Revista SOCAMPAR

ORIGINAL

FIABILIDAD DIAGNOSTICA DE LOS MODELOS DE LENGUAJE BASADOS EN INTELIGENCIA ARTIFICIAL

Diagnostic Reliability of Artificial Intelligence–Based Language Models

Autores: ***Qiheng Zhou [1], José Luis Izquierdo Alonso [1,2]***

1. ***Departamento de Medicina y Especialidades Médicas, Universidad de Alcalá, Alcalá de Henares. Madrid***
2. ***Servicio de Neumología, Hospital Universitario de Guadalajara, Guadalajara.***

RESUMEN:

Introducción: Los modelos de inteligencia artificial generativa basados en grandes volúmenes de datos textuales (LLM) han mostrado potencial en el ámbito médico, pero su rendimiento diagnóstico, especialmente en español, sigue poco explorado.

Métodos: Se evaluaron las herramientas ChatGPT 4o y 5, DeepSeek y Grok usando 20 casos clínicos del NEJM en su versión original en inglés y traducidos al castellano. Se introdujeron seis *prompts* por caso (tres en castellano, tres en inglés), evaluando diagnósticos diferenciales, diagnóstico final y consistencia en la respuesta.

Resultados: Los modelos alcanzaron un 83 % de aciertos en diagnósticos diferenciales. La precisión disminuyó al 50–60 % al solicitar un diagnóstico único, excepto en ChatGPT 5 logrando un 70% en castellano. No se hallaron diferencias significativas en el resto de los modelos y entre idiomas, aunque DeepSeek y Grok mostraron un ligero descenso en castellano.

Discusión: Las herramientas de inteligencia artificial basadas en LLM demostraron su utilidad en fases iniciales del razonamiento clínico, pero no fue suficientemente precisas para asegurar un diagnóstico final. ChatGPT destacó por su rendimiento estable entre idiomas, lo que lo convierte en una opción preferente en entornos hispanohablantes, especialmente la versión GPT 5.

Palabras clave: inteligencia artificial, diagnóstico clínico, chatbots médicos, ChatGPT, DeepSeek, Grok.

Resume:

Introduction: Natural language-based artificial intelligence (AI) models have shown potential in the medical field, but their diagnostic performance, especially in Spanish, remains underexplored.

Methods: ChatGPT 4o and 5, DeepSeek, and Grok were evaluated using 20 clinical cases from *The New England Journal of Medicine*, translated into Spanish. Six prompts per case (three in Spanish, three in English) assessed differential diagnoses, final diagnosis, and response consistency.

Results: Models achieved an 83% success rate in generating differential diagnoses. Accuracy decreased to 50–60% when asked for a final diagnosis, except for ChatGPT 5 that achieved a 70% in Spanish. No statistically significant differences were found between the rest of models or languages, though DeepSeek and Grok showed slightly lower performance in Spanish.

Discussion: AI proved useful in the early stages of clinical reasoning but was unreliable for definitive diagnoses. ChatGPT stood out for its consistent performance across languages, especially the version GPT 5, making it a preferred option in Spanish-speaking clinical settings.

Keywords: artificial intelligence, clinical diagnosis, medical chatbots, ChatGPT, DeepSeek, Grok.

Introducción:

La inteligencia artificial (IA) ha transformado múltiples aspectos de nuestra vida cotidiana, incluido el ámbito de la salud. Sus primeras aplicaciones en este campo se remontan a la década de 1970. Gracias a los avances en *deep learning*, *natural language processing* y *computer visión* se han desarrollado múltiples herramientas en campos tan diversos como el cribado de retinopatía diabética[1], la identificación de lesiones dermatológicas[2], el diagnóstico asistido en endoscopias[3] y resonancias magnéticas cardíacas. Algunos de estos sistemas han demostrado tal eficacia y eficiencia que han sido aprobados para uso clínico por la FDA (*Food and Drug Administration)*[4].

La relevancia de la inteligencia artificial (IA) basada en los modelos de LLM se ha visto amplificada con la aparición de los *chatbots*, abriendo nuevas posibilidades para su uso.

