SahuarUS
Revista electrónica de Matemáticas
Vol. 9 / No. 1 / Julio 2025, pp. 34-41
SahuarUS

Evaluación del Potencial de LLM (ChatGPT-4) como Métrica Objetiva para Calidad de Imagen en CT de Mama

Maritza Callejas-Cortés[1], Irving O. Ayala-Iturbe[1], María M. Méndez-González[1], Juan P. Cruz-Bastida[1]

[1]Departamento de Física, Escuela Superior de Física y Matemáticas, Instituto Politécnico Nacional, CDMX, México.

[1]E-mail: jcruzb@ipn.mx

Resumen

El cáncer de mama es una de las principales causas de mortalidad en mujeres a nivel mundial, lo que resalta la importancia de contar con imágenes médicas de calidad para mejorar su diagnóstico. Este estudio evalúa el uso de modelos de lenguaje grande (LLM), específicamente ChatGPT-4 para caracterizar la calidad diagnóstica de imágenes de tomografía computarizada de mama (bCT). Se diseñó un maniquí numérico que simula un corte transversal de la mama, a partir del cual se generaron 100 imágenes con variaciones en calidad de imagen. Se realizó una prueba de observador comparando el desempeño de ChatGPT-4 con seis lectores con conocimientos de Física Médica, calificando aspectos como contraste, nitidez, ruido y artefactos mediante una escala de Likert. Aunque el LLM mostró un desempeño limitado frente al observador de referencia, alcanzó concordancias ligeras cuando se comparó con observadores menos experimentados. Estos resultados sugieren que, con un protocolo de entrenamiento más robusto, los LLM podrían aproximarse al observador humano en tareas de caracterización de calidad de diagnóstica.

Palabras Clave: LLM; ChatGPT; Cáncer de Mama; breast CT.

DOI: 10.36788/sah.v9i1.160
Recibido: 30 de enero de 2025.
Aceptado: 23 de junio de 2025.

1. Introducción

El cáncer de mama (CaMa) es una enfermedad caracterizada por el crecimiento descontrolado de células en el tejido mamario, formando tumores que pueden invadir tejidos circundantes y diseminarse a otras partes del cuerpo. Aunque afecta principalmente a mujeres, también puede presentarse en hombres, aunque en menor proporción. Esta enfermedad representa una de las principales causas de muerte por cáncer en mujeres a nivel mundial, subrayando la importancia de las estrategias de prevención y diagnóstico temprano [9]. En

ISSN 2448 5365
Universidad de Sonora

México, el CaMa constituye un grave problema de salud pública. De acuerdo con cifras preliminares del Instituto Nacional de Estadística y Geografía (INEGI), en 2023 se registraron 8,034 defunciones por esta enfermedad en personas mayores de 20 años, de las cuales el 99.5 % correspondieron a mujeres [8].

El diagnóstico temprano del CaMa es fundamental para mejorar las tasas de supervivencia y reducir la mortalidad asociada a esta enfermedad. Detectar el cáncer en sus etapas iniciales permite tratamientos menos invasivos, más efectivos y menos costosos [15]. Existen diversos métodos para la detección temprana del CaMa. Las mamografías son la herramienta principal y más recomendada para mujeres a partir de los 40 años, ya que pueden identificar tumores antes de que sean palpables [3]. Aunque la mamografía es una herramienta esencial en la detección temprana del cáncer de mama, presenta ciertas limitaciones; por ejemplo, una menor sensibilidad en mujeres con mamas densas [5]. Esto ha impulsado la exploración de tecnologías más avanzadas, como la tomografía computarizada de mama (bCT). La bCT es una técnica de imagen avanzada que ofrece imágenes tridimensionales de alta resolución del tejido mamario. A diferencia de la mamografía convencional, la bCT elimina la superposición de tejidos, lo que facilita la detección de lesiones pequeñas y mejora la precisión diagnóstica [10].

