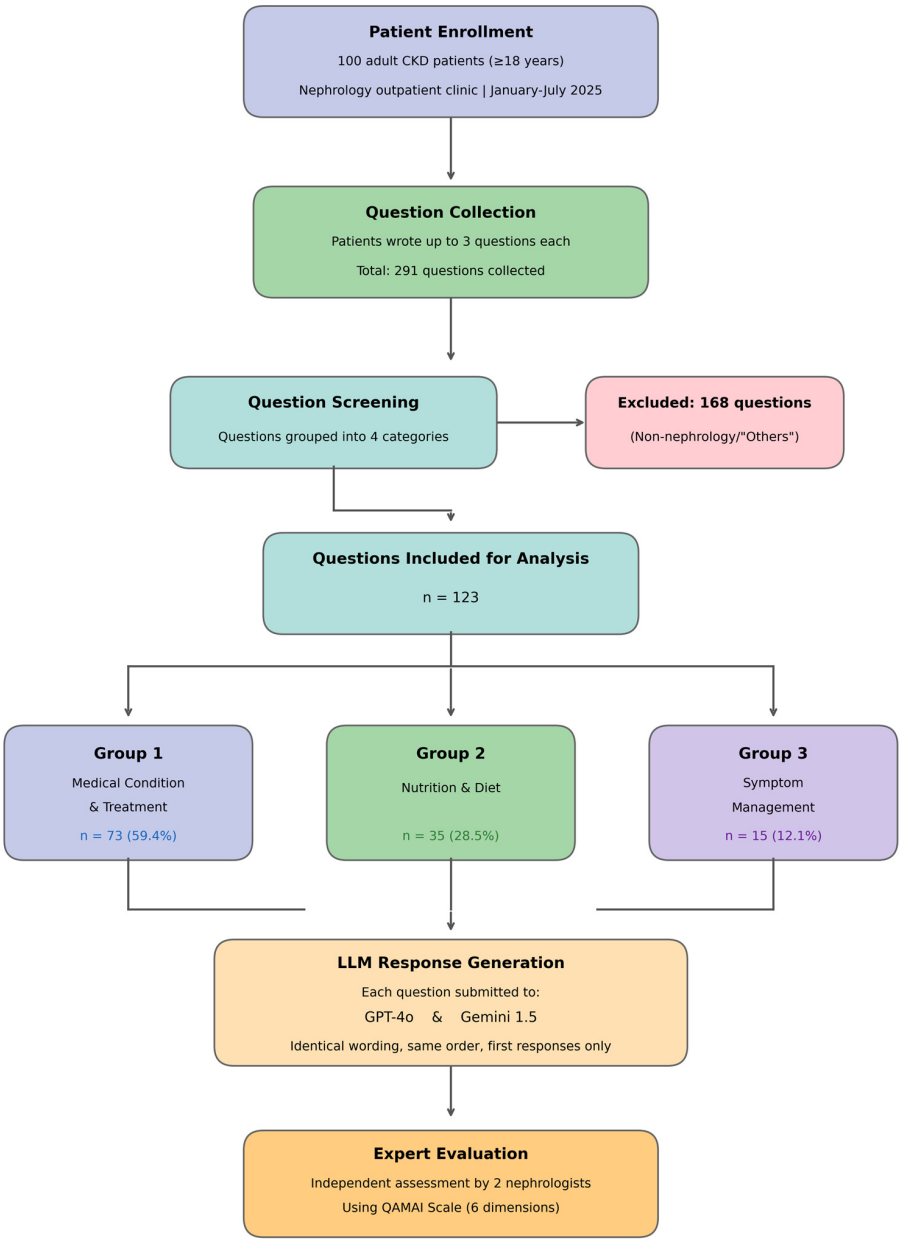

Figure 1. Study workflow. CKD: Chronic Kidney Disease; LLM: Large Language Model; QAMAI: Quality Assessment of Medical Artificial Intelligence.

2.4. LLM Scoring According to Nephrologists

We independently posed each patient question to both LLMs and recorded the first generated response. Two nephrologists served as evaluators: one professor of nephrology with over 22 years of clinical experience, and one associate professor with over seven years of experience. All six QAMAI dimensions were scored on a 5-point Likert scale (1 = very poor, 5 = excellent), in accordance with the original validation study. Both evaluators scored all dimensions for every response. The evaluators were blinded to the identity of the LLM. For analysis, final scores were calculated as the mean of the two evaluators' ratings. Inter-rater reliability was assessed using the intraclass correlation coefficient (two-way random effects model, absolute agreement).