ChatGPT, Grok y DeepSeek son ejemplos de herramientas de inteligencia artificial conversacional —comúnmente conocidas como chatbots de IA— basadas en modelos LLM que, a diferencia de los chatbots tradicionales, son capaces de comprender el contexto, generar respuestas coherentes y realizar tareas complejas.

De todos ellos el más popular es ChatGPT, lanzado en noviembre de 2022 por OpenAI[5], conocido por su accesibilidad y facilidad de uso. Este modelo ha sido capaz de

superar con altos niveles de acierto pruebas formativas como el examen MIR (Médico Interno Residente), USMLE (*United States Medical Licensing Examination*) o CNMLE (*China National Medical Licensing Examination*)[6]. Sin embargo, su capacidad diagnóstica aún está poco explorada, y los resultados son dispares según la especialidad, con una precisión generalmente baja para su uso autónomo[7-9].

DeepSeek, lanzado en enero de 2025 con código parcialmente abierto, ha ganado popularidad por su buena relación coste-efectividad y su capacidad de razonamiento[10,11]. No obstante, los estudios que comparan su rendimiento médico con otros modelos son escasos y en muchos casos obsoletos, debido a la rápida evolución de estas tecnologías.

Grok, desarrollado por xAI y lanzado en noviembre de 2023, se distingue por su acceso a datos en tiempo real a través de la plataforma X (antes Twitter)[12], utilizada diariamente por millones de usuarios, incluidos organismos oficiales y profesionales sanitarios. Genera respuestas más simples y comprensibles, lo cual podría ser útil para usuarios no especializados que buscan orientación médica básica[13].

Dado que estos modelos están entrenados en varios idiomas, su rendimiento varía según la calidad de los datos disponibles para cada lengua[14.15]. En general, su eficacia es mayor en inglés, lo que plantea dudas sobre su aplicabilidad en otras lenguas como el castellano. La escasez de estudios que evalúen su capacidad diagnóstica en español es precisamente lo que motiva este trabajo.

Partimos de la hipótesis de que los modelos actuales de IA basados en procesamiento del lenguaje natural pueden ofrecer diagnósticos médicos correctos a partir de información clínica.

A partir de esta hipótesis, se plantean los siguientes objetivos específicos: 1) evaluar la capacidad de los modelos de LLM para generar diagnósticos diferenciales a partir de casos clínicos; 2) analizar su capacidad para llegar a un diagnóstico único y preciso; 3) valorar la consistencia con la que los modelos basados en lenguaje natural emiten sus respuestas diagnósticas; 4) comparar la precisión diagnóstica entre los distintos modelos de LLM; 5) examinar las diferencias en el rendimiento diagnóstico de los modelos al utilizar el idioma inglés frente al castellano.

Material y métodos:

Modelos de IA empleados

Se han utilizado cuatro modelos de inteligencia artificial basados en procesamiento del lenguaje natural, desarrollados por tres empresas distintas: ChatGPT 4o, lanzado el 13 de mayo de 2024 por OpenAI; ChatGPT 5, lanzado el 7 de agosto de 2025; DeepSeek-V3, con fecha de lanzamiento el 26 de diciembre de 2024; y Grok-3, con fecha de lanzamiento oficial el 17 de febrero de 2025, desarrollado por xAI. En todas las versiones, se ha empleado la versión accesible de manera gratuita.

Muestra de casos clínicos

Se han seleccionado los 20 casos clínicos más recientes disponibles en la revista The New England Journal of Medicine (NEJM) hasta la fecha de inicio del trabajo[16-35]. Dado que todos los casos están publicados originalmente en inglés, fue necesario realizar una traducción al castellano. En los casos en los que se incluían tablas con resultados analíticos u otros datos complementarios, estos fueron incorporados al cuerpo del texto en formato narrativo.

Prompts

Un *prompt* es el texto de entrada que proporciona el usuario, y que el modelo de utiliza para generar una respuesta. Basándonos en las recomendaciones de OpenAI[36], establecemos los siguientes *prompts* en cada conversación:

PROMPT1	Eres un prestigioso doctor y estás ante un paciente en una consulta médica. Necesito que me des un listado de 5 posibles diagnósticos diferenciales para el paciente con la siguiente historia clínica: [CASO CLÍNICO]. Recuerda que tienes que darme un listado de 5 posibles diagnósticos diferenciales.
PROMPT2	¿Cuál darías como diagnóstico final para el paciente anterior?
PROMPT3	¿Estás seguro de que ese es su diagnóstico final?