Independientemente de la modalidad de imagen utilizada, el éxito en el diagnóstico médico depende en gran medida de la calidad de las imágenes obtenidas. Una imagen de alta calidad permite una lectura precisa y confiable, facilitando la identificación de patologías. Por lo tanto, la evaluación de la calidad de imagen en radiología es crucial para garantizar diagnósticos confiables. Para medir esta calidad, se emplean diversas métricas objetivas, como la razón contraste-ruido (CNR), que evalúa la capacidad de diferenciar estructuras de interés del fondo. Sin embargo, aunque estas métricas brindan información valiosa, no siempre correlacionan directamente con la percepción del observador humano, el cual se considera estándar de oro [4]. Los estudios basados en la interpretación humana, como los análisis de curvas de características operativas del receptor (ROC), son esenciales para evaluar la calidad de imagen; no obstante, es un enfoque costoso en recursos y tiempo [13]. Por ello, existe un interés creciente en desarrollar métricas alternativas que puedan aproximarse al juicio humano, haciendo más eficiente la evaluación de calidad de imagen.

Por lo tanto, el objetivo de este trabajo es evaluar el potencial de los modelos de lenguaje grande (LLM), particularmente ChatGPT-4 [2], como observadores subrogados para la evaluación de la calidad diagnóstica de imágenes de bCT. En este estudio, se investigó si las capacidades avanzadas de procesamiento de lenguaje y razonamiento contextual de ChatGPT pueden ser aprovechadas para analizar características clave de las imágenes médicas como el ruido, el contraste, la resolución espacial y la presencia de artefactos.

2. Materiales y Métodos

La Fig. 1 presenta un resumen visual de la metodología empleada en este trabajo. Como primer paso, se diseñó un maniquí numérico que simula un corte transversal de la mama, tomando en cuenta las propiedades radiológicas de la piel, el tejido adiposo y el tejido glan-

Figura 1: Diagrama de flujo de la metodología empleada en el estudio. Se muestra el diseño del maniquí numérico, la simulación de imágenes de bCT y los estudios de observador.

dular. Este maniquí, con un diámetro de 11 cm, está definido en términos de coeficientes de atenuación lineal (μ) [7]. Además, se incluyeron en la estructura dos grupos de lesiones: cúmulos de microcalcificaciones redondas y carcinomas espiculados, ambos con tamaños que van desde 1 mm hasta 1 cm.

Posteriormente, se realizaron dos simulaciones de adquisición de imágenes de bCT de dicho maniquí utilizando el software FastCat [12]. La simulación replicó las condiciones geométricas y los parámetros de adquisición típicos de un equipo clínico de bCT [1], considerando dos niveles de dosis de radiación: 6 y 1.5 mGy. Las imágenes simuladas fueron reconstruidas empleando métodos de reconstrucción analítica, como FDK [6], y métodos iterativos como SART-TV [14], variando los parámetros de reconstrucción. Como resultado, se creó una base de datos con 100 imágenes del mismo maniquí para las cuales se variaron parámetros que pueden influir en la calidad de la imagen adquirida, generando variaciones en aspectos como contraste, resolución, presencia de artefactos y niveles de ruido.

Con ayuda de dicha base de datos, se llevó a cabo una prueba de observador en la que se evaluó el desempeño de un LLM: **GPT-4-turbo** (OpenAI, versión del modelo: gpt-4-2024-04-09) en comparación con seis observadores humanos con formación en Física Médica, considerando al más experimentado como referencia. Los observadores humanos analizaron un conjunto de 50 imágenes y calificaron cinco aspectos de calidad: nitidez, magnitud del ruido, contraste de las lesiones, presencia de artefactos y textura del ruido. La evaluación se realizó utilizando una escala de Likert, como se ilustra en la Fig. 1.

Para la evaluación del LLM, se inició una conversación para contextualizar la tarea. Se le proporcionó información sobre radiología, con un enfoque específico en bCT, y se le mostraron imágenes del maniquí numérico como referencia. Posteriormente, se le explicó la prueba de manera similar a los observadores humanos, presentándole seis imágenes de entrenamiento acompañadas de los puntajes asignados por el observador de referencia. Un ejemplo de las imágenes de entrenamiento se muestra en la Fig. 2.