To illustrate the scoring process, Supplementary Table S2 presents representative examples of patient questions, corresponding LLM responses, and the individual QAMAI scores assigned by each nephrologist.

2.5. 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, and minimum–maximum) were calculated. Normality was tested using the Shapiro–Wilk test. Since assumptions of normality were not met, comparisons between ChatGPT and Gemini were performed using the Wilcoxon signed-rank test.

For each sub-dimension, the proportion of high scores ($\geq$4) was reported as frequency and percentage, with between-model differences tested by McNemar's test. Inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC) with 95% confidence intervals (two-way random effects model, absolute agreement). Effect sizes were calculated using Cohen's d. Statistical significance was set at $p < 0.05$. Analyses were conducted using IBM SPSS version 21 (Chicago, IL, USA).

3. Results

A total of 100 patients with CKD were included. The mean age was 59.4 $\pm$ 16.1 years, and 50% were male. A total of 291 questions were collected, of which 123 were included in the final analysis after excluding those unrelated to nephrology. Of these, 73 (59.4%) addressed medical condition and treatment (group 1), 35 (28.5%) focused on nutrition and diet (group 2), and 15 (12.1%) concerned symptom management (group 3).

3.1. QAMAI Scores

Overall, the QAMAI assessment showed that both models generally provided high-quality responses, with mean total scores of 23.68 $\pm$ 1.78 for ChatGPT and 23.21 $\pm$ 2.42 for Gemini ($p = 0.106$) (Table 1). Assessment of total QAMAI scores demonstrated that 96.7% of ChatGPT responses and 87.8% of Gemini responses were classified as good (21–25 points) or excellent (26–30 points), indicating that both models generally provided high-quality information despite domain-specific differences.

Table 1. QAMAI mean scores.

Dimension	ChatGPT (Mean $\pm$ SD)	Gemini (Mean $\pm$ SD)	*p*-Value
Accuracy	4.44 $\pm$ 0.53	4.55 $\pm$ 0.53	0.095
Clarity	4.78 $\pm$ 0.32	4.50 $\pm$ 0.56	<0.001
Relevance	4.66 $\pm$ 0.43	4.58 $\pm$ 0.51	0.124
Completeness	4.28 $\pm$ 0.57	4.23 $\pm$ 0.70	0.634
Usefulness	4.51 $\pm$ 0.49	4.36 $\pm$ 0.63	0.009
Sources & References	1.00	1.00	NA
QAMAI Total Score	23.68 $\pm$ 1.78 (18–26)	23.21 $\pm$ 2.42 (14.5–26)	0.106

When individual dimensions were analyzed, ChatGPT performed significantly better for clarity (4.78 $\pm$ 0.32 vs. 4.50 $\pm$ 0.56, $p < 0.001$) and usefulness (4.51 $\pm$ 0.49 vs. 4.36 $\pm$ 0.63, $p = 0.009$). No significant differences were observed for accuracy (ChatGPT: 4.44 $\pm$ 0.53, Gemini: 4.55 $\pm$ 0.53, $p = 0.095$), relevance (4.66 $\pm$ 0.43 vs. 4.58 $\pm$ 0.51, $p = 0.124$), or completeness (4.28 $\pm$ 0.57 vs. 4.23 $\pm$ 0.70, $p = 0.634$). Both models consistently received the minimum score (1.00) for source citation across all responses (Table 1).

Subgroup analysis by question category revealed distinct performance patterns. For medical condition and treatment questions, ChatGPT scored significantly higher for clarity (4.72 ± 0.35 vs. 4.32 ± 0.59, $p < 0.001$), relevance (4.54 ± 0.47 vs. 4.36 ± 0.53, $p = 0.019$), completeness (4.24 ± 0.62 vs. 3.93 ± 0.69, $p = 0.001$), and usefulness (4.45 ± 0.52 vs. 4.11 ± 0.62, $p < 0.001$) (Table 2).

Table 2. QAMAI scores by question categories.

