ARTÍCULO ORIGINAL
ISSN 1870-7203

Desempeño comparativo entre modelos de inteligencia artificial y médicos residentes en un examen clínico tipo ABIM

Comparative performance between artificial intelligence models and medical residents on an ABIM-style clinical exam

César Adolfo Nieves Pérez,*,§§ Federico Leopoldo Rodríguez Weber,‡ Miguel Cuauhtémoc Molina Obana,*,¶¶ Juan Carlos Núñez Hernández,§ Alejandro Rivera Tapia,¶ Alejandro Rojas Montaño,|| Axel Corona Deschamps,** Enrique Juan Díaz Greene‡‡

Citar como: Nieves PCA, Rodríguez WFL, Molina OMC, Núñez HJC, Rivera TA, Rojas MA et al. Desempeño comparativo entre modelos de inteligencia artificial y médicos residentes en un examen clínico tipo ABIM. Acta Med GA. 2026; 24 (2): 118-124. https://dx.doi.org/10.35366/122614

Resumen

Este estudio evaluó el desempeño académico de cuatro modelos de inteligencia artificial (ChatGPT-4, Gemini 2.5, Claude 3.7 y DeepSeek R1) y médicos residentes de Medicina Interna al resolver un examen clínico tipo ABIM. Se compararon los promedios de respuestas correctas entre grupos, considerando también la consistencia de sus resultados. Gemini 2.5 obtuvo el puntaje más alto (98.3%, DE = 1.76), seguido de Claude 3.7 (93.3%, DE = 2.11), ChatGPT-4 (92.7%, DE = 2.00) y DeepSeek R1 (90.7%, DE = 3.06). En contraste, los residentes alcanzaron un promedio significativamente menor (60.4%, DE = 12.04). Todas las IA superaron estadísticamente a los residentes; Gemini 2.5 mostró diferencias significativas frente a los otros modelos. Las menores desviaciones estándar en los modelos de IA indican una mayor consistencia en sus respuestas frente a la amplia variabilidad observada en el grupo humano.

Palabras clave: inteligencia artificial, ChatGPT, modelos de lenguaje grande, desempeño académico, ABIM.

Abstract

This study evaluated the academic performance of four artificial intelligence language models (ChatGPT-4, Gemini 2.5, Claude 3.7, and DeepSeek R1) and Internal Medicine residents in solving an ABIM-style clinical examination. Mean accuracy rates were compared across groups, and within-group consistency was also assessed. Gemini 2.5 achieved the highest score (98.3%, SD = 1.76), followed by Claude 3.7 (93.3%, SD = 2.11), ChatGPT-4 (92.7%, SD = 2.00), and DeepSeek R1 (90.7%, SD = 3.06). In contrast, residents achieved a significantly lower mean score (60.4%, SD = 12.04). All AI models significantly outperformed residents; Gemini 2.5 also showed statistically significant differences compared with the other AI models. The lower standard deviations observed among AI models indicate greater response consistency relative to the wide variability in the human group.

Keywords: artificial intelligence, ChatGPT, large language models, academic performance, ABIM.

* Residente de Medicina Interna, Hospital Angeles Pedregal (HAP). Facultad Mexicana de Medicina de la Universidad La Salle. Ciudad de México, México.
‡ Profesor adjunto del curso de Medicina Interna, HAP. Ciudad de México, México. ORCID: 0000-0001-5680-4743
§ Residente de Oncología Médica, Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán". Ciudad de México, México. ORCID: 0009-0007-7137-0315
¶ Pasante médico de Servicio Social, Universidad del Valle de México. Ciudad de México, México. ORCID: 0000-0002-0671-371X
|| Adscrito de Nefrología, HAP. Ciudad de México, México. ORCID: 0009-0008-0602-9927
** Adscrito de Medicina Interna, HAP. Ciudad de México, México. ORCID: 0009-0007-9212-7322
‡‡ Profesor titular del curso de Medicina Interna, HAP. Ciudad de México, México. ORCID: 0000-0003-2449-9662

ORCID:
§§ 0009-0009-5554-6127
¶¶ 0000-0002-8030-3161

Correspondencia:
Dr. César Adolfo Nieves Pérez
Correo electrónico: nievescesar96@gmail.com

Recibido: 11-04-2025. *Aceptado:* 21-05-2025.

