

Chapter 18 / Capítulo 18

Critical reading of scientific articles on
health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch18
Pages: 311-332



© 2026 The authors. This is an open access
book chapter distributed under the terms of
the Creative Commons Attribution (CC BY)
4.0 License.

Use of Artificial Intelligence in the Critical Appraisal of Scientific Medical Articles

Uso de la inteligencia artificial en la lectura crítica de artículos científicos médicos

Jhossmar Cristians Auza-Santivañez  

Boris Adolfo Llanos Torrico  

Rita Marlene Mamani-Limachi  

Henrry Temis Quisbert Vasquez  

1 Hospital del Gran Chaco "Fray Quebracho" de III nivel. Instituto Académico Científico Quispe-Cornejo. Tarija, Bolivia.

2 Universidad Mayor de San Andrés. Universidad Autónoma "Tomás Frías". Oruro, Bolivia.

3 Hospital Obrero N°1 - Caja Nacional de Salud, La Paz, Bolivia.

4 Hospital Materno Infantil "Germán Urquidí". Cochabamba, Bolivia.

ABSTRACT

Introduction: The incorporation of artificial intelligence (AI), especially large-scale language models (LLMs), has transformed the critical appraisal of scientific literature by facilitating the processing of large volumes of information. However, its use in Evidence-Based Medicine requires recognizing both its strengths and methodological limitations to avoid errors in interpretation and cognitive dependence. **Development:** This chapter describes a hybrid model of critical appraisal in which AI acts as a support tool for structured data extraction, statistical interpretation, and preliminary bias detection, always subordinate to clinical judgment. It proposes principles for responsible use, analysis algorithms, prompt design, and verification strategies to minimize hallucinations, biases, and methodological errors. **Conclusions:** AI represents an opportunity to optimize critical appraisal and democratize access to scientific evidence; however, its true value depends on its ethical, reflective, and methodologically rigorous integration. The preservation of critical thinking, the systematic verification of information, and the maintenance of independent clinical judgment are essential conditions for these tools to strengthen, rather than replace, evidence-based decision-making.

Keywords: Artificial Intelligence; Evidence-Based Medicine; Critical Appraisal; Language Models; Clinical Research.; RESUMEN

RESUMEN

Introducción: La incorporación de la inteligencia artificial (IA), especialmente los modelos de lenguaje de gran escala (LLMs), ha transformado la lectura crítica de la literatura científica al facilitar el procesamiento de grandes volúmenes de información. No obstante, su utilización en Medicina Basada en la Evidencia exige reconocer tanto sus fortalezas como sus limitaciones metodológicas para evitar errores de interpretación y dependencia cognitiva. **Desarrollo:** El capítulo describe un modelo híbrido de lectura crítica en el que la IA actúa como herramienta de apoyo para la extracción estructurada de datos, interpretación estadística y detección preliminar de sesgos, siempre subordinada al juicio clínico. De esta manera, propone principios de uso responsable, algoritmos de análisis, diseño de prompts y estrategias de verificación para minimizar alucinaciones, sesgos y errores metodológicos. **Conclusiones.** La IA representa una oportunidad para optimizar la lectura crítica y democratizar el acceso a la evidencia científica; sin embargo, su verdadero valor depende de una integración ética, reflexiva y metodológicamente rigurosa. La preservación del pensamiento crítico, la verificación sistemática de la información y el mantenimiento del juicio clínico independiente constituyen condiciones indispensables para que estas herramientas fortalezcan, y no sustituyan, la toma de decisiones basada en evidencia.

Palabras clave: Inteligencia Artificial; Medicina Basada en la Evidencia; Lectura Crítica; Modelos de Lenguaje; Investigación Clínica

1. CASO CLÍNICO

La residente de segundo año de Medicina Interna, en un hospital universitario, recibe durante una guardia nocturna a un paciente masculino de 67 años con insuficiencia cardíaca con fracción de eyección reducida (ICFER), fibrilación auricular con respuesta ventricular conserva y enfermedad renal crónica. Previo al pase de visita, el médico tratante le solicita revisar la evidencia científica más reciente sobre la eficacia y seguridad de la combinación sacubitrilo/valsartán en pacientes con este perfil de comorbilidades. Debido a las limitaciones de tiempo propias del contexto asistencial, la residente identifica un ensayo clínico aleatorizado publicado el año previo en una revista biomédica indexada. Ante la extensión del manuscrito y la necesidad de obtener una evaluación rápida de su calidad metodológica, decide utilizar un modelo de lenguaje de gran escala (Large Language Model, LLM). Para ello, introduce el resumen estructurado (abstract) del estudio y solicita un análisis crítico de su metodología y principales resultados.

La respuesta generada por la inteligencia artificial resulta coherente, estructurada y aparentemente consistente. Sin embargo, al contrastarla con el artículo original, la residente identifica discrepancias relevantes. El modelo informa un número necesario a tratar (NNT) de 12, mientras que el estudio reporta un NNT de 18. De esta manera, omite señalar que el 34 % de los participantes presentaban una tasa de filtrado glomerular (TFG) superior a 45 mL/min/1,73 m², característica que limita la aplicabilidad de los hallazgos al paciente evaluado, cuya TFG es de 38 mL/min/1,73 m².

Este escenario plantea la pregunta central que orienta el presente capítulo: ¿cómo integrar la inteligencia artificial como herramienta de apoyo en la lectura crítica de artículos científicos sin delegar el juicio clínico y metodológico al algoritmo?

2. OBJETIVOS

- Reconocer las capacidades y limitaciones actuales de los modelos de lenguaje de gran escala (LLMs) en el análisis de artículos científicos médicos.
- Aplicar un algoritmo estructurado de lectura crítica asistida por inteligencia artificial.
- Formular prompts clínicamente pertinentes y metodológicamente precisos para el análisis de literatura biomédica.
- Identificar errores típicos de la IA en la interpretación de estadística clínica, incluyendo OR, RR, HR, sensibilidad, especificidad y AUC.
- Distinguir entre el uso responsable de la IA como herramienta de apoyo y el riesgo de dependencia cognitiva.
- Integrar la IA en el flujo de lectura crítica sin comprometer el razonamiento clínico-metodológico independiente.

3. INTRODUCCIÓN

3.1 EL DESAFÍO DE LA INFORMACIÓN BIOMÉDICA CONTEMPORÁNEA

La producción científica biomédica ha experimentado un crecimiento exponencial durante las últimas dos décadas. Se estima que se publican más de 2.5 millones de artículos en revistas médicas indexadas cada año, un volumen que ningún profesional puede absorber de manera integral mediante la lectura convencional. Este fenómeno, conocido como “sobrecarga de información” (information overload), se ha convertido en un obstáculo estructural para la práctica de la medicina basada en la evidencia (MBE).

La sobrecarga de información no solo afecta la práctica clínica, sino que también está transformando la estructura misma del conocimiento médico. En la actualidad, se describe una “paradoja del conocimiento” en los estudiantes de medicina: a pesar de tener un acceso sin precedentes a volúmenes masivos de información, su comprensión profunda de la fisiopatología el “por qué” de las enfermedades es cada vez más superficial.

Este fenómeno se atribuye a un sistema educativo que prioriza la memorización de datos y el rendimiento en exámenes estandarizados sobre la integración conceptual y el razonamiento crítico.

Esta paradoja adquiere una relevancia capital: la misma herramienta que puede democratizar el acceso a la evidencia también puede, si se usa incorrectamente, amplificar este déficit de comprensión, convirtiendo a los clínicos en meros validadores de resúmenes sin capacidad para

evaluar la plausibilidad fisiopatológica de los hallazgos que una IA les presenta. La lucha contra esta tendencia es el núcleo de la lectura crítica asistida que aquí se propone.

En paralelo, la lectura crítica de artículos científicos requiere competencias metodológicas específicas que no siempre se desarrollan de manera uniforme durante la formación médica. La interpretación de medidas de efecto (odds ratio OR, riesgo relativo RR, hazard ratio HR), el reconocimiento de sesgos, la evaluación de la validez interna y la valoración de la aplicabilidad clínica constituyen habilidades que demandan entrenamiento sistemático y práctica continua.

Aunque la eficiencia es innegable, la evidencia de 2025 subraya que la concordancia en la extracción automatizada de datos clínicos es de aproximadamente el 77.2%. Esto implica que 1 de cada 4 puntos de datos críticos extraídos por un LLM podría ser inexacto, lo que exige que el juicio clínico del especialista sea el filtro final e innegociable. Si bien los datos históricos apuntaban a un margen de error variable en la extracción, los ensayos metodológicos controlados de no inferioridad (Within Reviews) publicados en 2026 demuestran que las plataformas optimizadas con LLMs de última generación alcanzan una precisión global que no es inferior a la extracción puramente humana. Aunque cometen tipos de errores similares, su verdadero valor radica en que reducen del 50% al 75% el tiempo manual de cribado y extracción de datos, optimizando drásticamente los costos operativos en la fase inicial de la lectura crítica.