Dimension	Group 1			Group 2			Group 3		
	ChatGPT	**Gemini**	***p* Value**	**ChatGPT**	**Gemini**	***p* Value**	**ChatGPT**	**Gemini**	***p* Value**
Accuracy	4.36 ± 0.58	4.35 ± 0.57	0.781	4.51 ± 0.43	4.80 ± 0.27	0.001	4.73 ± 0.32	4.93 ± 0.26	0.058
Clarity	4.72 ± 0.35	4.32 ± 0.59	<0.001	4.88 ± 0.25	4.68 ± 0.41	0.009	4.93 ± 0.18	4.97 ± 0.13	0.564
Relevance	4.54 ± 0.47	4.36 ± 0.53	0.019	4.88 ± 0.24	4.88 ± 0.25	1.000	4.73 ± 0.37	4.97 ± 0.13	0.038
Completeness	4.24 ± 0.62	3.93 ± 0.69	0.001	4.29 ± 0.49	4.58 ± 0.42	0.004	4.43 ± 0.42	4.93 ± 0.18	0.002
Sources & References	1.00	1.00	NA	1.00	1.00	NA	1.00	1.00	NA
Usefulness	4.45 ± 0.52	4.11 ± 0.62	0.001	4.69 ± 0.47	4.68 ± 0.51	0.847	4.40 ± 0.51	4.87 ± 0.23	0.004
QAMAI Total Score	23.31 ± 2.00	22.07 ± 2.35	<0.001	24.26 ± 1.18	24.63 ± 1.26	0.145	24.23 ± 1.25	25.67 ± 0.68	0.001

Group 1: Medical condition- and treatment-related questions. Group 2: Nutrition- and diet-related questions. Group 3: Symptom management-related questions.

In contrast, for nutrition and diet questions, Gemini performed better in accuracy (4.80 ± 0.27 vs. 4.51 ± 0.43, $p = 0.001$) and completeness (4.58 ± 0.42 vs. 4.29 ± 0.49, $p = 0.004$). For symptom management questions, Gemini outperformed ChatGPT in relevance (4.97 ± 0.13 vs. 4.73 ± 0.37, $p = 0.038$), completeness (4.93 ± 0.18 vs. 4.43 ± 0.42, $p = 0.002$), and usefulness (4.87 ± 0.23 vs. 4.40 ± 0.51, $p = 0.004$) (Table 2). Mean QAMAI subdimension scores and mean total QAMAI scores by question groups are shown in Figures 2 and 3.

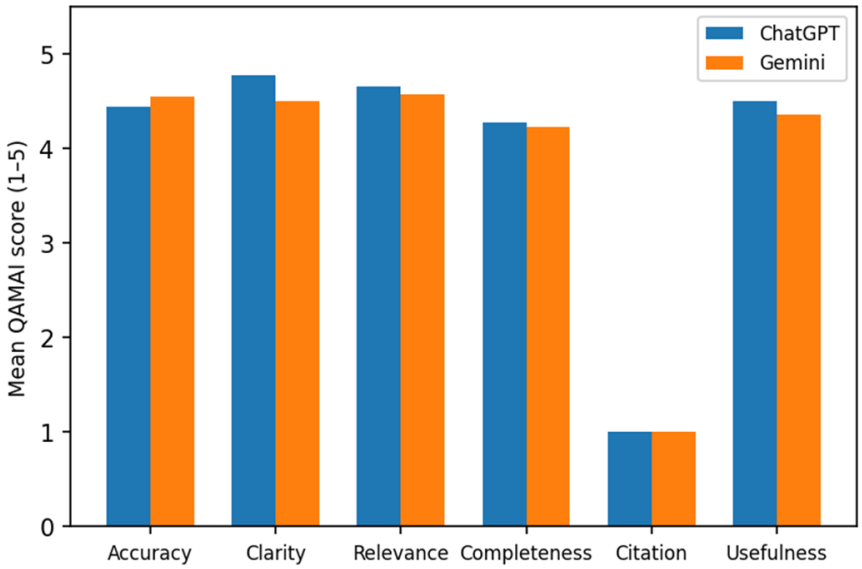

Figure 2. Mean QAMAI subdimension scores.