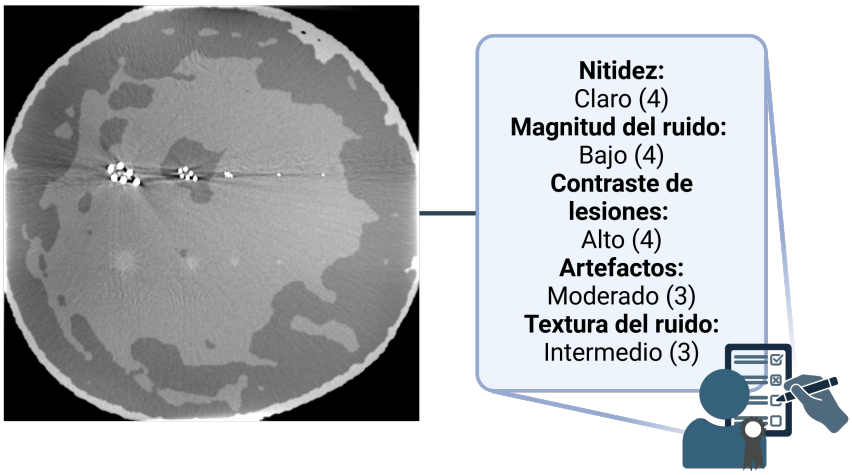

Figura 2: Imagen de entrenamiento del LLM acompañada de los puntajes asignados por el observador de referencia. El prompt utilizado en el entrenamiento fue: *La siguiente imagen fue calificada en Nitidez: Claro (4) Magnitud del ruido: Bajo (4) Contraste de lesiones: Alto (4) Artefactos: Moderado (3) Textura del ruido: Intermedio (3)*
.

Finalmente, después del entrenamiento, se utilizó el siguiente prompt para la evaluación: *"Te mostraré varias imágenes y de acuerdo con lo que has aprendido en los ejemplos que te proporcioné anteriormente, haz una evaluación de la calidad de imagen evaluando los siguientes criterios con la siguiente escala. Nitidez: en una escala del 1 al 5, siendo 1 nada, 2 pobre, 3 moderado, 4 claro y 5 muy claro; Magnitud del ruido: en una escala del 1 al 5, siendo 1 muy alto, 2 alto, 3 moderado, 4 bajo y 5 muy bajo; Contraste de lesiones: en una escala del 1 al 5, siendo 1 muy bajo, 2 bajo, 3 moderado, 4 alto y 5 muy alto; Presencia de artefacto afectando tu visibilidad: en una escala del 1 al 5, siendo 1 bastante, 2 considerable, 3 moderado, 4 muy poco y 5 nada; Granulosidad del ruido: en una escala del 1 al 5, siendo 1 muy grueso, 2 grueso, 3 intermedio, 4 fino y 5 muy fino."*

Con esta indicación, el modelo calificó el resto de las 50 imágenes de prueba para comparar su concordancia con los observadores humanos.

Se calculó el índice κ [11] con el fin de evaluar la concordancia entre el LLM y los observadores, así como en comparación con el observador de referencia. El índice κ se define como

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{1}$$

donde: P_0 representa la proporción de concordancia observada entre los evaluadores. P_e es la proporción de concordancia esperada al azar. El índice κ tiene un rango de -1 a 1, un valor de 1 indica concordancia perfecta; un valor 0 indica concordancia equivalente al azar y valores negativos sugieren desacuerdo entre los evaluadores.

3. Resultados

La Fig. 3 presenta un resumen detallado de los resultados obtenidos.