Tabla 1. Prompts en castellano

PROMPT4	You are a prestigious doctor, and you are with a patient inside a clinic. I need you to give me a list of 5 possible differential diagnosis for the patient with the following clinic history: [CASO CLÍNICO]. Remember you must provide a list of 5 possible differential diagnosis.
PROMPT5	What would you consider to be the final diagnosis of the previous patient?
PROMPT6	Are you sure that that is the final diagnosis?

Tabla 2. Prompts en inglés

Es importante destacar que en el primer *prompt*, siguiendo pautas de ingeniería de *prompts*, se repite la tarea al modelo para que no la olvide tras un *prompt* tan largo, obteniendo respuestas más precisas, relevantes y útiles.

Cada grupo de *prompts* correspondiente a cada idioma (PROMPT1–3 en castellano y PROMPT4–6 en inglés) se introdujo de forma encadenada en una única conversación. Este enfoque busca simular un contexto conversacional realista, manteniendo la coherencia en el desarrollo del caso clínico.

Sesgo de memoria

Los cuatro modelos emiten respuestas en función de su preentrenamiento, de dónde extrae los datos de fuentes de información y de las interacciones de los usuarios con el chat (37). Para evitar este sesgo, en los 4 modelos se han desactivado las funciones correspondientes que garantizan la no retención de información entre sesiones. En caso de ChatGPT el chat temporal; en DeepSeek se desactivó la función de mejora de modelo; y en Grok se activó el chat privado.

	N	ChatGPT 4o				ChatGPT 5				DeepSeek				Grok				p-valor
		Bien	%	Mal	%	Bien	%	Mal	%	Bien	%	Mal	%	Bien	%	Mal	%	
PROMPT1	20	17	**85**	3	15	18	**90**	2	10	17	**85**	3	15	16	**80**	4	20	0.3916
PROMPT2	20	11	**55**	9	45	14	**70**	6	30	10	**50**	10	50	10	**50**	10	50	0,1048
PROMPT3	20	10	**50**	10	50	15	**75**	5	25	8	**40**	12	60	10	**50**	10	50	0,0415
PROMPT4	20	15	**75**	5	25	17	**85**	3	15	17	**85**	3	15	18	**90**	2	10	0,1589
PROMPT5	20	11	**55**	9	45	14	**70**	6	30	13	**65**	7	35	12	**60**	8	40	0,1718
PROMPT6	20	11	**55**	9	45	9	**45**	11	55	13	**65**	7	35	12	**60**	8	40	0,4173

Tabla 2. Aciertos y fallos de cada modelo en cada prompt. N = número de respuestas

Análisis estadístico

Para comparar el rendimiento de los cuatro modelos de IA en cada *prompt*, se aplicó la prueba Q de Cochran. Para comparar el rendimiento en los dos idiomas (castellano e inglés) dentro de cada *prompt* y modelo, se aplicó la prueba de McNemar con corrección de continuidad de Yates. Se consideraron estadísticamente significativos los valores con $p<0,05$. Los datos han sido procesados en la plataforma de OpenEpi, y la prueba Q de Cochran se realizó a través de Google Colab.

Resultados:

Con 20 casos seleccionados, 6 *prompts* introducidos por cada caso y en los 4 modelos escogidos para la comparativa, obtuvimos un total de 480 respuestas.

En los cuatro modelos evaluados, se observan tasas de acierto que oscilan entre el 75% y el 90% al solicitar un listado con posibles diagnósticos diferenciales. No obstante, cuando se les solicita que emitan un diagnóstico definitivo final, la tasa de aciertos se reduce a aproximadamente un 50%, excepto en ChatGPT 5 que logra una tasa de 70%. Al analizar la consistencia en sus respuestas, se observa que el rendimiento permanece en torno al mismo nivel que en el diagnóstico definitivo, salvo en ChatGPT 5 en inglés. Comparando los modelos entre sí, no se observan diferencias significativas en todos los *prompts* menos en el tercero, donde se observa una tasa de aciertos de ChatGPT 5 muy superior al resto de versiones comparadas.