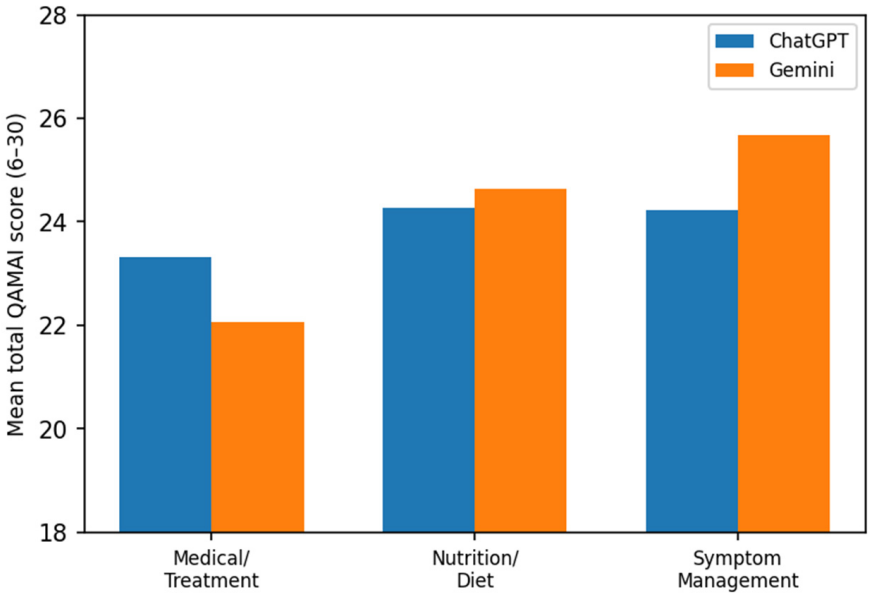

Figure 3. Mean total QAMAI scores by question groups.

3.2. Effect Size Analysis

Effect size analyses using Cohen's d are shown in Table 3. ChatGPT showed a moderate advantage in clarity for medical/treatment-related questions (d = 0.806). In contrast, Gemini demonstrated larger effect sizes in completeness (d = 1.329 in symptom management) and usefulness (d = 1.142 in symptom management).

Table 3. Effect sizes by question groups.

	Group 1			Group 2			Group 3		
Dimension	**ChatGPT**	**Gemini**	**Cohen's d**	**ChatGPT**	**Gemini**	**Cohen's d**	**ChatGPT**	**Gemini**	**Cohen's d**
Accuracy	4.36 ± 0.58	4.35 ± 0.57	0.014	4.51 ± 0.43	4.80 ± 0.27	0.768	4.73 ± 0.32	4.93 ± 0.26	0.702
Clarity	4.72 ± 0.35	4.32 ± 0.59	0.806	4.88 ± 0.25	4.68 ± 0.41	0.572	4.93 ± 0.18	4.97 ± 0.13	0.270
Relevance	4.54 ± 0.47	4.36 ± 0.53	0.380	4.88 ± 0.24	4.88 ± 0.25	0.000	4.73 ± 0.37	4.97 ± 0.13	0.938
Completeness	4.24 ± 0.62	3.93 ± 0.69	0.462	4.29 ± 0.49	4.58 ± 0.42	0.651	4.43 ± 0.42	4.93 ± 0.18	1.329
Usefulness	4.45 ± 0.52	4.11 ± 0.62	0.607	4.69 ± 0.47	4.68 ± 0.51	0.018	4.40 ± 0.51	4.87 ± 0.23	1.142

Group 1: Medical condition- and treatment-related questions. Group 2: Nutrition- and diet-related questions. Group 3: Symptom management-related questions.

3.3. Inter-Rater Reliability

Inter-rater reliability for QAMAI scores was moderate across domains. ICC values for ChatGPT ranged from 0.137 (clarity) to 0.430 (completeness), while Gemini showed slightly higher consistency, ranging from 0.376 (usefulness) to 0.575 (completeness). The overall ICC for QAMAI was 0.395 (95% CI, 0.136–0.577) for ChatGPT and 0.560 (95% CI, 0.371–0.692) for Gemini.

3.4. Consensus Analysis

Consensus analysis, based on the proportion of responses receiving high ratings (scores 4–5), showed that ChatGPT achieved 100% agreement for clarity, compared with 87.8% for Gemini ($p < 0.001$). For completeness, ChatGPT responses were rated highly in 85.4% of evaluations, versus 76.4% for Gemini ($p = 0.054$). Neither model received high scores for source citation, with both consistently obtaining the minimum score across all evaluations.

4. Discussion

This study compared the performance of ChatGPT and Gemini in answering CKD patients' questions. We found that both models produced generally high-quality responses, but with clear, domain-specific strengths. ChatGPT performed better for medical condition and treatment questions, offering clearer and more practically useful explanations, whereas Gemini gave more accurate and complete answers for nutrition and symptom management. Effect size analyses indicated that these differences were clinically meaningful: ChatGPT showed a moderate advantage in clarity for medical/treatment items, while Gemini demonstrated large effects in completeness and usefulness for symptom-related content. Together, these findings support a complementary use of the two models for patient education, provided they are deployed under careful clinical supervision.

Our results are consistent with broader evidence that LLM performance is highly task-dependent rather than dominated by a single "best" model. A recent scoping review of 69 comparative LLM studies reported that ChatGPT-4 was the most accurate model in 21 studies, while ChatGPT-4 and Bard/Gemini tied for readability in 14 studies, although nearly half of the studies did not use standardized evaluation tools [24]. In another study, ChatGPT-4 correctly triaged 93% of simulated nephrology patient messages with high internal consistency, suggesting potential to support patient portal workflows [25]. Together with our QAMAI scores, these data indicate that current LLMs may be ready to augment, but not replace, nephrologists in both patient communication and selected aspects of clinical reasoning.