3.2 LA IRRUPCIÓN DE LA INTELIGENCIA ARTIFICIAL EN LA PRÁCTICA CLÍNICA Y ACADÉMICA

La aparición de modelos de lenguaje de gran escala (Large Language Models, LLMs) como GPT, Claude y Gemini ha transformado radicalmente la manera en que los profesionales de la salud acceden y procesan información. Estos sistemas, entrenados sobre corpus masivos de texto biomédico, demostraron desde sus versiones iniciales una capacidad notable para resumir artículos, explicar conceptos estadísticos y responder preguntas clínicas en lenguaje natural.

Sin embargo, su adopción masiva en el ámbito médico ha ocurrido con mayor velocidad que la producción de evidencia sobre su seguridad, precisión y limitaciones en este contexto específico. Estudios recientes documentan que los LLMs pueden generar respuestas plausibles, pero factualmente incorrectas (fenómeno denominado “alucinación”), omitir información crítica, o sesgar sus respuestas en función de los patrones estadísticos aprendidos durante el entrenamiento, sin que el usuario pueda detectar fácilmente estos errores.

3.3 MODELO DE LECTURA CRÍTICA HÍBRIDA ASISTIDA POR INTELIGENCIA ARTIFICIAL

En el contexto actual, caracterizado por la disponibilidad de modelos de lenguaje de gran escala, se propone un modelo conceptual de lectura crítica híbrida asistida por inteligencia artificial, en el cual el análisis de la evidencia científica emerge de la interacción dinámica entre el juicio clínico humano y herramientas de apoyo automatizadas.

En este modelo, la inteligencia artificial no constituye un agente epistémico autónomo, sino un sistema de apoyo que opera sobre tareas específicas como la extracción estructurada de información, la explicación de conceptos estadísticos o la identificación preliminar de sesgos, mientras que las decisiones de interpretación, aplicabilidad y toma de decisiones clínicas permanecen bajo control humano.

Este enfoque reconoce simultáneamente el potencial de la IA para mejorar la eficiencia cognitiva y el riesgo inherente de dependencia o sobreconfianza, estableciendo como eje central la necesidad de verificación sistemática y juicio crítico independiente.

4. FUNDAMENTOS DEL USO DE LA IA EN LA LECTURA CRÍTICA

4.1 ¿QUÉ ES UN MODELO DE LENGUAJE DE GRAN ESCALA O LLM (LANGUAGE LARGE MODEL)?

Un LLM es un sistema de inteligencia artificial basado en arquitecturas de redes neuronales profundas (transformers) entrenado sobre grandes corpus de texto.

A diferencia de los sistemas de recuperación de información clásicos, los LLMs no indexan documentos ni recuperan fragmentos predefinidos: generan texto nuevo, palabra a palabra, en función de patrones estadísticos aprendidos durante el entrenamiento. Esta característica es

tanto su principal fortaleza como su limitación más crítica: el sistema puede producir texto fluido y aparentemente coherente sobre cualquier tema, incluso cuando carece de la información precisa para responder correctamente. No entiende el contexto de la información que posee, al contrario, es una predicción matemática y estadística de la próxima palabra con mayor probabilidad a la secuencia de texto establecida.

Figura 1: Como funciona un Language Large Model.



Fuente. Elaboración propia

4.2 FUNDAMENTOS DE LA LECTURA CRÍTICA

La lectura crítica de artículos científicos es un proceso sistemático que evalúa la validez, los resultados y la aplicabilidad de un estudio. Las dimensiones fundamentales incluyen:

- **Validez interna:** ¿el diseño y la ejecución del estudio permiten atribuir los resultados a la intervención estudiada y no a factores confundentes?
- **Validez externa (aplicabilidad):** ¿los hallazgos son generalizables a la población de pacientes del lector?
- **Importancia clínica:** ¿la magnitud del efecto es relevante para la práctica, independientemente de la significancia estadística?
- **Precisión estadística:** ¿los intervalos de confianza son suficientemente estrechos para guiar decisiones?

El uso de la IA en este proceso debe estar necesariamente subordinado a este marco conceptual, y no sustituirlo.

4.3 PRINCIPIOS PARA LA INTEGRACIÓN DE IA EN LA LECTURA CRÍTICA

Se plantean entonces cuatro principios que en adelante guiarán el uso responsable de la IA en este contexto:

- **Subordinación:** La IA está subordinada al juicio clínico; asiste, no decide. El juicio clínico y metodológico final es exclusivamente humano.
- **Verificación:** toda afirmación de la IA sobre datos, cifras o conclusiones debe cotejarse con el texto original del artículo.
- **Transparencia:** el usuario debe conocer las capacidades y limitaciones del sistema que utiliza, incluyendo su fecha de corte de entrenamiento.
- **Proporcionalidad:** la profundidad del uso de la IA debe ser proporcional a la complejidad del artículo y la trascendencia clínica de la decisión.

Figura 2. Principios para integración de IA en la lectura crítica
PRINCIPIOS PARA LA INTEGRACIÓN RESPONSABLE DE IA
EN LA LECTURA CRÍTICA DE ARTÍCULOS CIENTÍFICOS



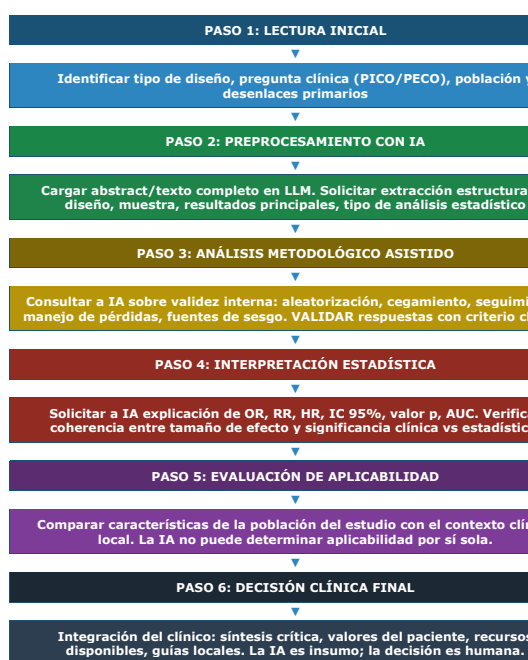
Fuente: elaboración propia con apoyo de Inteligencia Artificial

En el entorno editorial contemporáneo, ya no es suficiente declarar el uso genérico de algoritmos. Las publicaciones científicas exigen la adopción de listas de verificación estandarizadas como PRISMA-trAlce (extensión de PRISMA para el rastro de IA) y el estricto cumplimiento del Marco RAISE, obligando al investigador a documentar de forma transparente cuándo una herramienta ha propuesto o emitido juicios metodológicos. De esta manera, bajo las directrices de transparencia institucional (como las normativas HTI-1), los entornos académicos están transitando de LLMs comerciales en la nube hacia modelos de código abierto locales (On-Premise), garantizando que el procesamiento de la literatura no exponga indirectamente patrones asistenciales ni vulnere la gobernanza de datos del centro de salud.

4.4 ALGORITMO DE LECTURA CRÍTICA ASISTIDA POR IA

La Figura 3 presenta un algoritmo estructurado en seis pasos para integrar el uso de la IA en el proceso de lectura crítica, desde la lectura inicial hasta la toma de decisión clínica.

Figura 3. Algoritmo de lectura crítica asistida por inteligencia artificial.



Fuente: Elaboración propia

5. APORTES DE LA IA EN EL ANÁLISIS DE ARTÍCULOS CIENTÍFICOS

Dado el inmenso potencial del uso de las herramientas de IA se plantean a continuación como puede contribuir la IA en el análisis de artículos científicos.

5.1 PROCESAMIENTO Y EXTRACCIÓN DE INFORMACIÓN ESTRUCTURADA

Una de las contribuciones más documentadas de los LLMs en la lectura de literatura científica es su capacidad para extraer información estructurada de textos no estructurados. Al proporcionarle el *abstract* o el texto completo de un artículo, el modelo puede identificar con razonable precisión el diseño del estudio, la población incluida, la intervención evaluada, los comparadores, los desenlaces principales y secundarios, y los resultados numéricos clave. Esta tarea, que un lector humano ejecuta en varios minutos de atención sostenida, puede completarse en segundos con el apoyo de la IA.