En la Tabla 4, se recogen los datos obtenidos agrupados por idiomas. Los *prompts* 1-3 corresponden a los que están en castellano, y los *prompts* 4-6 al inglés.

En el caso de ChatGPT 4o, se observa una tasa de aciertos casi idénticas entre ambos idiomas.

En ChatGPT 5 la tasa es prácticamente idéntica entre idiomas, excepto en el *prompt 3* donde se observa una diferencia estadísticamente significativa, siendo superior la tasa de aciertos en castellano.

Respecto a DeepSeek, se observa que ambos idiomas presentan el mismo número de aciertos en la fase de diagnóstico diferencial. Sin embargo, al solicitar un diagnóstico final, el rendimiento disminuyó un 10% en mayor medida en castellano. A pesar de estas diferencias observadas, no se alcanzó significación estadística

		N	Castellano				Inglés			
			Bien	%	Mal	%	Bien	%	Mal	%
ChatGPT 4o	PROMPT1	20	17	**85**	3	15	15	**75**	5	25
	PROMPT2	20	11	**55**	9	45	11	**55**	9	45
	PROMPT3	20	10	**50**	10	50	11	**55**	9	45
ChatGPT 5	PROMPT1	20	18	**90**	2	10	17	**85**	3	15
	PROMPT2	20	14	**70**	6	30	14	**70**	6	30
	PROMPT3	20	15	**75**	5	25	9	**45**	11	55
DeepSeek	PROMPT1	20	17	**85**	3	15	17	**85**	3	15
	PROMPT2	20	10	**50**	10	50	13	**65**	7	35
	PROMPT3	20	8	**40**	12	60	13	**65**	7	35
Grok	PROMPT1	20	16	**80**	4	20	18	**90**	2	10
	PROMPT2	20	10	**50**	10	50	12	**60**	8	40
	PROMPT3	20	10	**50**	10	50	12	**60**	8	40

Tabla 4. Resultados agrupados por idiomas. N = número de respuestas

Por último, los resultados de Grok siguen un patrón similar al de DeepSeek. Aunque el rendimiento fue superior en inglés, las diferencias entre idiomas no fueron estadísticamente significativas.

Discusión:

Herramienta para la actividad clínica

Según los datos obtenidos, la IA puede ser una buena herramienta en las consultas, especialmente en la fase inicial de generación de diagnósticos diferenciales, donde alcanzan una elevada precisión que oscila entre el 80% y 90%. Sin embargo, a la hora de realizar un diagnóstico definitivo, su rendimiento disminuye considerablemente. Este nivel de precisión es insuficiente para su uso autónomo en la toma de decisiones clínicas. Además, en varias ocasiones, los propios modelos manifestaron incertidumbre sobre sus respuestas, lo que refuerza la necesidad de supervisión humana en su uso clínico.

Es importante destacar que la mayoría de los casos clínicos utilizados en el estudio presentan un alto grado de complejidad y, en muchos de ellos, el diagnóstico definitivo depende de pruebas complementarias o datos no

incluidos en los *prompts* introducidos. Por tanto, estos resultados no reflejan necesariamente el rendimiento que podrían alcanzar los modelos en casos de patología común o en contextos con información clínica más completa.

Elección del modelo de LLM y del idioma de uso

Si nos basamos en los resultados, de manera global no hay diferencias estadísticamente significativas entre los 4 modelos evaluados ni en el idioma empleado, salvo en uno de los *prompts*, siendo favorable para ChatGPT 5.

ChatGPT 4o tiene prácticamente la misma tasa de respuestas acertadas en inglés y castellano en todos los *prompts* introducidos, lo que sugiere que su rendimiento es consistente tanto en castellano como en inglés. Sin embargo, los resultados de ChatGPT 5 superan en todos los aspectos de los de su predecesora, difiriendo únicamente en la consistencia de las respuestas en ambos idiomas, siendo superior en castellano. En DeepSeek como en Grok, se observó un mejor rendimiento en inglés. En el caso de DeepSeek, la precisión en el diagnóstico final fue del 65% en inglés, frente al 50% en castellano. En Grok, esta diferencia fue similar. Estos resultados pueden atribuirse al hecho de que estos modelos están entrenados predominantemente con datos en inglés, lo cual sigue siendo una limitación común en muchos sistemas de IA actuales.

