Rev Nefrol Dial Traspl. 2025;45(4):173-178

EVALUACIÓN DEL DESEMPEÑO DE UNA HERRAMIENTA BASADA EN MODELOS DE LENGUAJE GRANDE PARA RECOMENDACIONES NUTRICIONALES EN ENFERMEDAD RENAL CRÓNICA

PERFORMANCE EVALUATION OF A LARGE LANGUAGE MODEL-BASED TOOL FOR NUTRITIONAL RECOMMENDATIONS IN CHRONIC KIDNEY DISEASE

Dr. Carlos Matías Callegari[1], Dr. Gonzalo García [2], Lic. Cristina Milano[1], Lic. Judith Leibovich[1], Lic. Florencia Cardone[1]

ABSTRACT

Introduction: Nutritional management of chronic kidney disease (CKD) is an essential component of treatment; however, its implementation faces multiple challenges due to the complexity of dietary recommendations and the shortage of specialized professionals. Large language models (LLMs) offer the possibility of complementing professional consultations through virtual assistance tools, but their specific performance in the area of renal nutrition has not yet been adequately evaluated. **Objective:** To evaluate the performance of NutriRenal, a virtual assistant based on a large language model adjusted using a prompt designed by experts, through an evaluation by nutritionists specializing in CKD in response to nutrition-related queries from patients with CKD. **Methods:** A descriptive, cross-sectional study was conducted in which three specialized nutritionists evaluated 211 responses generated by NutriRenal to questions formulated by nephrologists. Responses were classified into three dimensions: comprehensibility, completeness, and consistency with scientific evidence, using a scale of 1 to 3. Differences were analyzed before and after the prompt's adjustment, as well as by CKD stage, presence of diabetes, and evaluator. **Results:** After the prompt's adjustment, NutriRenal demonstrated high performance: 99% of responses were rated as adequate in comprehensibility, 86.7% in completeness, and 95.2% in consistency with scientific evidence. These improvements were statistically significant compared to the original prompt. Performance was consistent across the different subgroups evaluated, with patients with diabetes showing the best scores. **Conclusions:** NutriRenal demonstrated robust performance after the rapid adjustment, generating high-quality responses according to the evaluated professional criteria. Its implementation could be a valuable complement to traditional nutritional consultations in patients with CKD. However, further studies in real-world clinical settings are needed to validate its impact on daily clinical practice.

Keywords: Chronic kidney disease; nutrition; large language

Correspondencia:
Dr. Carlos M. Callegari
ORCID:
0000-0001-2009-6957
carlosm.callegari@gmail.com

Financiamiento:
Ninguno.

Conflicto de intereses:
Ninguno que declarar.

Recibido: 08-09-2025
Corregido: 26-09-2025
Aceptado: 08-10-2025

1) Asociación Nefrológica de Buenos Aires
2) Sociedad Argentina de Nefrología

RESUMEN

Introducción: El manejo nutricional de la enfermedad renal crónica (ERC) constituye un componente esencial del tratamiento; sin embargo, su implementación enfrenta múltiples desafíos debido a la complejidad de las recomendaciones dietéticas y la escasez de profesionales especializados. Los modelos de lenguaje grande (LLM, por sus siglas en inglés) ofrecen la posibilidad de complementar la consulta profesional mediante herramientas de asistencia virtual, pero su desempeño específico en el área de nutrición renal aún no ha sido evaluado adecuadamente. **Objetivo:** Evaluar el desempeño de NutriRenal, un asistente virtual basado en un modelo de lenguaje grande ajustado mediante un prompt diseñado por expertos, mediante la evaluación realizada por nutricionistas especializadas en ERC en la respuesta a consultas relacionadas con nutrición en pacientes con ERC. **Métodos:** Se realizó un estudio descriptivo, transversal, donde tres nutricionistas especializadas evaluaron 211 respuestas generadas por NutriRenal a preguntas elaboradas por médicos nefrólogos. Las respuestas fueron clasificadas en tres dimensiones: comprensibilidad, completitud y coherencia con la evidencia científica, utilizando una escala del 1 al 3. Se analizaron las diferencias antes y después del ajuste del prompt, así como por estadio de ERC, presencia de diabetes y evaluador. **Resultados:** Tras el ajuste del prompt, NutriRenal presentó un desempeño elevado: el 99% de las respuestas fueron calificadas como adecuadas en comprensibilidad, el 86,7% en completitud y el 95,2% en coherencia con la evidencia científica. Estas mejoras fueron estadísticamente significativas respecto al prompt original. El desempeño fue homogéneo entre los distintos subgrupos evaluados, destacándose una mejor puntuación en pacientes con diabetes. **Conclusiones:** NutriRenal mostró un desempeño robusto tras el ajuste del prompt, generando respuestas de alta calidad según los criterios profesionales evaluados. Su implementación podría constituir un valioso complemento a la consulta nutricional tradicional en pacientes con ERC. No obstante, son necesarios estudios adicionales en contextos clínicos reales para validar su impacto en la práctica clínica cotidiana.