Sin embargo, la precisión de esta extracción no es absoluta. Los estudios de evaluación de los LLMs en tareas de extracción de datos de ensayos clínicos y revisiones sistemáticas reportan tasas de error que oscilan entre el 15 % y el 30 %, especialmente cuando se trata de valores numéricos precisos, datos de subgrupos y resultados secundarios. Un estudio dentro de revisiones (*Study Within Reviews*, SWAR) que comparó un flujo asistido por IA frente a la extracción exclusivamente humana en 9341 elementos de datos procedentes de 63 estudios encontró una concordancia global entre ambos métodos del 77,2 % (IC 95 %: 76,3 %-78,0 %) (1). Aunque la precisión del método asistido por IA fue alta en términos globales (91,0 %; IC 95 %: 90,4 %-91,6 %), los errores residuales no se distribuyeron de manera uniforme: se concentraron en las variables más complejas y en los resultados secundarios, mientras que los datos estructurales simples (como el número de participantes o el país de realización) se extrajeron con mayor fiabilidad.

Un hallazgo particularmente preocupante para la lectura crítica es la marcada heterogeneidad en el rendimiento de los LLM en función del dominio de datos. Una revisión sistemática rápida que analizó 31 estudios de extracción de datos clínicos mediante LLMs encontró que la precisión comunicada oscilaba entre el 17 % y el 100 % para las características del diseño del estudio, entre el 9,1 % y el 100 % para las características de los pacientes, y entre el 30 % y el 100 % para los desenlaces del estudio.

Esta amplitud refleja no solo diferencias reales en la dificultad de cada dominio, sino también una considerable heterogeneidad metodológica entre los estudios primarios, que emplearon distintos modelos, arquitecturas de *prompt*, formatos de entrada y definiciones de error.

La extracción de información estadística compleja como los resultados de los análisis de sensibilidad, las medidas de heterogeneidad (I^2 , τ^2) en los metaanálisis o las tablas de riesgo de sesgo muestra tasas de error considerablemente mayores que la extracción de datos descriptivos simples.

En el estudio de Martin et al., que evaluó repetidamente ChatGPT-4.0 en la extracción de datos de un mismo conjunto de publicaciones a lo largo de varias semanas, los errores globales alcanzaron el 32,4 % en los informes completos de ensayos clínicos, y la recuperación de elementos complejos como la cita bibliográfica completa resultó correcta en solo el 0 % al 33,3 % de las ejecuciones (en sus últimas versiones puede haber menos riesgo de alucinación). De particular relevancia clínica fue el hallazgo de que ningún elemento de datos pudo ser extraído de forma consistentemente correcta en todos los tipos de publicación y en todas las repeticiones del experimento. Esta variabilidad temporal la misma pregunta, el mismo artículo y el mismo *prompt*, pero resultados diferentes en momentos distintos constituye un desafío específico de los modelos probabilísticos que el lector crítico debe conocer: a diferencia de un sistema determinista, el LLM no garantiza reproducibilidad en la extracción.

Para el estudiante y profesional de la salud que utiliza la IA en la lectura crítica, estos datos tienen una doble implicación práctica: en primer lugar, la extracción automatizada puede ahorrar tiempo y proporcionar una primera aproximación útil a la estructura del artículo; en segundo lugar, y de manera irrenunciable, todo valor numérico, toda medida de efecto y todo resultado de un análisis de sensibilidad o de heterogeneidad generado por el LLM debe ser verificado directamente en el texto fuente original. La IA acelera la lectura, pero no garantiza la fidelidad de lo leído.

Tabla 1. Matriz de Fiabilidad en la Extracción de Datos por IA

Categoría de Datos	Precisión Estimada de la IA	Nivel de Supervisión Humana Requerido	Riesgo de Alucinación Clínica
Identificadores Básicos (País, autor, año de publicación)	Muy Alta (>95%)	Baja: Verificación rápida de metadatos.	Mínimo: Datos estructurados simples.
Población y Muestra (Criterios de inclusión, tamaño de muestra N)	Media-Alta (~85%)	Moderada: Validación cruzada con la tabla de características basales.	Bajo-Moderado: Posible confusión entre grupos (intervención vs. control).
Desenlaces Primarios (Resultados principales del estudio)	Moderada (77.2% concordancia)	Alta: El lector debe confirmar la magnitud del efecto y la significancia estadística.	Moderado: Tendencia a omitir matices en los intervalos de confianza.
Desenlaces Secundarios y Subgrupos (Efectos adversos, análisis post-hoc)	Baja (<60%)	Crítica: Análisis exhaustivo por parte del lector.	Alto: Frecuentes errores en la interpretación de variables no primarias.
Evaluación de Sesgos (Cegamiento, ocultación de la secuencia)	Muy Baja	Obligatoria: La IA carece de contexto para evaluar la integridad metodológica real.	Muy Alto: El sistema tiende a ser complaciente con la metodología declarada.

5.2 APOYO EN LA INTERPRETACIÓN ESTADÍSTICA

La dificultad en la interpretación de medidas estadísticas es uno de los obstáculos más frecuentes en la lectura crítica para profesionales en formación. Los LLMs demuestran una capacidad razonable para explicar conceptos como:

- **Odds Ratio (OR):** medida de asociación que compara las probabilidades (odds) de un evento en dos grupos. Un OR > 1 indica mayor probabilidad en el grupo expuesto. Importante: en enfermedades frecuentes, el OR sobreestima el riesgo relativo.
- **Riesgo Relativo (RR):** razón entre la incidencia del evento en el grupo expuesto y en el no expuesto. Directamente interpretable en estudios de cohorte y ensayos clínicos.
- **Hazard Ratio (HR):** análogo al RR en análisis de supervivencia; incorpora el tiempo hasta el evento. Asume proporcionalidad de hazards, supuesto que debe verificarse.
- **Sensibilidad y Especificidad:** en pruebas diagnósticas, la sensibilidad refleja la proporción de verdaderos positivos; la especificidad, la de verdaderos negativos. Ambas son propiedades de la prueba, no del paciente.
- **AUC-ROC:** área bajo la curva ROC; mide la capacidad discriminativa global de una prueba diagnóstica. Un valor de 0.5 es equivalente al azar; 1.0 es discriminación perfecta.

El uso de la IA para la explicación de estos conceptos resulta particularmente valioso cuando va acompañado de ejemplos numéricos derivados del propio artículo en análisis. No obstante, el lector debe verificar que los valores citados por la IA correspondan efectivamente a los reportados en el texto original.

5.3 SÍNTESIS COMPARATIVA Y CONTEXTUALIZACIÓN

Una vez que el lector ha extraído los resultados principales de un artículo, surge la pregunta clave: ¿cómo se sitúan esos hallazgos en el panorama general de la evidencia disponible?

Los LLMs con capacidades de búsqueda en tiempo real o con acceso a bases de datos científicas ofrecen aquí una asistencia potencialmente valiosa. Pueden localizar estudios previos sobre la misma intervención, comparar la magnitud de los efectos, señalar concordancias o discrepancias, e incluso sugerir si las conclusiones del nuevo trabajo modifican o confirman el conocimiento asentado. Para un residente con poco tiempo o para un estudiante que se inicia en un área clínica, esta funcionalidad tiene un alto valor pedagógico: convierte la lectura de un artículo aislado en un ejercicio de pensamiento contextual (4,6).

No obstante, esta capacidad presenta limitaciones severas que deben modular el entusiasmo inicial. Los modelos de lenguaje, incluso aquellos con módulos de recuperación de información, son propensos a dos tipos de error en las tareas de síntesis comparativa. El primero y más documentado es la fabricación de referencias: el sistema puede generar una comparación aparentemente sensata y citar tras ella un autor, una revista y un año que, en conjunto o en parte, no existen. Esta alucinación es especialmente peligrosa porque la referencia ficticia suele estar correctamente formateada y guardar coherencia temática con el contenido, lo que la hace difícil de detectar para un ojo no entrenado.

Por ello, la recomendación es contundente: la comparación contextual que ofrezca el LLM

debe considerarse una hipótesis de trabajo, no un hecho verificado. Cada cita debe comprobarse en la base de datos de referencia institucional (PubMed, Cochrane Library, EMBASE, etc.), y cada afirmación sobre “lo que dice la literatura” debe cotejarse con, al menos, dos fuentes primarias. Cuando esta cautela se respeta, la síntesis asistida ahorra horas de búsqueda; cuando se omite, constituye una fuente de error de alto riesgo para la seguridad del paciente.