Por lo tanto, en caso de querer usar ChatGPT, se recomienda emplear la última versión en castellano. Sin embargo, si se plantea utilizar Grok o DeepSeek, es recomendable utilizarlo en inglés, aunque el rendimiento en castellano sea razonablemente bueno. Esta diferencia idiomática podría tener implicaciones en el contexto clínico español. Hasta un 77% de la población española no hablan nada de inglés (38), y un 84% de las facultades de Medicina en España no priorizan enseñar inglés (39).

Aspectos legales

El uso clínico de modelos de LLM plantea importantes interrogantes en cuanto a la privacidad y protección de datos personales. Hoy en día, no existe un marco legal plenamente definido que regule de forma específica el uso de la IA en medicina, ni tampoco una normativa uniforme sobre el almacenamiento, tratamiento y uso compartido de datos clínicos (10). Uno de los principales problemas reside en la necesidad de introducir la mayor cantidad posible de información clínica en los *prompts* para que el modelo proporcione una respuesta precisa. Aunque no se incluyan datos directamente identificativos, siempre existe un riesgo potencial de reidentificación o exposición de información sensible, especialmente si el tratamiento de los datos no se realiza en un entorno controlado.

Limitaciones y posibles líneas de investigación

Una de las principales limitaciones de los modelos de inteligencia artificial basados en LLM es la naturaleza del propio entrenamiento, el cual se fundamenta en datos previamente existentes y validados externamente. Estos modelos reformulan patrones previos, y el rendimiento puede verse comprometido en situaciones en las que no ha habido entrenamiento previo.

Un ejemplo claro de este sesgo se encuentra en el estudio de Kamulegeya et al, donde a la hora de clasificar lesiones cutáneas en personas no blancas la precisión diagnóstica era muy inferior que los pacientes de piel blanca (40). De forma análoga, en nuestro estudio al basarse en casos clínicos ya publicados, los modelos podrían haber estado expuestos previamente a dicha información, lo que podría influir positivamente en sus respuestas y generar una sobreestimación de su rendimiento real. De esta manera, sería interesante estudiar directamente la fiabilidad diagnóstica en la clínica, en casos nuevos y no documentados previamente, para observar realmente su utilidad.

Otra limitación importante es que en los casos clínicos que incluían pruebas de imagen, se introdujeron las interpretaciones realizadas por médicos en lugar de usar las imágenes originales. Esto introduce un sesgo añadido, al depender de un intermediario humano en la transmisión de la información.

Por último, cabe destacar que el campo de la inteligencia artificial está en constante evolución. Con la aparición frecuente de nuevas versiones y modelos, los resultados obtenidos en este estudio podrían quedar obsoletos en un corto período de tiempo. Por tanto, es fundamental continuar evaluando de forma periódica la precisión y fiabilidad diagnóstica de las nuevas generaciones de modelos, así como explorar propuestas emergentes de distintas empresas desarrolladoras. En nuestro estudio, esta evolución se ve claramente entre las versiones 4.o y 5 de ChatGPT. Recientemente, con un planteamiento similar, la plataforma MAI-DxO de Microsoft, tomando como referencia 300 casos clínicos publicados en el New England Journal of Medicine obtuvo una un porcentaje de aciertos del 85 % mientras que los médicos solo llegaron al 20 % (41).

Conclusión

La IA presenta un gran potencial como apoyo en la práctica clínica, con un valor especifico en los procesos diagnósticos de los modelos basados en LLM. Actualmente, en la práctica clínica, en España, sería preferible emplear ChatGPT 5, siempre como una herramienta complementaria y bajo supervisión médica. Los datos presentados en este artículo son los disponibles en este momento, aunque seguro que estos resultados mejorarán de forma muy relevante en los próximos años.