Palabras Clave: Enfermedad renal cronica; nutrición, lenguaje grande.

INTRODUCCIÓN

La enfermedad renal crónica (ERC) afecta a aproximadamente el 10% de la población mundial y se asocia a una elevada morbilidad, mortalidad y costos sanitarios [1]. La alimentación es uno de los pilares fundamentales en el manejo de esta enfermedad, con impacto directo en la progresión del daño renal, el control de las complicaciones metabólicas y la calidad de vida de los pacientes [2]. Sin embargo, ajustar las recomendaciones nutricionales según el estadio de ERC, considerando además comorbilidades como la diabetes tipo 2, representa un desafío clínico importante que requiere conocimientos específicos y actualizados [3].

En muchos entornos, especialmente en contextos con recursos limitados, los pacientes tienen escaso acceso a nutricionistas especializados en enfermedad renal [4]. Además, las recomendaciones dietéticas suelen variar entre profesionales y no siempre están alineadas con las guías más recientes ni con el entorno alimentario local [5]. Este escenario resalta la necesidad de herramientas escalables, consistentes y contextualizadas que puedan apoyar la toma de decisiones clínicas y mejorar la educación del paciente.

En los últimos años, los modelos de lenguaje de gran escala (LLM) como ChatGPT han demostrado un potencial creciente para aplicaciones en salud, incluyendo el asesoramiento nutricional [6]. Estudios recientes han evaluado su desempeño en diabetes, síndrome metabólico y dieta renal, con resultados prometedores pero también con limitaciones en términos de precisión y consistencia [7,8]. En el ámbito de la nefrología, donde las recomendaciones dietéticas han cambiado significativamente en la última década (por ejemplo, reconsiderando restricciones rígidas de potasio o fósforo), estos modelos requieren una adaptación específica para ser clínicamente útiles [9].

En este contexto se desarrolló NutriRenal, un asistente virtual basado en un LLM ajustado con un prompt diseñado por expertos para proporcionar recomendaciones alimentarias

personalizadas en personas con ERC. El presente estudio tiene como objetivo evaluar su desempeño según criterios cualitativos establecidos por nutricionistas especializadas.

OBJETIVOS

Objetivo General

Evaluar el desempeño de un asistente virtual por nutricionistas especializadas en enfermedad renal crónica para la respuesta a consultas relacionadas a la nutrición en enfermedad renal crónica.

Objetivos Primarios

Evaluar la comprensibilidad, completitud y coherencia con la evidencia científica de las respuestas generadas por el asistente virtual en cada una de las categorías preespecificadas de preguntas.

Objetivos Secundarios

Evaluar el cambio en la comprensibilidad, completitud y coherencia con la evidencia científica posterior a la modificación del prompt.

MATERIALES Y METODOS

Diseño del Estudio

El presente trabajo corresponde a un estudio descriptivo de corte transversal realizado entre marzo y mayo de 2025 en la Ciudad Autónoma de Buenos Aires, Argentina.

Descripción del Asistente Virtual

NutriRenal es un asistente virtual desarrollado sobre el modelo ChatGPT versión 4o, un modelo de lenguaje grande (LLM) del tipo transformador generativo preentrenado. Se le integró un prompt específicamente diseñado para restringir y contextualizar las respuestas dentro del ámbito de la nutrición renal. Este prompt fue elaborado por dos nefrólogos especializadas en nutrición y validado por nutricionistas renales.

La configuración por defecto de temperatura del modelo [10] no fue modificada. La temperatura es un parámetro que regula el grado de aleatoriedad de las respuestas generadas: valores más altos promueven variabilidad, mientras que valores más bajos aumentan la consistencia.

Participantes

Dos médicos nefrólogos especialistas participaron en la creación del prompt y en la elaboración de las 211 preguntas que simulan consultas frecuentes realizadas por pacientes con ERC. Las evaluaciones fueron realizadas por tres licenciadas en nutrición con más de diez años de experiencia en nutrición renal, todas activas en ámbitos académicos y asistenciales.