CKD-focused studies further clarify model-specific strengths. A recent systematic review synthesized 23 nephrology LLM studies and found that Gemini (formerly Bard) achieved the highest Global Quality Score for CKD questions, while Copilot (Bing Chat) outperformed both ChatGPT-3.5 and Gemini (formerly Bard) for nephrology laboratory interpretation [18]. In pediatric CKD, Naz et al. showed that both ChatGPT-3.5 and Gemini (formerly Bard) achieved high F1 scores for diagnosis and lifestyle counseling when benchmarked against Kidney Disease Improving Global Outcomes (KDIGO) guidelines, with Gemini attaining the highest overall quality score [17]. In line with these findings, Gemini in our study outperformed ChatGPT for nutrition and symptom management questions, which closely mirror lifestyle counseling and day-to-day self-management, whereas ChatGPT retained an advantage in clarity and perceived usefulness for treatment-related issues. These converging findings suggest that Gemini may be particularly well-suited for CKD education and day-to-day guidance, while ChatGPT remains strong for abstract medical reasoning and structured explanation [26,27]. Studies in nephrology also show that AI chatbots can be useful educational tools. One study reported that chatbots designed to follow current nephrology guidelines can provide CKD patients with personalized, easy-to-understand information, helping support their education and self-management [28].

From a patient education perspective, clarity and usefulness are particularly important dimensions. Even highly accurate information may not lead to better self-care if patients find it confusing, overly technical, or poorly organized. Our data indicate that ChatGPT's strength lies precisely in translating complex concepts into clear, accessible language. Given that limited health literacy affects nearly one in four CKD patients, ChatGPT's strength in clarity may help reduce knowledge gaps that contribute to poor adherence and worse outcomes [8–10,29]. Recent work suggests that AI use can reduce the reading complexity of kidney donation documents from a 9th-grade to a 4th-grade level using GPT-4, suggesting that LLMs may help address health literacy barriers when properly guided [15]. Likewise, our data suggest that ChatGPT's clearer explanations of disease and treatment may be

particularly useful for patients with limited health literacy, who struggle with complex medical terminology.

However, clarity alone is not sufficient. For domains like nutrition and symptom management—where day-to-day decisions are nuanced and highly contextual—accuracy and completeness are critical. Small errors or omissions in dietary advice may have serious consequences for patients with advanced CKD, especially regarding potassium, phosphorus, sodium, or fluid intake. Gemini's stronger performance in these domains in our study is therefore clinically meaningful. Its more thorough and precise responses in symptom management suggest that it may be particularly useful for addressing common concerns such as pain control, pruritus, or fluid overload, which often drive unplanned healthcare use and reduce quality of life.

A consistent limitation observed in our study and throughout the literature is the lack of reliable citations. Previous studies have shown that many references generated by ChatGPT-3 are fabricated or inaccurate, despite appearing convincing [30]. In clinical settings, one analysis found that 69% of ChatGPT-3.5-generated references were fabricated, and another reported that only 7% were both real and correct [14,31]. This limitation represents a major barrier to safe clinical adoption, as unverifiable sources undermine trust and increase the risk of misinformation [13,30,32]. In our evaluation, both models scored at the minimum level for source citation across all questions, reinforcing concerns that current default configurations are not suitable for unsupervised use as providers of evidence-linked information. It is also important to note that, during our study, advanced retrieval-augmented features such as deep research modes were not used for either model. Future implementations integrated into clinical systems should prioritize retrieval-augmented generation with curated nephrology guidelines (e.g., KDIGO, European Renal Association, and national recommendations) and transparent citation of up-to-date sources.

Our study adds to the still-limited evidence on LLM performance in Turkish. Prior work in Korean, Japanese, Chinese, and Arabic medical settings suggests that accuracy is generally lower for non-English questions and that performance can drop further for open-ended or free-text tasks [19,20,33,34]. These findings highlight important language-related gaps when applying LLMs outside English. Despite these challenges, our data indicate that both ChatGPT and Gemini can produce high-quality Turkish responses to CKD questions, as evaluated by nephrologists using a standardized tool. However, we did not directly compare Turkish and English performance, and QAMAI scores reflect expert evaluations rather than patient comprehension. As a result, subtle linguistic errors, cultural nuances, or misunderstandings may, therefore, be under-recognized. Future studies should explicitly compare Turkish and English answers to the same CKD questions and incorporate patient-reported measures of comprehensibility, trust, and behavioral impact.