Bibliografía:

1.Gargeya R, Leng T. Automated Identification of Diabetic Retinopathy Using Deep Learning. Ophthalmology 2017 -07-01;124(7):962–969.

2.Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist–level classification of skin cancer with deep neural networks. Nature 2017 February 2;542(7639):115–118.

3.Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. Lancet Gastroenterol Hepatol 2020 -04;5(4):352–361.

4.Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointestinal Endoscopy 2020 October 1;92(4):807–812.

5.Marr B. A Short History Of ChatGPT: How We Got To Where We Are Today. Accessed Apr 19, 2025.

6.Meléndez D, Izquierdo JL. Revista SOCAMPAR ORIGINAL CAPACIDAD DE CHATGPT EN LA RESOLUCCIÓN CORRECTA DE LAS PREGUNTAS DE NEUMOLOGÍA DEL EXAMEN MIR CHATGPT'S ABILITY IN THE CORRECT RESOLUTION OF THE PNEUMOLOGY QUESTIONS OF THE MIR EXAM.

7.Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int J Environ Res Public Health 2023 February 15;20(4):3378.

8.Lin JC, Younessi DN, Kurapati SS, Tang OY, Scott IU. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. Eye (Lond) 2023 -12;37(17):3694–3695.

9.Shemer A, Cohen M, Altarescu A, Atar-Vardi M, Hecht I, Dubinsky-Pertzov B, et al. Diagnostic capabilities of ChatGPT in ophthalmology. Graefes Arch Clin Exp Ophthalmol 2024 -07;262(7):2345–2352.

10.Peng Y, Malin BA, Rousseau JF, Wang Y, Xu Z, Xu X, et al. From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare. Journal of Biomedical Informatics 2025 -03-01;163:104791.

11.Chen Y, Shen J, Ma D. DeepSeek's impact on thoracic surgeons' work patterns—past, present and future. J Thorac Dis 2025 -2-28;17(2):1114–1117.

12.Gupta R, Park JB, Ragsdale LB, Meggers K, Eimani A, Mailey BA. The Intersection of AI Grok With Aesthetic Plastic Surgery. Aesthetic Surgery Journal 2024 -06-01;44(6):NP437–NP440.

13.Şahin MF, Topkaç EC, Doğan Ç, Şeramet S, Özcan R, Akgül M, et al. Still Using Only ChatGPT? The Comparison of Five Different Artificial Intelligence Chatbots' Answers to the Most Common Questions About Kidney Stones. Journal of Endourology 2024;38(11):1172–1177.

14.Wu J. The rise of DeepSeek: technology calls for the "catfish effect". J Thorac Dis 2025 -2-28;17(2):1106–1108.

15.Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: Multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-Redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc 2025 -05;33(5):1553–1556.

16.Corey KE, Dudzinski DM, Guimaraes AR, Mino-Kenudson M. Case 9-2025: A 59-Year-Old Man with Hepatocellular Carcinoma. N Engl J Med 2025 -03-27;392(12):1216.

17.Arrillaga-Romany I, Ford JN, Dunn GP, Kotton CN, Mount CW, Latham KA. Case 33-2024: A 71-Year-Old Woman with Confusion, Aphasia, and a Brain Mass. N Engl J Med 2024 -10-24;391(16):1529.

18.Leaf RK, Messick BH, Meador CB, Loneman D. Case 7-2025: A 65-Year-Old Woman with Weakness, Back Pain, and Pancytopenia. N Engl J Med 2025 -02-27;392(9):903.

19.Yancy CW, Guseh JS, Ghoshhajra BG, Falk RH, Yee AJ, Hutchison BM. Case 3-2025: A 54-Year-Old Man with Exertional Dyspnea and Chest Pain. N Engl J Med 2025 -01-23;392(4):383.

20.Simmons LH, Willett R, Haydu JE, Fitzpatrick MJ. Case 11-2025: A 79-Year-Old Woman with Cough and Weight Loss. N Engl J Med 2025 -04-17;392(15):1532.

21.Bromberg GK, Rasmussen RG, Sherman KE, Kalva SP, Goodarzi K, Glickman JN. Case 6-2025: A 62-Year-Old Man with Abdominal Pain. N Engl J Med 2025 -02-20;392(8):807.