Evaluación de Desempeño

Las respuestas generadas por NutriRenal fueron evaluadas de forma cualitativa en tres dimensiones:

Comprensibilidad

Claridad y facilidad de interpretación por parte del paciente.

Completitud

Inclusión de los elementos relevantes para una recomendación nutricional adecuada.

Coherencia con la Evidencia Científica

Grado de alineación con guías clínicas y literatura actualizada.

Cada dimensión fue evaluada con una escala ordinal de 1 (baja) a 3 (alta), en base al juicio experto de las nutricionistas. Las respuestas fueron asignadas aleatoriamente y distribuidas equitativamente entre evaluadoras.

Las preguntas fueron clasificadas según tres categorías clínicas de ERC: no avanzada, avanzada y terapia sustitutiva renal, y se consideró también la presencia o ausencia de diabetes tipo 2.

Modificación del Prompt

Tras una primera ronda de evaluación, los autores identificaron errores frecuentes —principalmente la recomendación inadecuada de restringir alimentos no procesados— y reformularon el prompt para enfatizar la diferencia en biodisponibilidad de fósforo y potasio entre alimentos frescos y procesados. Esta modificación dio lugar a una segunda ronda de evaluación con los mismos criterios.

Cálculo Muestral

La muestra de 210 preguntas se determinó estimando una proporción esperada de

concordancia del 80%, un nivel de confianza del 95% y un margen de error del 6%, añadiendo un 20% para cubrir posibles respuestas no evaluables.

Análisis Estadístico

Se calcularon proporciones de respuestas calificadas como "3" en cada dimensión, comparando resultados entre el prompt inicial y el ajustado, entre categorías clínicas, presencia de diabetes y entre evaluadoras. El análisis se realizó utilizando Python en Google Colab.

RESULTADOS

Desempeño global antes y después del ajuste del prompt

Se obtuvieron 211 respuestas para la evaluación cualitativa. Con el prompt inicial, la proporción de respuestas calificadas como "3" fue de 75,9% en comprensibilidad, 52,2% en completitud y 64% en coherencia con la evidencia científica. Luego del ajuste del prompt, estas proporciones aumentaron a 99%, 86,7% y 95,2% respectivamente. Las diferencias fueron estadísticamente significativas en las tres dimensiones ($p < 0,001$) **(Gráfico 1).**

Este mejor desempeño del prompt ajustado se mantuvo en cada categoría individual evaluada, con diferencias estadísticamente significativas en las proporciones de respuestas calificadas como 1, 2 y 3, favoreciendo sistemáticamente al prompt ajustado **(Gráfico 2)**.

Grafico 1. Evaluación global del prompt 1 vs prompt 2.

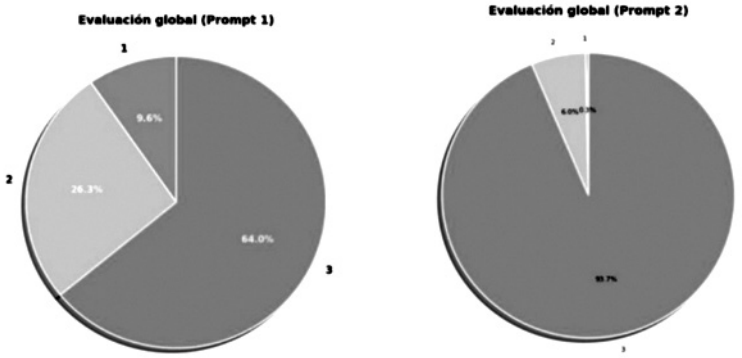
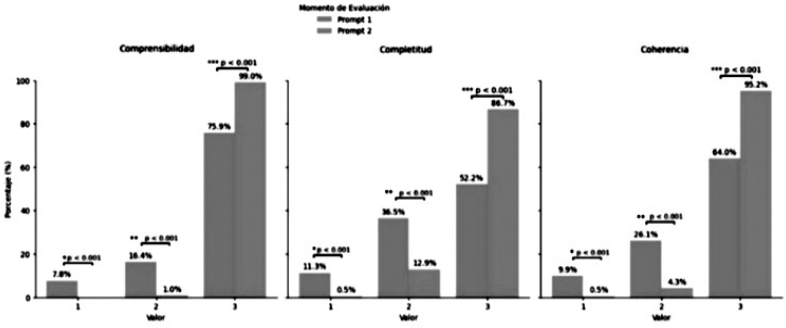


Gráfico 2. Desempeño del prompt 1 vs prompt 2 por categoría.