Beyond technical performance, our findings have several practical and ethical implications. First, the complementary strengths of ChatGPT and Gemini suggest that a "model-agnostic" strategy may be preferable in clinical practice. Rather than committing to a single LLM, nephrology teams could select or combine models based on the type of information requested—for example, using ChatGPT to generate clear explanations of diagnosis and prognosis and Gemini to provide detailed guideline-aligned recommendations on diet and symptom control. Such a modular approach would require robust governance, including consistent human oversight, clear disclaimers, and mechanisms for continuous quality monitoring. Finally, our study emphasizes the importance of equity considerations. While LLMs could potentially reduce disparities by providing clear explanations to patients who have limited access to specialist care, they might also exacerbate inequities if only digitally literate or affluent patients can use them effectively. Older adults, people with limited literacy, and those with limited internet access may struggle to benefit from these

tools. To ensure equitable implementation, LLM-based education should be integrated into supervised clinical settings—such as dialysis units or CKD clinics—where staff can assist patients in formulating questions, interpreting answers, and applying information to their own circumstances.

Another point is that moderate ICC values for some QAMAI subdimensions do not necessarily indicate poor inter-rater agreement. For areas such as clarity, evaluators frequently assigned very similar scores, leading to low between-item variance and consequently lower ICC estimates, particularly in the context of a two-rater design and a 5-point rating scale.

This study has some limitations. The cross-sectional design captures model performance at a single time point in a rapidly evolving technological landscape. The number of symptom-related questions was relatively small; although statistically significant differences were observed in this subgroup, these findings should be interpreted cautiously due to the limited sample size. Furthermore, evaluations relied on nephrologists rather than patients, which may not fully reflect real-world understanding or behavior change. Future research should include larger and more diverse question sets, integrate patient perspectives, and assess the effects of AI-generated information on clinical outcomes such as adherence, clinic attendance, and hospitalization.

Future research should build on these insights in several ways. Larger, multicenter studies that include diverse CKD populations and languages are needed to confirm our findings and explore regional variations. Longitudinal designs could evaluate how LLM-generated information influences clinical outcomes such as adherence, blood pressure control, dietary adherence, clinic attendance, hospitalization, or time to dialysis initiation. Mixed-methods studies incorporating qualitative interviews and focus groups with patients and caregivers would help illuminate how they perceive, trust, and use AI-generated advice in daily life. Comparative effectiveness studies could examine hybrid models where nephrologists oversee and refine LLM outputs before sharing them with patients, versus traditional education alone. Finally, technical work is required to develop nephrology-specific, retrieval-augmented LLM systems that provide transparent citations and are aligned with up-to-date guidelines.

5. Conclusions

Both ChatGPT and Gemini produced high-quality answers to Turkish CKD patient questions but showed complementary strengths. ChatGPT was superior in clarity and perceived usefulness for medical and treatment queries, whereas Gemini provided more accurate and comprehensive guidance for nutrition and symptom management. The persistent absence of trustworthy citations remains a critical barrier, emphasizing the need for careful implementation, verification mechanisms, and ongoing nephrologist oversight. Continuous monitoring and re-evaluation will be essential to LLMs as their clinical integrations continue to evolve.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/kidneydial6010009/s1. Table S1: Questions collected from the patients; Table S2: Representative examples illustrating the QAMAI scoring process by two independent nephrology specialists.

Author Contributions: S.G.O., Y.B.S. and N.S. conceptualized and designed the study. Y.B.S. was involved in data collection. M.T.D., S.T. and N.S. evaluated the LLMs responses. Z.A. conducted the statistical analysis. S.G.O. and Y.B.S. drafted the manuscript. N.S. and all authors critically reviewed and revised the manuscript for intellectual content. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was approved by the Istanbul University-Cerrahpasa, Cerrahpasa Medical Faulty ethics committee (approval no: E-74555795-050.04-1274848, approval date: 12 March 2025) and conducted according to the Declaration of Helsinki.