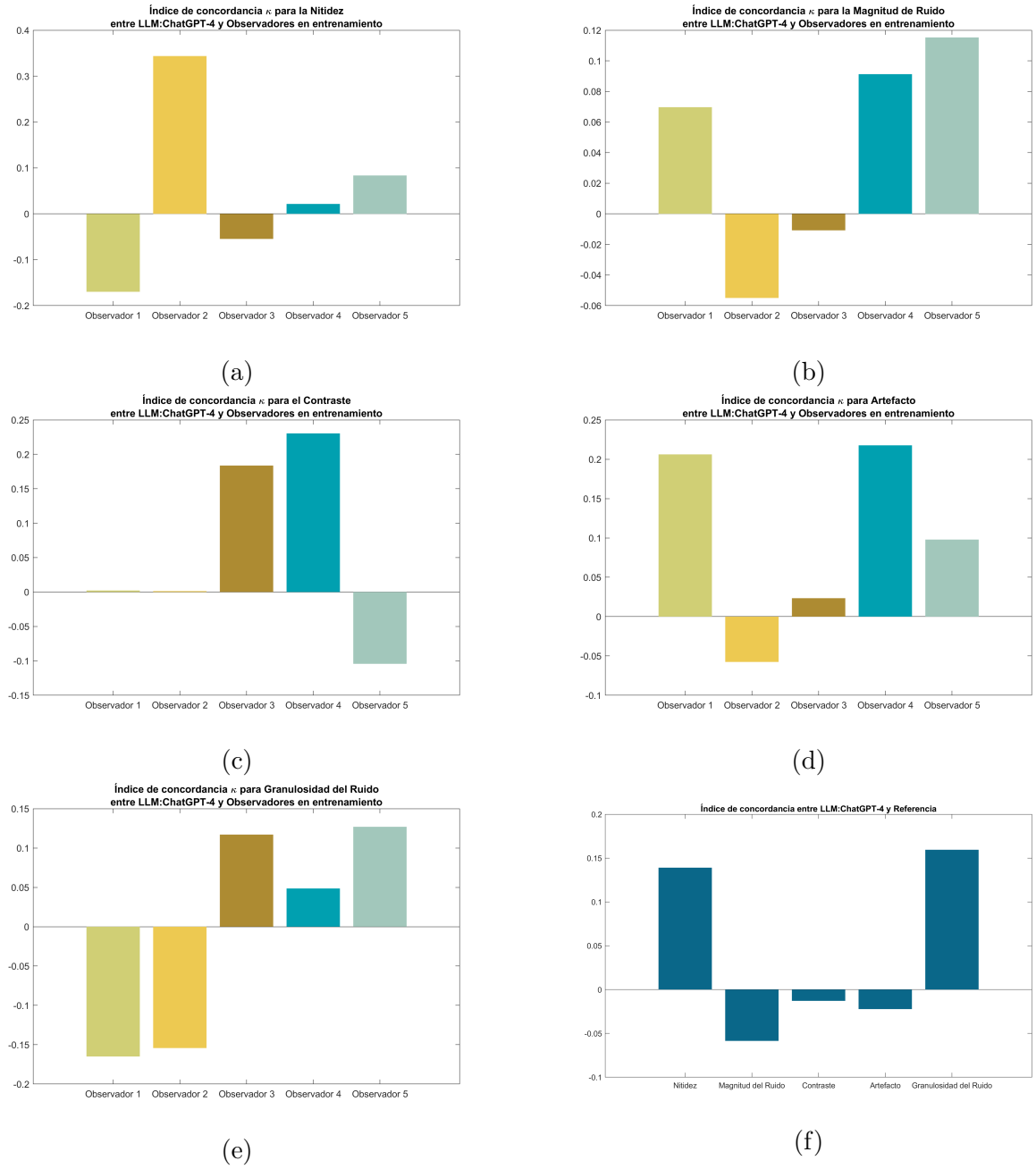

Figura 3: Resumen de los resultados de la evaluación de concordancia entre observadores. Las Figs. 3a and 3e muestran los valores de κ obtenidos para un observador específico en los aspectos evaluados. La Fig. 3f muestra el desempeño global del observador de referencia.

Las Figs. 3a and 3e ilustran los valores de κ para un observador específico en cada uno de

los siguientes aspectos evaluados: nitidez, magnitud del ruido, contraste de lesiones, presencia de artefactos y textura del ruido, respectivamente. La Fig. 3f muestra una integración de los valores de κ correspondiente al observador de referencia, proporcionando una visión global de su desempeño.

Según los criterios de interpretabilidad de Landis y Koch [11], el LLM muestra un desempeño pobre en la mayor parte de la prueba comparado con el observador de referencia, aunque logra una ligera concordancia en nitidez y granularidad de ruido. Sin embargo, al compararlo con otros observadores, su desempeño varía entre ligero y aceptable en al menos 60 % de las comparaciones, lo que sugiere cierta concordancia entre el LLM y los observadores más inexpertos.

4. Discusión y Conclusiones

Aunque el desempeño global del LLM para replicar la respuesta de observadores humanos fue limitado, es importante señalar que solo se utilizaron 6 imágenes de referencia, además del entrenamiento base, para que el modelo aprendiera a realizar esta tarea de imitación. Si bien la decisión de utilizar un conjunto relativamente pequeño de imágenes responde a los objetivos de una prueba de concepto, el bajo desempeño del modelo puede interpretarse como una consecuencia de la asimetría entre la limitada experiencia del modelo y el entrenamiento de al menos 6 meses con el que cuentan los observadores humanos. Sin embargo, al margen del desempeño global, el nivel de concordancia en categorías específicas, como la nitidez y la granulosidad del ruido, sugiere que un protocolo de aprendizaje más exhaustivo y robusto del LLM podría permitir una mayor aproximación a la respuesta en evaluadores humanos.