**Diferencia estadísticamente significativa entre categoría "1" del Prompt 1 y Prompt 2; **Diferencia estadísticamente significativa entre categoría "2" del Prompt 1 y Prompt 2; ***Diferencia estadísticamente significativa entre categoría "3" del Prompt 1 y Prompt 2*

Desempeño por Grupo de Pacientes

No se observaron diferencias estadísticamente significativas en la proporción de respuestas calificadas como "3" en ninguna de las tres dimensiones entre las distintas categorías de enfermedad renal crónica (no avanzada, avanzada y terapia sustitutiva renal).

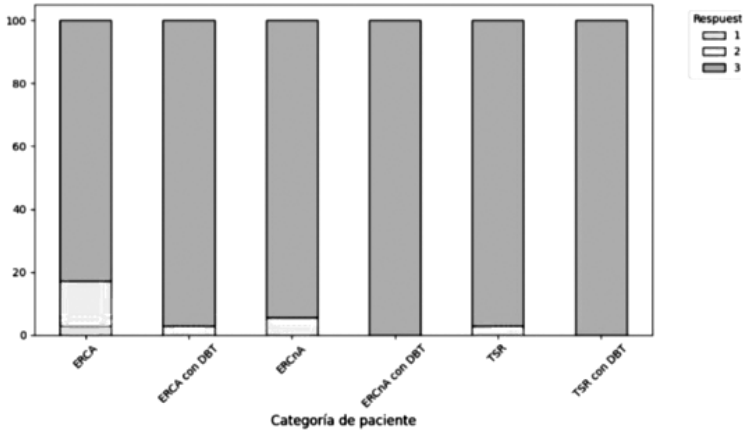
Desempeño Según Presencia de Diabetes Tipo 2

Las respuestas dirigidas a pacientes con diabetes tipo 2 fueron mejor calificadas en coherencia con la evidencia científica (99%) que las dirigidas a pacientes sin diabetes (91%), con una diferencia estadísticamente

 ISSN 0326-3428

significativa (p = 0,018; OR = 9,75). En las otras dimensiones no se observaron diferencias significativas **(Gráfico 3).**

Gráfico 3. Coherencia con evidencia científica por categoría de paciente



Desempeño Según Etapa del Modelo (antes y después del ajuste del prompt)

Las evaluaciones realizadas tras la modificación del prompt mostraron un desempeño consistentemente superior en todas las dimensiones. Las proporciones de respuestas calificadas como "3" aumentaron significativamente, lo cual confirma la efectividad del ajuste realizado en el diseño del prompt.

Variabilidad Entre Evaluadoras

Se observó una buena concordancia general entre las nutricionistas. No se detectaron diferencias estadísticamente significativas en las puntuaciones promedio otorgadas por cada evaluadora.

Interacción Simulada: Caso de Repregunta

Aunque el estudio evaluó respuestas únicas por pregunta, se documentó un ejemplo que sugiere un posible mejor desempeño en escenarios iterativos. Ante la consulta "¿puedo consumir pollo diariamente?", la respuesta inicial fue adecuada pero incompleta. Al re-preguntar por cantidad, el modelo amplió correctamente la información, sugiriendo que su utilidad podría ser mayor en interacciones clínicas reales.

DISCUSIÓN

Este estudio evaluó el desempeño de NutriRenal, un asistente virtual basado en un modelo de lenguaje de gran escala (LLM) ajustado con un prompt diseñado por expertos, para responder consultas relacionadas con la nutrición en pacientes con enfermedad renal crónica (ERC). Los resultados evidenciaron un alto desempeño del modelo tras la modificación del prompt, especialmente en términos de comprensibilidad, completitud y coherencia con la evidencia científica.

A diferencia de otros trabajos recientes que exploran el uso de LLMs en nutrición, NutriRenal se apoya en un proceso deliberado de ajuste del prompt con participación activa de profesionales especialistas en nefrología y nutrición. Esto contrasta con estudios previos donde los modelos fueron evaluados sin un entrenamiento o personalización específicos, lo cual puede explicar los resultados más inconsistentes reportados en esas investigaciones [7,8,9].

El abordaje de NutriRenal —centrado en aspectos cualitativos y alineados con principios actualizados de la nutrición renal— evitó errores frecuentes documentados en otros modelos, como la subestimación de nutrientes críticos o la aplicación de restricciones alimentarias obsoletas [9]. Además, la inclusión de criterios como el índice potasio/fibra, la biodisponibilidad del fósforo, el índice inflamatorio dietético y el índice glucémico, aporta solidez clínica a las recomendaciones generadas.