Informed Consent Statement: Informed consent was taken from all patients.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bikbov, B.; Purcell, C.; Levey, A.; Smith, M.; Abdoli, A.; Abebe, M.; Adebayo, O.M.; Afarideh, M.; Agarwal, S.K.; Agudelo-Botero, M.; et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **2020**, *395*, 709–733. [CrossRef]
2. Foreman, K.J.; Marquez, N.; Dolgert, A.; Fukutaki, K.; Fullman, N.; McGaughey, M.; Pletcher, M.A.; Smith, A.E.; Tang, K.; Yuan, C.-W.; et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–40 for 195 countries and territories. *Lancet* **2018**, *392*, 2052–2090. [CrossRef]
3. Besarab, A.; Levin, A. Defining a renal anemia management period. *Am. J. Kidney Dis.* **2000**, *36*, S13–S23. [CrossRef]
4. Moe, S.; Drüeke, T.; Cunningham, J.; Goodman, W.; Martin, K.; Olgaard, K.; Ott, S.; Sprague, S.; Lameire, N.; Eknoyan, G. Definition, evaluation, and classification of renal osteodystrophy: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney Int.* **2006**, *69*, 1945–1953. [CrossRef] [PubMed]
5. Lee, G.H.; Benner, D.; Regidor, D.L.; Kalantar-Zadeh, K. Impact of kidney bone disease and its management on survival of patients on dialysis. *J. Ren. Nutr.* **2007**, *17*, 38–44. [CrossRef]
6. Foley, R.N.; Parfrey, P.S.; Sarnak, M.J. Clinical epidemiology of cardiovascular disease in chronic renal disease. *Am. J. Kidney Dis.* **1998**, *32*, S112–S119. [CrossRef]
7. Bibbins-Domingo, K.; Chertow, G.M.; Fried, L.F.; Odden, M.C.; Newman, A.B.; Kritchevsky, S.B.; Harris, T.B.; Satterfield, S.; Cummings, S.R.; Shlipak, M.G. Renal function and heart failure risk in older black and white individuals: The Health, Aging, and Body Composition Study. *Arch. Intern. Med.* **2006**, *166*, 1396–1402. [CrossRef]
8. Billany, R.E.; Thopte, A.; Adenwalla, S.F.; March, D.S.; Burton, J.O.; Graham-Brown, M.P. Associations of health literacy with self-management behaviours and health outcomes in chronic kidney disease: A systematic review. *J. Nephrol.* **2023**, *36*, 1267–1281. [CrossRef] [PubMed]
9. Taylor, D.M.; Fraser, S.D.; Bradley, J.A.; Bradley, C.; Draper, H.; Metcalfe, W.; Oniscu, G.C.; Tomson, C.R.; Ravanan, R.; Roderick, P.J.; et al. A systematic review of the prevalence and associations of limited health literacy in CKD. *Clin. J. Am. Soc. Nephrol.* **2017**, *12*, 1070–1084. [CrossRef]
10. Miao, J.; Thongprayoon, C.; Kashani, K.B.; Cheungpasitporn, W. Artificial intelligence as a tool for improving health literacy in kidney care. *PLoS Digit. Health* **2025**, *4*, e0000746. [CrossRef] [PubMed]
11. Jin, Q.; Leaman, R.; Lu, Z. Retrieve, summarize, and verify: How will ChatGPT affect information seeking from the medical literature? *J. Am. Soc. Nephrol.* **2023**, *34*, 1302–1304. [CrossRef] [PubMed]
12. Yuan, Q.; Zhang, H.; Deng, T.; Tang, S.; Yuan, X.; Tang, W.; Xie, Y.; Ge, H.; Wang, X.; Zhou, Q.; et al. Role of artificial intelligence in kidney disease. *Int. J. Med. Sci.* **2020**, *17*, 970. [CrossRef]
13. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef] [PubMed]
14. Gravel, J.; D'Amours-Gravel, M.; Osmanlliu, E. Learning to fake it: Limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin. Proc. Digit. Health* **2023**, *1*, 226–234. [CrossRef]
15. Valencia, O.A.G.; Thongprayoon, C.; Miao, J.; Suppadungsuk, S.; Krisanapan, P.; Craici, I.M.; Jadlowiec, C.C.; Mao, S.A.; Mao, M.A.; Leeaphorn, N.; et al. Empowering inclusivity: Improving readability of living kidney donation information with ChatGPT. *Front. Digit. Health* **2024**, *6*, 1366967. [CrossRef] [PubMed]
16. Vaira, L.A.; Lechien, J.R.; Abbate, V.; Allevi, F.; Audino, G.; Beltramini, G.A.; Bergonzani, M.; Boscolo-Rizzo, P.; Califano, G.; Cammaroto, G.; 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. *Eur. Arch. Oto-Rhino-Laryngol.* **2024**, *281*, 6123–6131. [CrossRef]
17. Naz, R.; Akacı, O.; Erdoğan, H.; Açıkgöz, A. Can large language models provide accurate and quality information to parents regarding chronic kidney diseases? *J. Eval. Clin. Pract.* **2024**, *30*, 1556–1564. [CrossRef]
18. Unger, Z.; Soffer, S.; Efros, O.; Chan, L.; Klang, E.; Nadkarni, G.N. Clinical applications and limitations of large language models in nephrology: A systematic review. *Clin. Kidney J.* **2025**, *18*, sfaf243. [CrossRef]