Es importante destacar que, aunque los LLM basados en redes transformer, como ChatGPT, están diseñados para interpretar imágenes en general, no están optimizados específicamente para la interpretación de imágenes médicas. Por ello, podría ser conveniente comenzar evaluándolos en tareas más sencillas. En trabajos futuros, se planea explorar el potencial de estos modelos en actividades de detección y clasificación básica, evaluando si su desempeño puede alcanzar niveles de concordancia con los observadores humanos en estos contextos simplificados.

Agradecimientos

MCC e IOAI agradecen el apoyo de la SECIHTI para la realización de sus estudios de maestría.

Referencias

[1] "nu:view – AB-CT – Advanced Breast-CT," https://ab-ct.com/nuview/, 2024.

[2] "OpenAI," https://openai.com, 2024.

[3] American Cancer Society, "Detección temprana y diagnóstico del cáncer de seno," https://www.cancer.org/es/cancer/tipos/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno.html, 2024.

[4] H. H. Barrett and K. J. Myers, *Foundations of Image Science*. Wiley-Interscience, 2003.

[5] N. F. Boyd, H. Guo, L. J. Martin, L. Sun, J. Stone, E. Fishell, and M. J. Yaffe, "Mammographic density and the risk and detection of breast cancer," *New England Journal of Medicine*, vol. 356, no. 3, pp. 227–236, 2007. DOI: 10.1056/NEJMoa062790

[6] L. Feldkamp, L. Davis, and J. Kress, "Practical cone-beam algorithm," *Journal of the Optical Society of America A*, vol. 1, no. 6, pp. 612–619, 1984. DOI: 10.1364/JOSAA.1.000612

[7] J. H. Hubbell and S. M. Seltzer, "Table 3 of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients (version 1.4)," https://physics.nist.gov/PhysRefData/XrayMassCoef/tab3.html, 2004, National Institute of Standards and Technology, Gaithersburg, MD.

[8] Instituto Nacional de Estadística y Geografía (INEGI), "A propósito del Día Internacional de la Lucha contra el Cáncer de Mama," https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2024/EAP_LuchaCMama24.pdf, 2024.

[9] Instituto Nacional del Cáncer, "Cáncer de mama," https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/cancer-de-mama, 2024.

[10] T. E. Komolafe, C. Zhang, O. A. Olagbaju, G. Yuan, Q. Du, M. Li, J. Zheng, and X. Yang, "Comparison of diagnostic test accuracy of cone-beam breast computed tomography and digital mammography for detecting breast cancer: A systematic review and meta-analysis," *Sensors*, vol. 22, no. 9, p. 3594, 2023. DOI: 10.3390/s22093594

[11] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. DOI: 10.2307/2529310

[12] J. O'Connell and M. Bazalova-Carter, "fastCAT: Fast cone beam CT (CBCT) simulation," *Medical Physics*, vol. 48, no. 8, pp. 4448–4458, 2021. DOI: 10.1002/mp.15007

[13] S. Park, J. Goo, and C. Jo, "Receiver operating characteristic (ROC) curve: practical review for radiologists," *Korean Journal of Radiology*, vol. 5, no. 1, pp. 11–18, 2004. DOI: 10.3348/kjr.2004.5.1.11

[14] E. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," *Physics in Medicine & Biology*, vol. 53, no. 17, pp. 4777–4807, 2008. DOI: 10.1088/0031-9155/53/17/021

[15] World Health Organization (WHO), "Breast cancer – key facts," https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2024.

Como citar este artículo: M. Callejas Cortés, I. O. Ayala Iturbe, M. M. Méndez-González, & J. P. Cruz Bastida, "Evaluación del Potencial de LLM (ChatGPT-4) como Métrica Objetiva para la Calidad de Imagen en CT de Mama", SahuarUS. Revista Electrónica de Matemática, vol. 9, no. 1, pp. 34–41, 2025. https://doi.org/10.36788/sah.v9i1.161