A nivel subgrupal, el modelo mostró un desempeño homogéneo entre estadios de ERC y modalidades de tratamiento, lo cual refuerza su aplicabilidad transversal. La mejor performance

observada en pacientes con diabetes podría estar relacionada con una mayor disponibilidad de guías estandarizadas o un mayor nivel de precisión en los criterios nutricionales para este grupo.

Una de las fortalezas metodológicas del estudio fue la evaluación del modelo en seis perfiles clínicos diferentes, lo que permite explorar su robustez frente a variabilidad en el input clínico. Asimismo, el proceso de evaluación por parte de nutricionistas expertas en ERC otorga validez técnica a los resultados obtenidos.

Sin embargo, el estudio presenta limitaciones relevantes. En primer lugar, NutriRenal fue implementado como un GPT sin posibilidad de configurar la temperatura, un parámetro que afecta la consistencia de las respuestas. En contextos médicos, se recomienda trabajar con temperatura cero para favorecer la reproducibilidad, algo que podría mejorarse en futuras versiones mediante la creación de un agente con control de parámetros [10].

En segundo lugar, la evaluación se basó en respuestas únicas por pregunta. En la práctica clínica, las interacciones suelen ser iterativas, lo que podría mejorar aún más la calidad de las recomendaciones, como se observó en la repregunta del caso sobre consumo diario de pollo.

Por último, el modelo aún no ha sido evaluado en interacción con pacientes reales, y su impacto en la adherencia dietética, la comprensión de las recomendaciones o la relación médico-paciente no ha sido medido. Estos aspectos serán clave en futuras investigaciones.

En conjunto, los resultados sugieren que herramientas como NutriRenal pueden alcanzar niveles muy altos de desempeño cuando se restringen a un dominio clínico bien definido y se ajustan mediante conocimiento experto. Su integración como soporte en la consulta nutricional renal es prometedora, pero requiere validación adicional en entornos clínicos reales y control de variabilidad en la generación de respuestas.

CONCLUSIONES

NutriRenal es una herramienta basada en un modelo de lenguaje de gran escala ajustado por expertos que ha demostrado un alto desempeño en la generación de recomendaciones nutricionales para personas con enfermedad renal crónica. Su desempeño mejoró significativamente tras la modificación del prompt, alcanzando niveles altos de comprensibilidad, completitud y coherencia con la evidencia científica.

Estos resultados respaldan su utilidad potencial como herramienta de apoyo en el asesoramiento nutricional renal, particularmente en contextos donde el acceso a nutricionistas especializados es limitado. No obstante, serán necesarios estudios adicionales que evalúen su impacto en escenarios clínicos reales, su desempeño en interacciones iterativas, y la implementación de versiones más consistentes mediante agentes configurables.

BIBLIOGRAFÍA

1) García-García G, Jha V, Li PKT, Couser WG, Okpechi IG. Chronic kidney disease (CKD) in disadvantaged populations. *Nat Rev Nephrol.* 2023;19(1):1–2.
2) KDIGO. Clinical Practice Guideline for Diabetes Management in CKD. *Kidney Int.* 2020;98(4S):S1–S115.
3) National Kidney Foundation. *Nutrition and Chronic Kidney Disease*. 2022. https://www.kidney.org/nutrition
4) Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management: A review. *JAMA*. 2018;320(12):1294–1304.
5) Campbell KL, Ash S, Bauer JD, Davies PS, Johnson DW. Impact of nutrition intervention on quality of life in pre-dialysis CKD patients. *Clin Nutr.* 2022;41(2):403–410.
6) Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. *Basic Books; 2019.*
7) Naja F, Taktouk M, Matbouli D, *et al*. AI chatbots for nutrition management of diabetes and metabolic syndrome. *Eur J Clin Nutr.* 2024;78(10):887–896.
8) Qarajeh A, Tangpanithandee S, Thongprayoon C, *et al*. AI-powered renal diet support: ChatGPT, Bard AI, Bing Chat. *Clin Pract.* 2023;13(5):1160–1172.
9) Ponzo V, Rosato R, Scigliano MC, *et al*. Accuracy and consistency of AI chatbots in nutritional advice. *J Clin Med.* 2024;13(24):7810.
10) Miao Y, *et al*. Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. 2024.