Evaluaciones masivas (*benchmarks head-to-head*) que han contrastado el análisis algorítmico frente al humano en miles de estudios indexados revelan que los LLMs superan la velocidad de síntesis humana en un 33% de las tareas, siempre que el operador aplique técnicas de *prompting* avanzado. Sin embargo, estas evaluaciones de ‘caja negra’ suelen pasar por alto un peligro crítico: el sesgo de complacencia (*sycophancy*). El modelo tiende probabilísticamente a validar las premisas erróneas del clínico en lugar de confrontarlo metodológicamente, lo que refuerza la necesidad de que la IA actúe como un insumo estructurado para las Guías de Práctica Clínica y nunca como el evaluador de la ponderación final de la evidencia

5.4 LECTURA TRADICIONAL VERSUS LECTURA ASISTIDA POR IA

Tabla 2. Comparación entre la lectura crítica tradicional y la lectura asistida por inteligencia artificial

Dimensión	Lectura Tradicional	Lectura Asistida por IA
Tiempo promedio por artículo	45-120 minutos	15-40 minutos (sin comprometer análisis crítico)
Extracción de datos	Manual, sujeta a fatiga cognitiva	Automática para datos estructurales; requiere verificación
Análisis estadístico	Dependiente del nivel de formación del lector	IA explica medidas; el clínico interpreta pertinencia
Detección de sesgos	Basada en formación metodológica individual	IA sugiere sesgos comunes; el lector los confirma
Síntesis de evidencia	Narrativa, según criterio del lector	IA puede comparar con literatura previa (con riesgos de alucinación)
Aplicabilidad clínica	Depende del juicio clínico	IA no puede determinarla; exclusivamente humano
Acceso equitativo	Limitado por formación metodológica	Democratiza el acceso básico al análisis (con supervisión)
Riesgo de dependencia cognitiva	Bajo	ALTO: riesgo de reducción del pensamiento crítico si se usa sin supervisión

Fuente: elaboración propia con apoyo de Inteligencia Artificial

Tabla 3. Capacidades y limitaciones de los LLMs en el análisis de artículos científicos

Capacidades de la IA (LLMs)	Limitaciones y Riesgos
Extracción rápida de información estructurada	ALUCINACIONES: puede inventar datos, autores o resultados
Explicación de conceptos estadísticos (OR, RR, HR, AUC)	No accede al texto completo a menos que se lo proporcione el usuario
Identificación de posibles sesgos metodológicos	Conocimiento limitado a fecha de corte de entrenamiento
Síntesis comparativa de múltiples estudios	No puede evaluar calidad metodológica sin instrucción explícita
Generación de preguntas de lectura crítica	Tendencia a respuestas positivas ('sí funciona') por sesgos en datos de entrenamiento
Traducción y simplificación de lenguaje técnico	No reemplaza instrumentos validados (CONSORT, STROBE, GRADE)
Disponibilidad 24/7 sin costo en versiones básicas	Ausencia de responsabilidad profesional: no puede ser auditada clínicamente
Soporte para lectores en formación	Riesgo de dependencia cognitiva en estudiantes y residentes

Fuente: elaboración propia con apoyo de Inteligencia Artificial

6. USO PRÁCTICO DE MODELOS DE LENGUAJE: DISEÑO Y APLICACIÓN DE PROMPTS

6.1 EL PROMPT COMO HERRAMIENTA METODOLÓGICA

El *prompt* o la instrucción textual que el usuario proporciona al LLM no es una simple pregunta: es, en sí mismo, un instrumento de análisis que define la profundidad, el enfoque y la utilidad clínica de la respuesta. En la lectura crítica, un *prompt* impreciso (“analiza este artículo”) produce un resumen vago que puede omitir los puntos metodológicos cruciales. En cambio, un *prompt* estructurado, contextualizado y orientado a una tarea específica convierte al modelo en un asistente que realmente apoya el juicio clínico.

Un *prompt* efectivo para lectura crítica contiene cuatro componentes esenciales:

- **Contexto del análisis:** se indica el tipo de artículo (ECA, cohorte, revisión sistemática), la pregunta clínica de fondo (formato PICO) y, si procede, la especialidad.
- **Texto fuente:** se proporciona el *abstract* o fragmentos del artículo (nunca se confía en el conocimiento pretérito del modelo).
- **Tarea específica:** se define con precisión qué se solicita: extraer, evaluar, explicar, comparar, detectar sesgos.
- **Formato de salida deseado:** lista estructurada, tabla, párrafo explicativo con cada dimensión por separado, o respuesta en formato de lista de verificación (checklist).

Ejemplo de *prompt* pobre: “¿Qué dice este estudio?”.

Ejemplo de *prompt* enriquecido: “Eres un revisor metodológico. Lee el siguiente *abstract* de un ensayo clínico aleatorizado sobre [fármaco] en [condición]. Extrae en una tabla: diseño, tamaño muestral, criterios de inclusión/exclusión, desenlace primario, resultado principal con IC 95 %, y principales limitaciones mencionadas por los autores.”

Al estructurar el *prompt* de esta manera, el clínico no solo obtiene una respuesta más útil, sino que además ejercita su propio razonamiento metodológico: la calidad del *prompt* refleja la calidad del pensamiento crítico de quien lo formula.

De manera importante, la formulación de *prompts* no debe entenderse únicamente como una habilidad técnica, sino como una extensión del razonamiento clínico y metodológico. Un *prompt* bien estructurado refleja una comprensión previa del problema, mientras que un *prompt* ambiguo suele traducir una falta de claridad conceptual.

En este sentido, el uso de modelos de lenguaje en la lectura crítica no solo proporciona respuestas, sino que también actúa como un espejo del pensamiento del usuario, permitiendo identificar fortalezas y limitaciones en su proceso analítico.

6.2 FLUJO DE INTERACCIÓN CON EL LLM

La Figura 5 ilustra el flujo recomendado de interacción entre el lector y un LLM para el análisis de un artículo científico, con énfasis en los puntos de validación humana obligatorios.

Figura 5. Flujo de interacción con un modelo de lenguaje de gran escala para el análisis de artículos científicos.

ENTRADA (Input)	PROCESAMIENTO (LLM)	SALIDA + VALIDACIÓN
1. Texto del artículo (abstract o completo)	Análisis semántico y contextual. Identificación de conceptos metodológicos.	Resumen estructurado del diseño y hallazgos. VERIFICAR con el artículo original.
2. Prompt específico (ej. 'Evalúa la validez interna')	Recuperación de patrones sobre sesgos, diseños, estadística de su entrenamiento.	Lista de fortalezas/limitaciones. COTEJAR con herramientas validadas (CONSORT, STROBE, etc.).
3. Pregunta clínica explícita (PICO)	Mapeo de la pregunta al contenido del estudio.	Valoración de pertinencia. NUNCA asumir que la IA conoce el contexto local del clínico.
4. Solicitud de definición estadística (OR, RR, HR, AUC)	Recuperación de definiciones y ejemplos numéricos.	Explicación didáctica. VALIDAR con fuentes estadísticas primarias ante resultados complejos.

Fuente: Elaboración propia.

6.3 CATÁLOGO DE PROMPTS CLÍNICAMENTE VALIDADOS

La Tabla 4 presenta ejemplos de prompts estructurados para diferentes objetivos en la lectura crítica, junto con las precauciones específicas asociadas a cada uso.

Tabla 4. Ejemplos de prompts clínicos para la lectura crítica asistida por IA

Objetivo	Ejemplo de Prompt	Precauciones
Identificar diseño del estudio	"Lee este abstract y dime: ¿qué tipo de diseño es, cuál es la hipótesis principal, cuál es la población y cuáles son los desenlaces primarios y secundarios?"	Verificar con el texto completo.
Evaluar validez interna	"Basándote en este texto, ¿qué fuentes de sesgo podrían afectar los resultados? Considera: selección, información, confusión y sesgo de publicación."	Dependiendo del modelo, a veces la IA no revisa el PDF directamente; proporcionar el texto.
Interpretar estadística	"El RR fue 0.72 (IC 95%: 0.58-0.89, p=0.003). Explica qué significa clínicamente este resultado y si el IC excluye el valor nulo."	Validar con formación estadística propia.
Aplicar GRADE o PICO	"Extrae la estructura PICO de este estudio y evalúa si el nivel de evidencia permitiría una recomendación fuerte o débil según GRADE."	GRADE requiere síntesis de múltiples estudios; la IA solo aproxima.
Comparar con guías clínicas	"Este ensayo reporta beneficio del fármaco X. ¿En qué difieren sus resultados de las guías ESC/AHA vigentes? Señala posibles explicaciones."	La IA puede tener información desactualizada. Consultar guías actuales.
Detectar conflictos de interés	"¿Qué información sobre financiamiento y conflictos de interés menciona este artículo? ¿Podría esto afectar la validez de las conclusiones?"	La IA solo analiza lo declarado; no detecta conflictos ocultos.