www.medigraphic.com/actamedica

Abreviaturas:
ABIM = *American Board of Internal Medicine* (Junta Estadounidense de Medicina Interna)
IA = inteligencia artificial
MKSAP = *Medical Knowledge Self-Assessment Program* (Programa de Autoevaluación del Conocimiento Médico)
USMLE = *United States Medical Licensing Examination* (Examen de Licencia Médica de los Estados Unidos)

INTRODUCCIÓN

El uso de la inteligencia artificial (IA) ha emergido como una herramienta clave en salud, mejorando la eficiencia clínica y los desenlaces. En 1950, Alan Turing propuso simular el pensamiento humano con el uso de máquinas.[1] En 1956, John McCarthy acuñó el término "inteligencia artificial", anticipando su potencial para igualar la inteligencia humana.[1,2]

Desde sus inicios, la IA ha dado lugar a desarrollos notables como el brazo robótico de General Motors, el programa ELIZA, bases de datos como PubMed y sistema diagnóstico como CASNET, MYCIN, INTERNIST-1 y DXplain.[1-3] En años más recientes, plataformas como IBM Watson han demostrado su capacidad para diagnosticar enfermedades complejas.[1,2]

Actualmente, el desarrollo de modelos de IA ha generado debate sobre su utilidad y la posibilidad de reemplazar funciones médicas humanas.[4,5] Su uso se ha enfocado al diagnóstico por imagen, electrodiagnóstico y pruebas genéticas, particularmente en enfermedades oncológicas, neurológicas y cardiovasculares.[4,6]

Desde su lanzamiento en 2022, ChatGPT (un modelo generativo preentrenado tipo transformer, basado en técnicas de aprendizaje automático y procesamiento de lenguaje natural, que permite interacciones conversacionales complejas)[3,6] ha sido ampliamente evaluado. GPT-4, entrenado con datos públicos hasta septiembre de 2021, ha demostrado conocimiento clínico.[7] Estudios compararon su desempeño con médicos en varios contextos. En Israel, GPT-4 superó a médicos en exámenes de certificación.[7] En Polonia, GPT-3.5 aprobó el examen final de medicina varias veces. En España, GPT-4 logró un 86.8% en el examen MIR (Médico Interno Residente), superando a GPT-3.5.[8,9] En Estado Unidos, GPT-4 obtuvo resultados cercanos al aprobado en el *United States Medical Licensing Examination* (USMLE), destacando en pasos clínicos.[10] En Alemania, GPT-4 alcanzó un 85% en el examen de licencia médica, superando el promedio estudiantil.[11] Se exploraron además sus habilidades interpersonales; en preguntas del USMLE sobre habilidades blandas, GPT-4 tuvo un 90% de precisión, mejor que GPT-3.5 y usuarios de AMBOSS.[12]

Estudios recientes evalúan modelos de lenguaje en medicina. ChatGPT o1 (septiembre 2024) mejoró en razonamiento complejo frente a GPT-4.[12,13] GPT-4 (73.3%) y Claude 2 (54.4%) superaron en nefrología a modelos abiertos, destacando su utilidad.[14,15] En el examen nacional de licencia médica de Japón, GPT-4o (89.2% general, 95% en preguntas fáciles) superó a Claude 3, Gemini 1.5 y GPT-4, respaldando su valor educativo.[15,16]

Pese a la creciente evidencia sobre ChatGPT-4, faltan estudios que comparen directamente su rendimiento académico con otros modelos avanzados (Claude 3.7, Gemini 2.5, DeepSeek R1) en evaluaciones médicas formales. Este estudio busca evaluar el desempeño académico de estos cuatro modelos de IA y el de residentes de Medicina Interna en un examen tipo ABIM (*American Board of Internal Medicine*), analizando las diferencias entre las IA para evaluar su precisión, consistencia y potencial educativo complementario.

MATERIAL Y MÉTODOS

Diseño del estudio

Se llevó a cabo un estudio observacional, de corte transversal, con el objetivo de evaluar el desempeño académico de modelos de inteligencia artificial (ChatGPT-4, Claude 3.7, Gemini 2.5 y DeepSeek R1) y de residentes de Medicina Interna, utilizando un instrumento tipo ABIM. El análisis evaluó precisión, variabilidad y diferencias estadísticas.