19. Yoon, S.-H.; Oh, S.K.; Lim, B.G.; Lee, H.-J. Performance of ChatGPT in the In-Training Examination for Anesthesiology and Pain Medicine Residents in South Korea: Observational Study. *JMIR Med. Educ.* **2024**, *10*, e56859. [CrossRef]
20. Fang, C.; Wu, Y.; Fu, W.; Ling, J.; Wang, Y.; Liu, X.; Jiang, Y.; Wu, Y.; Chen, Y.; Zhou, J.; et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. *PLoS Digit. Health* **2023**, *2*, e0000397. [CrossRef]
21. Ozturk, N.; Yakak, I.; Ağ, M.B.; Aksoy, N. Is ChatGPT reliable and accurate in answering pharmacotherapy-related inquiries in both Turkish and English? *Curr. Pharm. Teach. Learn.* **2024**, *16*, 102101. [CrossRef]
22. OpenAI. *GPT-4o Model Card and Technical Overview*; OpenAI: San Francisco, CA, UAS, 2024; Available online: https://platform.openai.com/docs/models (accessed on 10 January 2026).
23. Google. *Gemini Model Overview and Technical Documentation*; Google DeepMind: Mountain View, CA, USA, 2024; Available online: https://ai.google.dev/gemini-api/docs/models (accessed on 10 January 2026).
24. AlSammarraie, A.; Househ, M. The use of large language models in generating patient education materials: A scoping review. *Acta Inform. Med.* **2025**, *33*, 4. [CrossRef]
25. Pham, J.H.; Thongprayoon, C.; Miao, J.; Suppadungsuk, S.; Koirala, P.; Craici, I.M.; Cheungpasitporn, W. Large language model triaging of simulated nephrology patient inbox messages. *Front. Artif. Intell.* **2024**, *7*, 1452469. [CrossRef]
26. Rossettini, G.; Rodeghiero, L.; Corradi, F.; Cook, C.; Pillastrini, P.; Turolla, A.; Castellini, G.; Chiappinotto, S.; Gianola, S.; Palese, A. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: A cross-sectional study. *BMC Med. Educ.* **2024**, *24*, 694. [CrossRef]
27. Bahir, D.; Zur, O.; Attal, L.; Nujeidat, Z.; Knaanie, A.; Pikkel, J.; Mimouni, M.; Plopsky, G. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefe's Arch. Clin. Exp. Ophthalmol.* **2025**, *263*, 527–536. [CrossRef] [PubMed]
28. Acharya, P.C.; Alba, R.; Krisanapan, P.; Acharya, C.M.; Suppadungsuk, S.; Csongradi, E.; Mao, M.A.; Craici, I.M.; Miao, J.; Thongprayoon, C.; et al. AI-driven patient education in chronic kidney disease: Evaluating chatbot responses against clinical guidelines. *Diseases* **2024**, *12*, 185. [CrossRef]
29. Hancı, V.; Ergün, B.; Gül, Ş.; Uzun, Ö.; Erdemir, İ.; Hancı, F.B. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. *Medicine* **2024**, *103*, e39305. [CrossRef]
30. Athaluri, S.A.; Manthena, S.V.; Kesapragada, V.K.M.; Yarlagadda, V.; Dave, T.; Duddumpudi, R.T.S. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* **2023**, *15*, e37432. [CrossRef] [PubMed]
31. Bhattacharyya, M.; Miller, V.M.; Bhattacharyya, D.; Miller, L.E.; Miller, V. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus* **2023**, *15*, e39238. [CrossRef] [PubMed]
32. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.-L.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: Comparative analysis. *J. Med. Internet Res.* **2024**, *26*, e53164. [CrossRef]
33. Harigai, A.; Toyama, Y.; Nagano, M.; Abe, M.; Kawabata, M.; Li, L.; Yamamura, J.; Takase, K. Response accuracy of GPT-4 across languages: Insights from an expert-level diagnostic radiology examination in Japan. *Jpn. J. Radiol.* **2025**, *43*, 319–329. [CrossRef] [PubMed]
34. Samaan, J.S.; Yeo, Y.H.; Ng, W.H.; Ting, P.-S.; Trivedi, H.; Vipani, A.; Yang, J.D.; Liran, O.; Spiegel, B.; Kuo, A.; et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. *Arab J. Gastroenterol.* **2023**, *24*, 145–148. [CrossRef] [PubMed]