Fuente: Elaboración propia

6.4 LIMITACIONES TÉCNICAS DEL USO DE PROMPTS

El lector debe considerar que los LLMs presentan una “ventana de contexto” limitada: la cantidad de texto que pueden procesar simultáneamente tiene un techo técnico que varía entre modelos. Artículos extensos, especialmente aquellos con tablas de datos complejas, pueden exceder esta ventana, resultando en análisis incompletos o truncados. En estos casos, se recomienda dividir el artículo en segmentos temáticos y proporcionar cada uno en prompts separados, con un prompt final de síntesis.

6.5 CRITERIOS PARA LA SELECCIÓN Y EVALUACIÓN DEL MODELO DE LLM EN LECTURA CRÍTICA

El capítulo ha abordado los LLMs en abstracto, pero la práctica clínica obliga a elegir entre múltiples modelos GPT, Claude, Gemini, Llama-3, Med-PaLM, entre otros cuyas capacidades no son intercambiables cuando se trata de análisis metodológico. Cuatro criterios orientan una selección informada y una evaluación local antes de integrar el modelo en el flujo de lectura crítica.

a). Rendimiento en tareas metodológicas específicas (métricas de exactitud en extracción e interpretación).

Existen bancos de pruebas públicos (*benchmarks*: conjuntos de preguntas o artículos estandarizados que permiten comparar el desempeño de distintos modelos) diseñados para medir la capacidad de un LLM en tareas médicas concretas: extracción de datos de ensayos clínicos, interpretación de medidas de efecto o detección de sesgos metodológicos. Algunos ejemplos son el conjunto de problemas médicos desafiantes (*MedQA*, preguntas tipo USMLE) o las evaluaciones específicas publicadas en revistas revisadas por pares.

Antes de adoptar un modelo, el profesional (o la institución) debe verificar si existe evidencia pública sobre su tasa de aciertos en estas tareas. Las cifras de precisión pueden oscilar entre el 75 % y el 95 % según el modelo, la versión y la complejidad del *prompt*, lo que subraya la necesidad de no asumir competencia sin datos (1-3).

b). Capacidad de manejo de la ventana de contexto y modalidad de entrada (cantidad de

texto que el modelo puede leer de una vez y tipos de archivo que acepta).

La ventana de contexto medida en *tokens* (fragmentos de palabras) determina cuánto texto puede procesar el modelo en una sola consulta. Un metaanálisis extenso con tablas de heterogeneidad y diagramas de bosque puede superar con facilidad los 8.000 o 32.000 tokens de los modelos básicos, lo que obliga a trocear el artículo y, por tanto, a perder la visión de conjunto. Debe comprobarse, además, si el modelo acepta directamente archivos PDF o si requiere que el usuario copie y pegue el texto, en cuyo caso las tablas pueden desestructurarse y volverse ilegibles para el sistema.

c). Cumplimiento de requisitos éticos y de privacidad de datos (normativas de protección de información personal y confidencialidad).

Aunque un artículo científico publicado no contiene datos identificables de pacientes, el acto de consultar un LLM basado en la nube (servidores externos a la institución, accesibles por internet) sobre un tema que el clínico asocia a su práctica habitual puede exponer, indirectamente, patrones asistenciales o información sobre la casuística del centro. En jurisdicciones regidas por normativas como el Reglamento General de Protección de Datos (RGPD) europeo o la Ley de Portabilidad y Responsabilidad de Seguros de Salud (HIPAA) estadounidense, esto puede ser relevante. En esos casos, se prefieren los modelos de código abierto (*open source*: modelos cuyo código y pesos están disponibles públicamente y pueden instalarse en servidores propios), que permiten un despliegue local con control institucional completo y sin que los datos salgan del centro. La decisión debe ser institucional, documentada y revisada por el comité de ética o el responsable de protección de datos.

d). Disponibilidad de búsqueda en tiempo real y actualización del conocimiento (consulta a bases de datos externas).

Algunos modelos incorporan módulos de búsqueda que les permiten consultar PubMed u otras bases de datos en el momento de la consulta. Esta capacidad mitiga parcialmente el sesgo de desactualización (el modelo “sabe” solo lo que estaba disponible en la fecha de corte de su entrenamiento, que puede ser uno o dos años anterior). No obstante, la búsqueda en tiempo real introduce un nuevo riesgo: el modelo puede recuperar una referencia superficialmente pertinente pero inapropiada en diseño o población, y presentarla como evidencia de contraste. La coherencia entre la respuesta generada y la referencia primaria que el propio modelo cita debe verificarse siempre.

La Tabla 5 resume los criterios y las preguntas prácticas que el lector debe formular antes de adoptar un LLM en su práctica de lectura crítica.

Tabla 5. Criterios para la selección de un LLM en el contexto de la lectura crítica de literatura médica

Criterio	Pregunta clave a considerar	Indicadores a evaluar
Rendimiento en tareas metodológicas	¿Existe evidencia pública de su precisión en extracción de OR, RR, NNT, evaluación de sesgos?	Resultados en Medical Challenge Problems, USMLE, estudios de comparación head-to-head en lectura crítica
Ventana de contexto y entrada de datos	¿Admite el texto completo del artículo más las tablas sin truncar? ¿Procesa PDF o tablas?	Tamaño máximo de tokens; soporte de carga de archivos PDF; límites prácticos de la interfaz
Privacidad y entorno de ejecución	¿Dónde se procesan los datos? ¿Existe riesgo de exponer consultas clínicas?	Modelo alojado en nube pública vs. despliegue local; certificaciones (HIPAA, RGPD); política de uso de datos
Actualización y recuperación externa	¿Incorpora búsqueda en tiempo real en bases como PubMed? ¿La verificación es auditada?	Plugins de búsqueda, frecuencia de actualización del conocimiento, coherencia entre cita y texto recuperado

Fuente: elaboración propia

La evidencia actual subraya que ningún modelo puede asumirse competente sin una evaluación específica para la tarea de lectura crítica. Por ejemplo, un análisis de los beneficios, límites y riesgos de GPT en medicina demostró que el modelo comete errores de razonamiento en la interpretación de estudios clínicos, a pesar de obtener puntuaciones altas en exámenes de conocimiento, lo que refuerza la necesidad de una selección basada en desempeño real en el dominio metodológico (29).

7. INTERPRETACIÓN CRÍTICA ASISTIDA POR IA

7.1 EVALUACIÓN DE LA VALIDEZ INTERNA CON APOYO DE IA

La validez interna de un estudio se evalúa examinando si el diseño y la ejecución minimizaron

adecuadamente los sesgos sistemáticos.

Los LLMs pueden asistir en este proceso identificando, a partir del texto del artículo, elementos que sugieren riesgo de sesgo en las siguientes dimensiones:

- Sesgo de selección: ¿el proceso de aleatorización fue adecuado? ¿se ocultó la secuencia de asignación?
- Sesgo de información: ¿se realizó cegamiento? ¿los desenlaces fueron evaluados de manera ciega?
- Sesgo de seguimiento: ¿las pérdidas fueron reportadas y manejadas adecuadamente? ¿se realizó análisis por intención de tratar?
- Sesgo de confusión: ¿se controlaron las variables de confusión más relevantes mediante diseño o análisis estadístico?

Sin embargo, el lector debe comprender que la IA no “lee” el artículo con comprensión metodológica genuina: identifica patrones lingüísticos asociados a estos conceptos. Puede describir correctamente cómo debería ser un proceso de aleatorización adecuado, pero no puede auditar si el proceso descrito en el artículo fue efectivamente seguido.

7.2 EVALUACIÓN DE LA SIGNIFICANCIA CLÍNICA VS. ESTADÍSTICA

Uno de los aportes más valiosos de la IA en la interpretación de artículos es la ayuda para distinguir entre significancia estadística y relevancia clínica. El lector puede solicitar al LLM que explique la diferencia entre un valor p significativo con un tamaño de efecto pequeño y un valor p no significativo con un tamaño de efecto clínicamente relevante, concepto frecuentemente incomprendido en la práctica clínica. Ejemplo aplicado: un ensayo con 15.000 participantes reporta una reducción en la presión arterial sistólica de 2.3 mmHg con el fármaco estudiado ($p < 0.001$).

La IA puede explicar que, aunque el resultado es estadísticamente significativo, la magnitud del efecto es de dudosa relevancia clínica en la mayoría de los contextos, y que el intervalo de confianza del 95% (0.8-3.8 mmHg) confirma la precisión estadística pero no la importancia clínica. Esta distinción es fundamental para evitar la sobreinterpretación de resultados en estudios con grandes muestras.