Instrumento de evaluación

Se utilizó como instrumento de evaluación un cuestionario de 30 preguntas de opción múltiple, seleccionadas del banco de preguntas MKSAP (*Medical Knowledge Self-Assessment Program*), una herramienta reconocida y validada para la preparación del examen ABIM. El cuestionario fue diseñado para evaluar conocimientos clínicos en medicina interna y asegurar una distribución temática representativa. Para ello, se incluyeron tres preguntas de cada una de las siguientes 10 subespecialidades: Cardiología, Endocrinología, Gastroenterología, Hematología, Infectología, Nefrología, Neurología, Oncología, Neumología y Reumatología. Todas las preguntas seguían el formato de opción múltiple con una única respuesta correcta, buscando mantener una dificultad homogénea acorde a los estándares del ABIM.

Participantes

Se incluyó en el estudio a 38 médicos residentes del programa de Medicina Interna del *Angeles Health System*, distribuidos por año de residencia de la siguiente manera: 13 de primer año (R1), 10 de segundo año (R2), 9 de tercer año (R3) y 6 de cuarto año (R4). La selección se realizó

mediante un muestreo por conveniencia, asegurando la participación voluntaria y anónima, y obteniendo consentimiento informado previo. Se realizó una comparación entre cinco grupos: un grupo humano, conformado por residentes de Medicina Interna, y cuatro modelos de lenguaje de IA, los cuales fueron evaluados a través de sus interfaces web oficiales entre marzo y abril de 2025. Estos modelos incluyeron: ChatGPT-4 (OpenAI), Claude (Anthropic, versión 3.7), Gemini (Google DeepMind, versión 2.5) y DeepSeek (DeepSeek AI, versión R1).

Procedimiento

La recolección de datos siguió protocolos distintos para los modelos de IA y los participantes humanos. Para evaluar los modelos de IA, se administró el cuestionario de 30 preguntas a cada modelo en 10 pruebas independientes. Cada prueba se realizó en una sesión de interacción nueva (iniciada desde cero, con historial limpio o diferente cuenta) para evitar el arrastre de contexto o memoria conversacional entre evaluaciones. Se utilizó un *prompt* estandarizado para realizar cada pregunta a cada una de las IA. La opción de respuesta seleccionada por cada modelo fue registrada manualmente. Este enfoque de múltiples ensayos tuvo como finalidad evaluar la homogeneidad de las respuestas generadas con IA.

Para los médicos residentes, el cuestionario se administró en una única sesión por participante, utilizando la plataforma digital Socrative (Socrative Inc., USA). La prueba se realizó bajo condiciones controladas, con un límite de tiempo estricto de 40 minutos. Durante la evaluación, no se permitió a los residentes el acceso a materiales de consulta externos ni se les proporcionó ningún tipo de retroalimentación sobre el acierto o error en sus respuestas.

Análisis estadístico

El análisis estadístico de los datos se realizó utilizando el software IBM SPSS Statistics (versión 30.0). Se calcularon estadísticas descriptivas, incluyendo media y desviación estándar del porcentaje de respuestas correctas para cada uno de los cinco grupos (cuatro modelos de IA y el grupo de residentes). Para fines del análisis comparativo inferencial, las 10 puntuaciones obtenidas para cada modelo de IA se trataron como observaciones individuales, resultando en $N = 10$ por cada modelo de IA y $N = 38$ para el grupo de residentes.

Dado que la prueba de Levene para la homogeneidad de varianzas resultó significativa ($p < 0.001$), se identificó heterocedasticidad entre los grupos. Adicionalmente, se evaluó la normalidad de las distribuciones por grupo mediante la prueba de Shapiro-Wilk. Los resultados mostraron que los modelos ChatGPT-4, Claude 3.7 y Gemini 2.5 presentaron distribuciones no normales ($p < 0.001$), mientras que DeepSeek R1 ($p = 0.191$) y el grupo de residentes ($p = 0.431$) mostraron distribuciones compatibles con la normalidad. Por esta razón, se decidió emplear un análisis de varianza robusto de Welch para evaluar las diferencias globales en el rendimiento entre los grupos.

Posteriormente, se realizaron comparaciones múltiples *post hoc* entre pares de grupos mediante la prueba de Games-Howell, adecuada para varianzas desiguales. Se estableció un nivel de significancia alfa de $p < 0.05$ para todas las pruebas. Adicionalmente, se utilizó la desviación estándar intragrupo como una medida descriptiva de la variabilidad (consistencia) del desempeño dentro de cada grupo.