22.Kinane TB, Zucker EJ, Sparger KA, Kelleher CM, Shih AR. Case 35-2024: A Newborn with Hypoxemia and a Lung Opacity. N Engl J Med 2024 -11-14;391(19):1838.

23.Simic P, Dudzinski DM, Masuodi B, Colling C, Liu L. Case 8-2025: A 72-Year-Old Woman with Altered Mental Status and Acidemia. N Engl J Med 2025 -03-13;392(11):1121.

24.Benjamin S, Basovic L, Romero JM, Lam AD, Adams C. Case 37-2024: A 41-Year-Old Man with Seizures and Agitation. N Engl J Med 2024 -11-28;391(21):2036.

25.Tangren JS, Jeyabalan A, Klepeis VE. Case 1-2025: A 35-Year-Old Woman with Shortness of Breath and Edema in the Legs. N Engl J Med 2025 -01-09;392(2):186.

26.Casey A, Madhavan VL, Zucker EJ, Farmer JR. Case 39-2024: A 30-Month-Old Boy with Recurrent Fever. N Engl J Med 2024 -12-12;391(23):2256.

27.Restrepo D, Sultana S, Divakaran S, Sparks JA. Case 4-2025: A 41-Year-Old Man with Syncope, Ankle Swelling, and Abnormal Chest Imaging. N Engl J Med 2025 -01-30;392(5):495.

28.Mylonakis E, Zhang EW, Bertrand PB, Gurol ME, Triant VA, Chaudet KM. Case 38-2024: A 22-Year-Old Woman with Headache, Fever, and Respiratory Failure. N Engl J Med 2024 -12-05;391(22):2148.

29.Heaton K, Zern EK, Spahillari A, Barrett CD. Case 34-2024: A 69-Year-Old Man with Dyspnea after Old Myocardial Infarction. N Engl J Med 2024 -10-31;391(17):1633.

30.Shappell EF, Applewhite BP, Azar SS, Lin DJ. Case 2-2025: A 21-Year-Old Man with Loss of Consciousness and a Fall. N Engl J Med 2025 -01-16;392(3):268.

31.Zunt J, Barczak AK, Chang DY. Case 5-2025: A 30-Year-Old Woman with Headache and Dysesthesia. N Engl J Med 2025 -02-13;392(7):699.

32.Sherman SV, Marinacci LX, Rincon SP, Raynor EM. Case 32-2024: A 72-Year-Old Woman with Dyspnea, Dysphagia, and Dysarthria. N Engl J Med 2024 -10-17;391(15):1441.

33.Zella GC, Pourvaziri A, Greenberg EL, Leonard MM. Case 36-2024: A 16-Year-Old Girl with Abdominal Pain. N Engl J Med 2024 -11-21;391(20):1937.

34.Jeyabalan A, Czawlytko CL, Beck LH, Trivin-Avillach C. Case 10-2025: A 32-Year-Old Woman with Flank Pain, Fever, and Hypoxemia. N Engl J Med 2025 -04-10;392(14):1428.

35.Herzig SJ, Kozak BM, Kotton CN, Fogerty AE, Turbett SE. Case 40-2024: A 56-Year-Old Woman with End-Stage Liver Disease and Headache. N Engl J Med 2024 -12-19;391(24):2361.

36.GPT-4.1 Prompting Guide | OpenAI Cookbook. Accessed Apr 20, 2025.

37.What is Memory? | ChatGPT. Accessed 21/04, 2025.

38.El inglés, ¿qué tan bien lo hablan los españoles? - educa-web.com. Accessed Jun 16, 2025.

39.El 84% de las facultades de Medicina no priorizan enseñar inglés2024. Accessed Jun 16, 2025.

40.Kamulegeya L, Bwanika J, Okello M, Rusoke D, Nassiwa F, Lubega W, et al. Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. Afr Health Sci 2023 -06;23(2):753–763.

41. Nori H, Daswani M, Kelly CH, et al. Sequential Diagnosis with Language Models. Julio 2025. https://arxiv.org/abs/2506.22405v2.