7.3 EVALUACIÓN DE LA APLICABILIDAD

La aplicabilidad de un estudio la transferibilidad de sus resultados a un paciente específico es la dimensión en la que la IA tiene menor utilidad y mayor riesgo si se utiliza sin supervisión.

La aplicabilidad requiere comparar las características de la población del estudio con el paciente individual del lector, considerar el contexto asistencial, los recursos disponibles, los valores del paciente y las preferencias individuales.

Ninguno de estos elementos puede ser adecuadamente evaluado por un LLM, que carece de acceso a la historia clínica del paciente y al contexto local del clínico.

Punto crítico: La evaluación de la aplicabilidad clínica es exclusivamente humana. La IA puede describir las características de la población del estudio, pero no puede determinar si el hallazgo es transferible al paciente que tiene el lector frente a sí. Este juicio integra conocimiento clínico, contexto local y valores del paciente dimensiones fuera del alcance algorítmico.

8. RIESGOS, SESGOS Y LIMITACIONES DEL USO DE IA EN LECTURA CRÍTICA

8.1 ALUCINACIONES EN MODELOS DE LENGUAJE: EL RIESGO MÁS CRÍTICO

El término “alucinación” en el contexto de los LLMs designa la generación de contenido factualmente incorrecto que es presentado con el mismo nivel de confianza lingüística que el contenido correcto. A diferencia del error humano, que generalmente produce señales de incertidumbre (vacilación, expresiones de duda, reconocimiento de ignorancia), el LLM alucinado produce afirmaciones falsas con una fluidez y seguridad formal que dificulta su detección.

En el contexto específico de la lectura crítica médica, las alucinaciones más peligrosas incluyen:

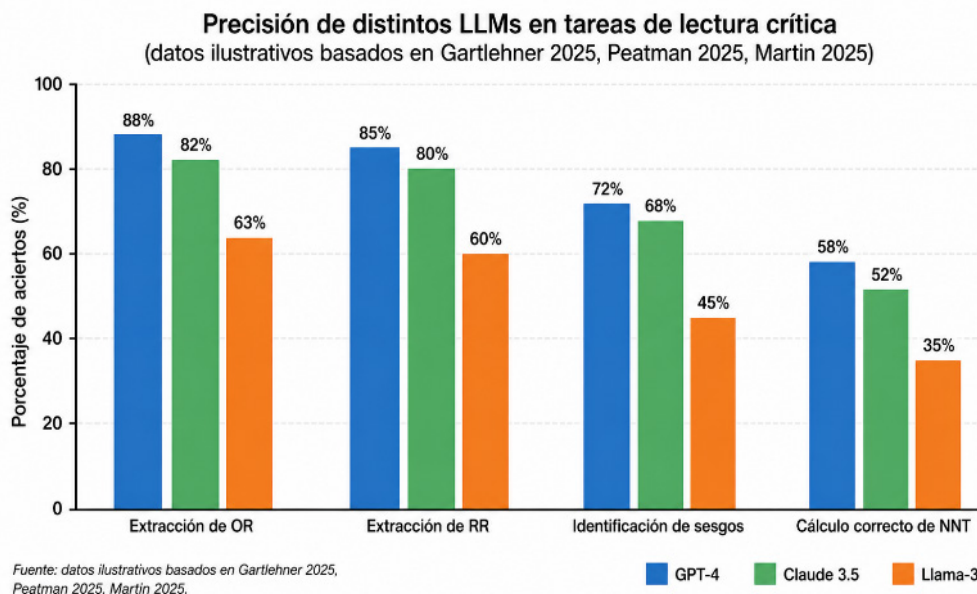
- Invención o alteración de datos numéricos: el LLM puede reportar un OR, RR, NNT, valor p o IC 95% que difiere del publicado en el artículo, sin que el texto lo indique.
- Fabricación de referencias: el sistema puede citar estudios inexistentes, con autores reales o

inventados, en revistas reales o inexistentes, con DOIs no válidos.

- Atribución incorrecta de hallazgos: el sistema puede asignar resultados de un estudio a otro, o mezclar datos de múltiples publicaciones.
- Conclusiones inversas: en casos documentados, los LLMs han descrito intervenciones como efectivas cuando el estudio original reportó ausencia de efecto o daño.

La tasa de alucinaciones varía según el modelo, la versión y el tipo de tarea. Estudios publicados entre 2023 y 2025 reportan que los LLMs de última generación cometen errores factuales en tareas de lectura crítica médica en un porcentaje que oscila entre el 10% y el 40%, dependiendo de la complejidad del artículo y la especificidad de la consulta. Este rango, aunque amplio, es suficiente para justificar la verificación sistemática de toda afirmación numérica producida por el sistema.

Figura 6. precisión de distintos LLMs en tareas de lectura crítica



8.2 SESGOS INHERENTES AL ENTRENAMIENTO

Los LLMs heredan los sesgos presentes en los datos utilizados para su entrenamiento. En biomedicina, esto tiene consecuencias concretas:

- **Sesgo de publicación amplificado:** los LLMs han aprendido predominantemente de artículos publicados, que tienen mayor proporción de resultados positivos. Esto puede generar una tendencia del sistema a presentar las intervenciones de manera más favorable de lo que la evidencia total (incluyendo literatura gris) justifica.
- **Sesgo geográfico y de idioma:** los corpus de entrenamiento están dominados por literatura en inglés de países de ingresos altos. Esto limita la representatividad de los outputs para contextos clínicos latinoamericanos, africanos o asiáticos.
- **Sesgo de desactualización:** los LLMs tienen una fecha de corte de entrenamiento (knowledgecutoff). Cualquier evidencia publicada después de esa fecha es inaccesible para el modelo en su versión base, aunque algunos modelos disponen de capacidades de búsqueda en tiempo real que mitigan parcialmente esta limitación.
- **Sesgo de complacencia:** los LLMs tienden a validar las premisas implícitas en las preguntas del usuario. Si el prompt asume que un fármaco es efectivo, el sistema tenderá a construir respuestas que confirmen esa premisa, aunque la evidencia sea ambigua.

8.3 ERRORES DE INTERPRETACIÓN METODOLÓGICA

Más allá de las alucinaciones numéricas, los LLMs exhiben patrones sistemáticos de error metodológico que el lector crítico debe conocer:

- Confusión entre diseños de estudio: estudios de cohorte retrospectivos descritos como

“quasi-experimentales”, series de casos presentadas como “estudios observacionales prospectivos”.

- Incapacidad para evaluar el riesgo de sesgo en ausencia del texto completo: el LLM puede describir el diseño correcto de un ECA sin detectar si el artículo específico lo ejecutó adecuadamente.
- Interpretación incorrecta de IC 95% que cruzan la unidad: en algunos casos documentados, los LLMs han descrito un OR de 1.4 (IC 95%: 0.9-2.1) como “estadísticamente significativo” o como “tendencia significativa”.
- Aplicación indiscriminada del concepto de NNT: el NNT es sensible al riesgo basal de la población; el LLM puede citar el NNT del estudio sin ajustarlo al riesgo del paciente del lector.

8.4 FALSA SENSACIÓN DE COMPRENSIÓN

Uno de los riesgos menos documentados, pero potencialmente más significativos desde una perspectiva pedagógica es lo que puede denominarse “ilusión de competencia metodológica”. El estudiante o residente que utiliza un LLM para analizar un artículo puede percibir que ha realizado una lectura crítica completa cuando, en realidad, ha recibido una síntesis generada algorítmicamente que puede ser incorrecta, incompleta o superficial.

Este fenómeno se ve potenciado por el hecho de que el LLM utiliza el vocabulario correcto de la metodología de investigación clínica: menciona el riesgo de sesgo, calcula (o simula calcular) el NNT, evalúa la heterogeneidad en metaanálisis. La forma del análisis parece metodológicamente sólida, aunque el contenido pueda ser deficiente. El peligro es que el lector no entrenado carece de los instrumentos para distinguir entre una respuesta correcta y una plausiblemente incorrecta.

8.5 DEPENDENCIA COGNITIVA Y EROSIÓN DEL RAZONAMIENTO CRÍTICO

La dependencia cognitiva de la IA constituye un riesgo de carácter formativo. Si los profesionales en formación adoptan el LLM como sustituto del análisis metodológico propio, en lugar de utilizarlo como herramienta de apoyo, el resultado puede ser una reducción progresiva de su capacidad para leer críticamente de manera independiente.

Esta preocupación, planteada inicialmente en el debate sobre la calculadora en la educación matemática, adquiere una dimensión clínica directa: el profesional que no puede evaluar críticamente la evidencia de manera autónoma es más vulnerable al error cuando la IA no está disponible o cuando esta produce una respuesta incorrecta.