RESULTADOS

Se evaluó el desempeño de cinco grupos en un examen médico tipo ABIM: cuatro modelos de inteligencia artificial (ChatGPT 4, Gemini 2.5, Claude 3.7 y DeepSeek R1) y un grupo de residentes humanos. En promedio, los modelos de IA obtuvieron mejores resultados que los residentes. El modelo con mayor puntaje fue Gemini 2.5, con una media de 98.33 puntos (DE = 1.76), seguido de Claude 3.7 (M = 93.33) y ChatGPT 4 (M = 92.67). Por otro lado, el grupo de residentes obtuvo un promedio considerablemente más bajo, con 60.43 puntos (DE = 12.04) *(Tabla 1 y Figura 1)*.

Dado que se encontró una diferencia significativa en la variabilidad de los resultados (prueba de Levene: $p < 0.001$), se aplicó un análisis estadístico (ANOVA de Welch) que confirmó diferencias importantes entre los grupos ($F[4, 22.20] = 85.29$, $p < 0.001$). El tamaño del efecto fue alto ($\eta^2 = 0.799$; $\omega^2 = 0.785$), lo que indica que el tipo de grupo (IA vs residentes) explica cerca del 80% de la variabilidad observada en el desempeño.

El análisis *post hoc* (Games-Howell) mostró que Gemini 2.5 superó significativamente a los demás modelos de IA, incluyendo a ChatGPT 4 (diferencia media = 5.66 puntos, IC95%: 3.03-8.30), a Claude 3.7 ($\Delta = 5.00$, IC95%: 3.13-6.87) y a DeepSeek R1 ($\Delta = 7.67$, IC95%: 3.83-11.50), con $p < 0.001$ en todos los casos *(Tabla 2)*.

En cambio, no se encontraron diferencias estadísticamente significativas entre ChatGPT 4, Claude 3.7 y DeepSeek R1, lo que sugiere un desempeño similar entre ellos ($p > 0.05$). Todos los modelos de IA tuvieron un rendimiento significativamente superior al de los residentes humanos, con diferencias que oscilaron entre 30 y 40 puntos ($p < 0.001$).

Finalmente, al analizar los subgrupos de residentes por año de formación (R1 a R4), se observó una progresión en el rendimiento académico, siendo los residentes de cuarto

año (R4) quienes obtuvieron el promedio más alto entre los humanos (69.6%, DE = 14.3), en contraste con los de primer año (R1), que registraron el puntaje más bajo (57.6%, DE = 9.1). Sin embargo, ninguno de los subgrupos alcanzó los resultados obtenidos por los modelos de inteligencia artificial, todos con promedios superiores al 90%. Las comparaciones *post hoc* mediante la prueba de Games-Howell confirmaron que los R1 fueron significativamente superados por todos los modelos de IA ($p < 0.001$), mientras que los R4 mostraron diferencias significativas sólo frente a Gemini 2.5 ($p = 0.039$) y DeepSeek R1 ($p = 0.048$), pero no frente a ChatGPT-4 ni Claude 3.7 ($p > 0.05$). La comparación directa entre los grupos R1 y R4 no alcanzaron significancia estadística ($p = 0.592$), aunque se identificó una tendencia a mejor desempeño con el avance en la formación clínica. Estas diferencias, no obstante, se vieron acompañadas de una alta variabilidad intragrupo entre los residentes, lo cual se refleja en sus amplios intervalos de confianza *(Tabla 3)*.

DISCUSIÓN

En este estudio se comparó el desempeño académico ante un examen clínico estandarizado de conocimientos entre cuatro modelos de inteligencia artificial (ChatGPT-4, Claude 3.7, Gemini 2.5 y DeepSeek R1) y residentes de Medicina Interna, utilizando un instrumento tipo ABIM. Los resultados mostraron que todos los modelos de inteligencia artificial obtuvieron puntajes superiores al del grupo de residentes, lo cual se correlaciona con hallazgos previos en Israel, España y Alemania.[7-11]

Un hallazgo relevante fue la menor variabilidad intragrupo en las respuestas de los modelos de IA, lo cual puede atribuirse a su consistencia algorítmica. Esto contrasta con la heterogeneidad natural entre humanos, influenciada por factores como experiencia clínica, preparación individual y estados emocionales. Entre los modelos, Gemini 2.5 fue el de mejor rendimiento, superando significativamente a los demás.