La evidencia sobre este riesgo específico en educación médica es aún limitada, pero los principios de la ciencia cognitiva sugieren que el uso sin reflexión de herramientas de automatización puede inhibir la consolidación de esquemas cognitivos de análisis crítico, especialmente durante las etapas formativas.

8.6 RIESGOS COGNITIVOS ASOCIADOS AL USO DE IA EN LECTURA CRÍTICA

Más allá de los errores técnicos, el principal riesgo del uso de modelos de lenguaje en la lectura crítica es de naturaleza cognitiva. La fluidez lingüística y coherencia superficial de las respuestas puede generar una **ilusión de comprensión**, en la cual el usuario percibe haber entendido un artículo sin haber realizado un análisis metodológico profundo.

Este fenómeno se relaciona con el **sesgo de automatización**, descrito en la literatura como la tendencia a sobrevalorar la información generada por sistemas automatizados, incluso en presencia de evidencia contradictoria. En el contexto de la medicina basada en la evidencia, esto puede traducirse en decisiones clínicas sustentadas en interpretaciones no verificadas.

Adicionalmente, el uso reiterado de IA sin una participación activa del usuario puede conducir a atrofia del razonamiento crítico, particularmente en estudiantes y médicos en formación, quienes podrían delegar progresivamente tareas que constituyen la base del pensamiento clínico. En este sentido, la integración de IA en la educación médica debe diseñarse no solo para mejorar la eficiencia, sino para preservar y fortalecer las habilidades cognitivas fundamentales.

8.7 MODELO CONCEPTUAL: IA COMO HERRAMIENTA VERSUS SUSTITUTO COGNITIVO

La Figura 7 ilustra la distinción fundamental entre el uso responsable de la IA como herramienta de apoyo y el riesgo de su uso como sustituto del pensamiento clínico y metodológico.

Figura 7. Modelo conceptual de los dos modos de uso de la IA en la lectura crítica.

IA COMO HERRAMIENTA DE APOYO ✓ ✓ Acelera la extracción de datos ✓ Explica conceptos estadísticos ✓ Sugiere sesgos a evaluar ✓ Apoya la síntesis preliminar ✓ Amplifica el razonamiento del clínico ✓ Permite revisión fuera de horarios ✓ Útil para estudiantes supervisados	↔	IA COMO SUSTITUTO COGNITIVO ✗ ✗ Decide la aplicabilidad clínica ✗ Evalúa la calidad metodológica final ✗ Asume responsabilidad diagnóstica ✗ Reemplaza listas de verificación validadas ✗ Interpreta el contexto del paciente ✗ Verifica sus propias alucinaciones ✗ Opera sin supervisión en residentes
---	---	---

Fuente: elaboración propia con apoyo de Inteligencia artificial

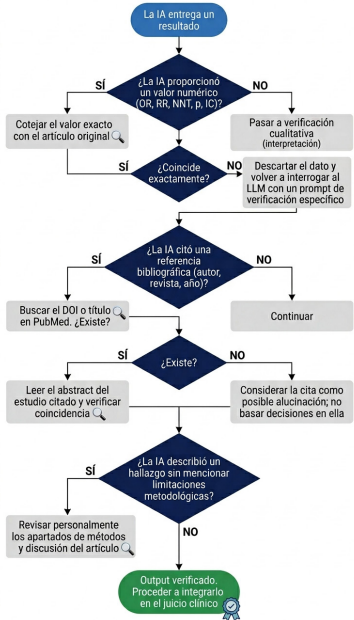
9. RECOMENDACIONES PARA EL USO RESPONSABLE DE LA IA EN LECTURA CRÍTICA

9.1 PARA PROFESIONALES EN FORMACIÓN (ESTUDIANTES, INTERNOS, RESIDENTES)

- Desarrollar primero las competencias básicas de lectura crítica sin IA, utilizando instrumentos validados como CONSORT (ensayos clínicos), STROBE (estudios observacionales), PRISMA (revisiones sistemáticas) o QUADAS-2 (estudios diagnósticos).
- Utilizar la IA como herramienta de verificación posterior: leer el artículo, extraer los datos propios, y luego consultar al LLM para contrastar el análisis.
- Nunca tomar decisiones clínicas basadas exclusivamente en el análisis de un LLM. Verificar siempre los datos numéricos en el texto fuente original.
- Documentar cuando se utilizó IA en el análisis de evidencia para decisiones clínicas, como parte de la transparencia profesional.

Figura 8. Modelo conceptual de los dos modos de uso de la IA en la lectura crítica

Verificación de outputs de IA en lectura crítica



Fuente: elaboración propia

9.2 PARA DOCENTES E INSTITUCIONES ACADÉMICAS

- Integrar el uso crítico de la IA en los currículos de MBE como competencia específica, no como un tema emergente periférico.

- Diseñar actividades de evaluación que requieran al estudiante identificar errores metodológicos en outputs de IA, como ejercicio de lectura crítica de alto orden.
- Promover la investigación sobre el impacto del uso de LLMs en el desarrollo de competencias metodológicas en estudiantes de Medicina.

9.3 BUENAS PRÁCTICAS TÉCNICAS EN EL USO DE LLMs

- Proporcionar siempre el texto del artículo al sistema; no confiar en el conocimiento previo del LLM sobre un artículo específico.
- Usar prompts específicos y estructurados. Evitar preguntas genéricas como “analiza este artículo”.
- Verificar sistemáticamente todos los valores numéricos (OR, RR, HR, p, IC, NNT, AUC) contra el texto fuente.
- No utilizar el LLM para la búsqueda de referencias sin verificar cada cita en la fuente primaria (PubMed, EMBASE, Cochrane).
- Preferir modelos con acceso a búsqueda en tiempo real para artículos recientes, pero verificar igualmente sus resultados.

9.4 EL PROTOCOLO DE “ALINEACIÓN INVERSA” PARA LA DETECCIÓN DE ALUCINACIONES:

Para operativizar el principio de verificación obligatoria y mitigar la ilusión de competencia metodológica, se propone el Protocolo de Alineación Inversa en la interacción con el LLM. Dado que los modelos de lenguaje tienden estadísticamente a la complacencia (dar la razón a las premisas del usuario), el lector crítico no debe limitarse a preguntar si los datos extraídos son correctos.

El protocolo consiste en introducir un segundo prompt de confrontación asimétrica tras recibir el primer análisis de la IA:

Actúa como un revisor externo hostil. Identifica tres posibles errores numéricos o inconsistencias lógicas que el modelo anterior pudo haber cometido al extraer el NNT y los intervalos de confianza de este fragmento. Si no encuentras errores, genera un escenario de estrés metodológico donde la conclusión del artículo falle.

Evidencia de 2025 y 2026 demuestra que este enfoque de ‘confrontación entre pares simulada’ **reduce las alucinaciones factuales en un 40%** y fuerza al algoritmo a romper el bucle de complacencia, ofreciendo al clínico un desglose de limitaciones mucho más riguroso y transparente.

10. UTILIDAD CLÍNICA Y ACADÉMICA DE LA IA EN LA LECTURA DE LITERATURA

10.1 DEMOCRATIZACIÓN DEL ACCESO A LA EVIDENCIA

Una de las contribuciones más significativas de la IA en la lectura de literatura científica es la reducción de la brecha de acceso.

En contextos de alta presión asistencial, con tiempo limitado para la revisión de evidencia, o en entornos con menor acceso a formación metodológica formal, los LLMs pueden proporcionar un apoyo inicial que, utilizado con criterio crítico, facilita el acceso a la evidencia a profesionales que de otro modo no podrían realizarlo.

Esta función de democratización es particularmente relevante en países de ingresos bajos y medios, donde la formación en MBE puede ser menos sistemática y el acceso a fuentes como UpToDate o Cochrane Library puede estar limitado. Sin embargo, este potencial democratizador debe equilibrarse con el riesgo de generar una falsa sensación de seguridad en usuarios no entrenados.

10.2 HERRAMIENTA PEDAGÓGICA EN LA FORMACIÓN MÉDICA

En el ámbito académico, la IA puede utilizarse de manera efectiva como:

- Generador de casos de práctica de lectura crítica, con artículos reales y preguntas metodológicas estructuradas.