Estos hallazgos tienen implicaciones relevantes para la educación médica, particularmente en el diseño de herramientas de apoyo al aprendizaje. Los modelos de IA podrían funcionar como tutores virtuales, asistentes para simulación clínica o instrumentos complementarios en la preparación para exámenes, siempre bajo supervisión crítica por parte de profesionales humanos.

El estudio presenta limitaciones importantes relacionadas con el tamaño y la representatividad de la muestra. El número de casos por grupo fue reducido, especialmente para los modelos de IA (n = 10 por modelo), y todos los residentes pertenecen a un sólo centro, lo que limita la generalización de los hallazgos. Aunque se utilizó el mismo instrumento de evaluación para todos los participantes, las

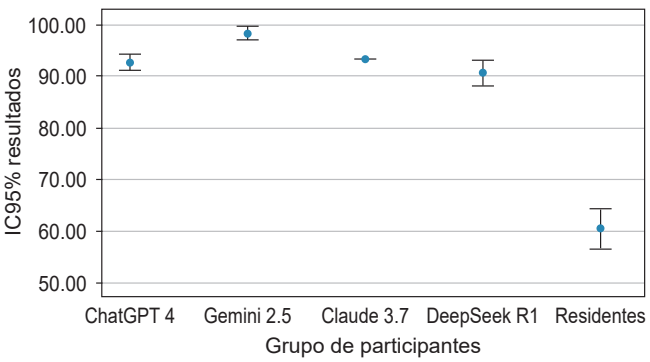

Figura 1: Medias del rendimiento. Puntajes promedio obtenidos por cada grupo participante (modelos de IA y residentes) con sus respectivos intervalos de confianza al 95% (IC95%).

Tabla 1: Estadísticos descriptivos del rendimiento por grupo.

Participantes	N	Media ± DE	EE	IC95%	Rango	Varianza entre componentes
ChatGPT 4	10	92.6667 ± 2.10819	0.66667	91.1586-94.1748	86.67-93.33	
Gemini 2.5	10	98.3333 ± 1.75682	0.55556	97.0766-99.5901	96.67-100.00	
Claude 3.7	10	93.3330 ± 0.00105	0.00033	93.3322-93.3338	93.33-93.33	
DeepSeek R1	10	90.6667 ± 3.44265	1.08866	88.2039-93.1294	86.67-96.67	
Residente	38	60.4321 ± 12.03695	1.95265	56.4757-64.3886	35.00-90.00	
Total	78	77.5182 ± 18.89347	2.13926	73.2583-81.7780	35.00-100.00	
Modelo: Efectos fijos		DE = 8.70782	0.98597	75.5531-79.4832		
Modelo: Efectos aleatorios			11.03054	46.8925-108.1438		39,823,361

Se presentan el número de casos (N), media ± desviación estándar (DE), error estándar (EE) e intervalo de confianza al 95% (IC9%) para la media del rendimiento en cada grupo (modelos de IA y residentes).

Tabla 2: Comparaciones múltiples *post hoc* entre grupos mediante el test de Games-Howell.

(I) Grupo de participantes	(J) Grupo de participantes	Diferencia media (I-J)	EE	p	IC95%
ChatGPT 4	Gemini 2.5	–5.66667	0.86781	**< 0.001**	–8.2997 - –3.0337
	Claude 3.7	–0.66663	1.06667	0.983	–3.9441 - 2.6108
	DeepSeek R1	0.7	1.27557	0.991	–3.1347 - 4.5347
	Residente	32.23456	1.22054	**< 0.001**	28.4443 - 36.0248
Gemini 2.5	ChatGPT 4	5.66667	0.86781	**< 0.001**	3.0337 - 8.2997
	Claude 3.7	5.00003	1.12222	**0.001**	1.9532 - 8.0468
	DeepSeek R1	6.36667	1.22522	**< 0.001**	3.1527 - 9.5806
	Residente	37.90123	1.26127	**< 0.001**	34.0462 - 41.7563
Claude 3.7	ChatGPT 4	0.66663	1.06667	0.983	–2.6108 - 3.9441
	Gemini 2.5	–5.00003	1.12222	**0.001**	–8.0468 - –1.9532
	DeepSeek R1	1.36664	1.31887	0.837	–2.9348 - 5.6681
	Residente	32.90120	1.36782	**< 0.001**	27.7304 - 38.0720
DeepSeek R1	ChatGPT 4	–0.7	1.27557	0.991	–4.5347 - 3.1347
	Gemini 2.5	–6.36667	1.22522	**< 0.001**	–9.5806 - –3.1527
	Claude 3.7	–1.36664	1.31887	0.837	–5.6681 - 2.9348
	Residente	31.53456	1.39328	**< 0.001**	27.3656 - 35.7035
Residente	ChatGPT 4	–32.23456	1.22054	**< 0.001**	–36.0248 - –28.4443
	Gemini 2.5	–37.90123	1.26127	**< 0.001**	–41.7563 - –34.0462
	Claude 3.7	–32.90120	1.36782	**< 0.001**	–38.4989 - –27.3034
	DeepSeek R1	–30.23456	2.23563	**< 0.001**	–36.5841 - –23.8850

Se presentan las diferencias de medias entre pares de grupos, con su error estándar (EE), significancia (p) y el intervalo de confianza al 95% (IC9%). Las diferencias estadísticamente significativas ($p < 0.05$) están marcadas con **negritas**. Comparaciones no significativas indican grupos con rendimiento similar.

condiciones bajo las cuales se administró la prueba fueron marcadamente distintas entre IA y humanos.

Las IA respondieron el examen en 10 ocasiones independientes, sin límite de tiempo, sin exposición a fatiga o ansiedad, y con acceso completo a su base de conocimiento entrenado. En contraste, los residentes realizaron el examen en una única sesión, bajo una estricta limitación de tiempo (40 minutos), sin acceso a recursos externos y sometidos a presión cognitiva y emocional. Esta asimetría metodológica favorece a las IA, por lo que los resultados deben interpretarse como una estimación de su techo de rendimiento en condiciones ideales, más que como una comparación directa del conocimiento clínico neto.

Si bien se observó una alta consistencia en las respuestas de los modelos, ésta no debe asumirse como una propiedad generalizable de toda la inteligencia artificial; aunque hayan alcanzado un nivel notable de desempeño en el ámbito médico, no todos los modelos ofrecen la misma capacidad para resolver problemas clínicos complejos. Cada modelo fue construido con diferentes arquitecturas, corpus de entrenamiento, mecanismos de alineación y principios éticos, lo que influye en su forma de razonar, interpretar preguntas y producir respuestas. También deben considerarse los sesgos relacionados con el *input*: pequeñas variaciones en la redacción de las preguntas pueden alterar significativamente las respuestas generadas por los modelos.

Asimismo, el presente estudio se centró exclusivamente en resultados cuantitativos, sin explorar dimensiones cualitativas como el razonamiento clínico, la toma de decisiones en contextos dinámicos o la interacción médico-paciente. Estos elementos son esenciales para evaluar la verdadera utilidad clínica de cualquier herramienta de apoyo basada en IA.

Estudios futuros deberían usar cuestionarios más amplios y por especialidad, evaluando el impacto de la IA en la práctica clínica, su capacidad para justificar diagnósticos y colaborar con médicos. También es valioso analizar si reformular preguntas incorrectas o pedir justificaciones mejora la comprensión de su lógica, evaluando así precisión y calidad explicativa para su integración educativa y clínica.

CONCLUSIONES

Los modelos de inteligencia artificial evaluados en este estudio demostraron un desempeño superior al de los residentes de Medicina Interna en una prueba tipo ABIM,

Tabla 3: Comparaciones Games-Howell entre modelos de IA y subgrupos de residentes (R1-R4).

(I) Nuevo	(J) Nuevo	Diferencia de medias (I-J)	EE	p	IC95%
ChatGPT 4	Gemini 2.5	–5.66667	0.86781	**< 0.001**	–8.6384 - –2.6949
	Claude 3.7	–0.66633	0.66667	0.963	–3.2270 - 1.8943
	DeepSeek R1	2	1.27657	0.762	–2.4626 - 6.4626
	R1	35.04897	2.62252	**< 0.001**	25.7565 - 44.3415
	R2	34.11467	3.95452	**< 0.001**	19.1239 - 49.1055
	R3	32.21444	4.34475	**< 0.001**	15.2329 - 49.1960
	R4	23.03333	5.85874	0.091	–3.9323 - 49.9989
Gemini 2.5	ChatGPT 4	5.66667	0.86781	**< 0.001**	2.6949 - 8.6384
	Claude 3.7	5.00033	0.55556	**< 0.001**	2.8665 - 7.1342
	DeepSeek R1	7.66667	1.22222	**< 0.001**	3.3238 - 12.0095
	R1	40.71564	2.59650	**< 0.001**	31.4615 - 49.9698
	R2	39.78133	3.93731	**< 0.001**	24.7981 - 54.7645
	R3	37.88111	4.32909	**< 0.001**	20.8989 - 54.8633
	R4	28.70000	5.84714	**0.039**	1.6972 - 55.7028
Claude 3.7	ChatGPT 4	0.66633	0.66667	0.963	–1.8943 - 3.2270
	Gemini 2.5	–5.00033	0.55556	**< 0.001**	–7.1342 - –2.8665
	DeepSeek R1	2.66633	1.08866	0.321	–1.5152 - 6.8478
	R1	35.71531	2.53637	**< 0.001**	26.5351 - 44.8955
	R2	34.78100	3.89792	**< 0.001**	19.8093 - 49.7527
	R3	32.88078	4.29330	**< 0.001**	15.8918 - 49.8698
	R4	23.69967	5.82068	0.083	–3.3921 - 50.7914
DeepSeek R1	ChatGPT 4	–2	1.27657	0.762	–6.4626 - 2.4626
	Gemini 2.5	–7.66667	1.22222	**< 0.001**	–12.0095 - –3.3238
	Claude 3.7	–2.66633	1.08866	0.321	–6.8478 - 1.5152
	R1	33.04897	2.76014	**< 0.001**	23.5009 - 42.5971
	R2	32.11467	4.04709	**< 0.001**	17.0584 - 47.1709
	R3	30.21444	4.42918	**0.001**	13.2151 - 47.2138
	R4	21.03333	5.92162	0.124	–5.7488 - 47.8154
R1	ChatGPT 4	–35.04897	2.62252	**< 0.001**	–44.3415 - –25.7565
	Gemini 2.5	–40.71564	2.5965	**< 0.001**	–49.9698 - –31.4615
	Claude 3.7	–35.71531	2.53637	**< 0.001**	–44.8955 - –26.5351
	DeepSeek R1	–33.04897	2.76014	**< 0.001**	–42.5971 - –23.5009
	R2	–0.93431	4.65048	1.000	–17.0252 - 15.1565
	R3	–2.83453	4.98654	0.999	–20.5372 - 14.8681
	R4	–12.01564	6.34930	0.592	–38.1539 - 14.1226
R2	ChatGPT 4	–34.11467	3.95452	**< 0.001**	–49.1055 - –19.1239
	Gemini 2.5	–39.78133	3.93731	**< 0.001**	–54.7645 - –24.7981
	Claude 3.7	–34.78100	3.89792	**< 0.001**	–49.7527 - –19.8093
	DeepSeek R1	–32.11467	4.04709	**< 0.001**	–47.1709 - –17.0584
	R1	0.93431	4.65048	1.000	–15.1565 - 17.0252
	R3	–1.90022	5.79881	1.000	–21.8803 - 18.0799
	R4	–11.08133	7.00530	0.751	–37.6942 - 15.5315
R3	ChatGPT 4	–32.21444	4.34475	**< 0.001**	–49.1960 - –15.2329
	Gemini 2.5	–37.88111	4.32909	**< 0.001**	–54.8633 - –20.8989
	Claude 3.7	–32.88078	4.29330	**< 0.001**	–49.8698 - –15.8918
	DeepSeek R1	–30.21444	4.42918	**0.001**	–47.2138 - –13.2151
	R1	2.83453	4.98654	0.999	–14.8681 - 20.5372
	R2	1.90022	5.79881	1.000	–21.8803 - 18.0799
	R4	–9.18111	7.23276	0.891	–36.2745 - 17.9122
R4	ChatGPT 4	–23.03333	5.85874	0.091	–49.9989 - 3.9323
	Gemini 2.5	–28.70000	5.84714	**0.039**	–55.7028 - –1.6972
	Claude 3.7	–23.69967	5.82068	0.083	–50.7914 - 3.3921
	DeepSeek R1	–21.03333	5.92162	0.124	–47.8154 - 5.7488
	R1	12.01564	6.34930	0.592	–14.1226 - 38.1539
	R2	11.08133	7.00530	0.751	–15.5315 - 37.6942
	R3	9.18111	7.23276	0.891	–17.9122 - 36.2745

Diferencias de medias, error estándar (EE), significancia (p) e intervalos de confianza al 95% (IC95%) para comparaciones par a par entre modelos de IA y subgrupos de residentes (R1-R4) en el examen tipo ABIM (*American Board of Internal Medicine*).
Las diferencias estadísticamente significativas ($p < 0.05$) están marcadas con **negritas**.

como era esperado. Estos hallazgos sugieren que las IA avanzadas poseen un potencial significativo como herramientas complementarias en la educación médica. Sin embargo, se observaron diferencias entre los modelos actuales en cuanto a su rendimiento y la variabilidad en las respuestas proporcionadas.

Este trabajo plantea la pregunta de si los residentes mejorarían sus resultados al repetir el examen en múltiples ocasiones o al realizarlo con libro abierto o con acceso a una base de datos médica confiable.

Futuras investigaciones deberán explorar no sólo el rendimiento de estas herramientas en otras áreas del conocimiento médico, sino también su utilidad en la práctica clínica, el desarrollo de habilidades blandas, el razonamiento clínico y su integración en entornos reales de enseñanza-aprendizaje. Una posible aplicación sería utilizar estos exámenes para evaluar a las IAs documentando el razonamiento detrás de cada respuesta incorrecta, así como la bibliografía que la respalda.

Las IA no logran un 100% de respuestas correctas debido a su dependencia de datos de entrenamiento, interpretación probabilística del lenguaje, falta de razonamiento clínico contextual y posibles sesgos algorítmicos. Algunos errores podrían originarse en el diseño del examen o en factores relacionados con los residentes al realizarlo.

Estas limitaciones subrayan la importancia de complementar el uso de la inteligencia artificial con el juicio crítico y la supervisión de profesionales humanos.

AGRADECIMIENTOS

Al Dr. Paolo Alberti Minutti por su orientación metodológica.

REFERENCIAS

1. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointest Endosc*. 2020; 92: 807-812.
2. Hirani R, Noruzi K, Khuram H, Hussaini AS, Aifuwa EI, Ely KE et al. Artificial intelligence and healthcare: a journey through history, present innovations, and future possibilities. *Life (Basel)*. 2024; 14 (5): 557.
3. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV et al. A review of the role of artificial intelligence in healthcare. *J Pers Med*. 2023; 13 (6): 951.
4. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S et al. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc Neurol*. 2017; 2: 230-243.
5. Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomed Mater Devices*. 2023: 1-8.
6. Chakraborty C, Bhattacharya M, Pal S, Lee SS. From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. *Curr Res Biotechnol*. 2024; 7: 100164.
7. Katz Katz U, Cohen E, Shachar E, Somer J, Fink A, Morse E et al. GPT versus resident physicians — a benchmark based on official board scores. *NEJM AI*. 2024; 1 (5): AIdbp2300192.
8. Suwala S, Szulc P, Guzowski C, Kaminska B, Dorobiala J, Wojciechowska K et al. ChatGPT-3.5 passes Poland's medical final examination-Is it possible for ChatGPT to become a doctor in Poland? *SAGE Open Med*. 2024; 12: 20503121241257777.
9. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W et al. Evaluating the efficacy of chatgpt in navigating the spanish medical residency entrance examination (MIR): promising horizons for ai in clinical medicine. *Clin Pract*. 2023; 13 (6): 1460-1487.
10. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT performance on USMLE sample items and implications for assessment. *Acad Med*. 2024; 99 (2): 192-197.
11. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. *JMIR Med Educ*. 2024; 10: e50965.
12. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023; 13 (1): 16492.
13. OpenAI. Introducing OpenAI o1 [Internet]. 2024. Available in: https://openai.com/index/introducing-openai-o1-preview/
14. OpenAI. Learning to reason with LLMs [Internet]. 2024. Available in: https://openai.com/index/learning-to-reason-with-llms/
15. Wu S, Koo M, Blum L, Black A, Kao L, Fei Z et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI*. 2024; 1 (2): AIdbp2300092.
16. Liu M, Okuhara T, Dai Z, Huang W, Okada H, Furukawa E et al. Performance of advanced large language models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese medical licensing examination: A comparative study [Internet]. medRxiv; 2024.

Si desea consultar los datos complementarios de este artículo, favor de dirigirse a **editorial.actamedica@saludangeles.mx**