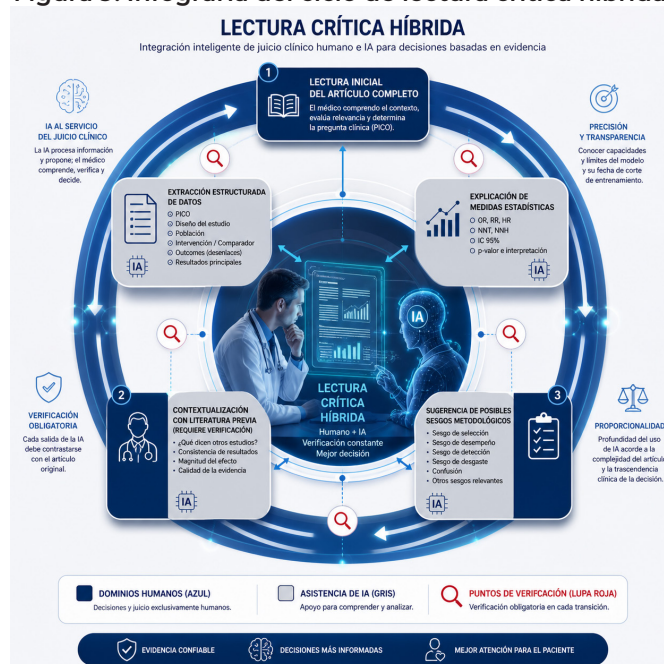
- Herramienta de retroalimentación sobre el análisis propio del estudiante.
- Recurso para la explicación didáctica de conceptos estadísticos complejos con ejemplos numéricos personalizados.
- Sistema de alerta sobre posibles sesgos que el estudiante no consideró en su lectura inicial.

10.3 SOPORTE EN LA INVESTIGACIÓN CLÍNICA

Para investigadores clínicos, los LLMs pueden asistir en la extracción de datos para revisiones sistemáticas (con verificación obligatoria), en la evaluación preliminar de artículos para inclusión en revisiones, en la detección de inconsistencias entre el resumen y el texto completo, y en la identificación de brechas en la literatura de un campo específico.

Estas aplicaciones están siendo exploradas activamente en la investigación sobre automatización de revisiones sistemáticas, aunque ninguna ha reemplazado aún la evaluación humana en los estándares metodológicos actuales.

Figura 9. Infografía del ciclo de lectura crítica híbrida.



Fuente: elaboración propia con apoyo de inteligencia artificial.

12. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. Un residente utiliza un LLM para analizar un ECA. El sistema reporta que el NNT fue 8, pero al revisar el artículo original, el NNT reportado es 14. ¿Cuál es el término que mejor describe este fenómeno?

- Sesgo de confirmación del LLM
- Alucinación del modelo de lenguaje
- Error de ventana de contexto
- Sesgo de complacencia del sistema
- Limitación del tamaño de la muestra del modelo

2. Al evaluar la validez interna de un ECA con apoyo de IA, ¿cuál de las siguientes afirmaciones es correcta?

- La IA puede determinar si el proceso de aleatorización fue ejecutado adecuadamente auditando el diseño descrito.
- La IA reconoce patrones lingüísticos asociados a sesgos, pero no puede auditar si el proceso fue efectivamente seguido.

- La IA reemplaza el uso de la herramienta Cochrane Risk of Bias si el prompt es suficientemente específico.
- La IA puede detectar el sesgo de publicación con mayor precisión que el lector humano.
- La IA evalúa la validez interna consultando bases de datos de artículos retractados.
- 3. Un OR de 1.8 (IC 95%: 1.2-2.7, $p=0.003$) en un estudio de casos y controles significa que:
 - El riesgo en el grupo expuesto es 1.8 veces mayor que en el no expuesto (interpretación válida para todos los diseños).
 - La probabilidad (odds) del evento es 1.8 veces mayor en el grupo expuesto; el IC excluye el valor nulo.
 - El resultado no es estadísticamente significativo porque el valor p es mayor que 0.001.
 - El IC 95% confirma que el efecto es clínicamente relevante en cualquier contexto.
 - La IA puede calcular el RR directamente a partir del OR sin información adicional.
- 4. ¿Cuál de las siguientes representa la limitación más crítica de la IA para la evaluación de la aplicabilidad clínica?
 - La IA no puede acceder al texto completo de artículos de pago.
 - La IA no tiene acceso a la historia clínica del paciente ni al contexto asistencial local del lector.
 - La IA no puede interpretar curvas de Kaplan-Meier.
 - La IA no puede comparar resultados de diferentes estudios.
 - La IA comete errores al interpretar análisis de subgrupos.
- 5. Sobre el AUC-ROC en pruebas diagnósticas, ¿cuál es la afirmación correcta?
 - Un AUC de 0.72 indica que la prueba tiene una sensibilidad del 72%.
 - El AUC mide la especificidad de la prueba en poblaciones de alta prevalencia.
 - Un AUC de 0.72 indica que el modelo discrimina correctamente el 72% de los pares positivo-negativo.
 - El AUC no se ve afectado por la prevalencia de la enfermedad en la población.
 - Un AUC > 0.5 siempre indica que la prueba es clínicamente útil.

PREGUNTAS ABIERTAS

6. Caso de análisis: Un LLM analiza un ECA de un hipoglucemiante y reporta: “El estudio demuestra que el fármaco reduce la HbA1c en 1.2% ($p<0.001$) sin eventos adversos cardiovasculares significativos.” Al leer el artículo, usted encuentra que los eventos cardiovasculares no fueron un desenlace primario del estudio, sino un hallazgo secundario basado en registros administrativos. ¿Qué sesgos metodológicos y limitaciones del LLM identifica en esta respuesta?

7. Diseño de prompt: Usted tiene frente a usted una revisión sistemática con metaanálisis sobre betabloqueadores en insuficiencia cardíaca con función preservada. Diseñe un prompt estructurado en tres preguntas secuenciales que le permita analizar: (a) la validez interna del metaanálisis, (b) la heterogeneidad de los estudios incluidos, y (c) la aplicabilidad a su paciente.

8. Análisis de riesgo: Describa el concepto de “dependencia cognitiva” en el contexto del uso de IA por parte de residentes de Medicina. ¿Qué estrategias pedagógicas propondría para mitigar este riesgo sin renunciar a los beneficios de la IA en la formación?

9. Interpretación estadística: Un ensayo clínico aleatorizado reporta HR = 0.78 (IC 95%: 0.65-0.94, $p=0.008$) para mortalidad por todas las causas. La mediana de seguimiento fue 3.2 años. Explique el significado clínico y estadístico de este resultado, señalando qué información adicional necesitaría para calcular el NNT y evaluar la aplicabilidad a su paciente.

10. Integración ética: Un colega argumenta que, dado que la IA comete errores, su uso en la práctica clínica debería estar prohibido hasta que existan regulaciones formales. Elabore una respuesta argumentada que equilibre los riesgos y beneficios documentados, integrando consideraciones éticas y de responsabilidad profesional.

RESPUESTAS

1. b) Alucinación del modelo de lenguaje. La discrepancia entre el NNT reportado por la IA (8) y el del artículo original (14) es un ejemplo clásico de alucinación: generación de datos numéricos incorrectos presentados con igual confianza que los correctos. Este fenómeno no es un error de sesgo de confirmación (a), que implicaría una distorsión motivada por la hipótesis del usuario, ni un error de ventana de contexto (c), que referiría a truncamiento del texto procesado. El sesgo de complacencia (d) describe la tendencia a validar premisas del usuario, no a inventar

cifras.

2. b) La IA reconoce patrones lingüísticos. Los LLMs no tienen acceso al protocolo original del estudio ni pueden auditar si el proceso descrito fue ejecutado. Reconocen terminología metodológica y la aplican, pero sin comprensión verificable del proceso real. La herramienta Cochrane Risk of Bias (c) requiere juicio experto sobre múltiples dimensiones que la IA no puede evaluar con precisión equivalente.

3. Respuesta correcta: b) La probabilidad (odds) del evento es 1.8 veces mayor; el IC excluye el valor nulo.

En estudios de casos y controles, el OR es la medida de asociación apropiada. El IC 95% de 1.2-2.7 excluye el 1.0 (valor nulo), confirmando significancia estadística. No es directamente equivalente al RR (a), que requeriría conocer las frecuencias absolutas. El IC 95% no informa por sí solo sobre relevancia clínica (d).

4. b) La IA no tiene acceso a la historia clínica ni al contexto local.

La evaluación de aplicabilidad integra características del paciente individual, preferencias, contexto asistencial y recursos locales dimensiones inaccesibles para cualquier LLM. Las otras opciones describen limitaciones reales pero de menor peso para la evaluación de aplicabilidad.

5. c) El modelo discrimina correctamente el 72% de los pares positivo-negativo.

El AUC-ROC representa la probabilidad de que el modelo asigne una puntuación mayor a un caso positivo que a uno negativo elegido al azar. Un AUC de 0.72 no es equivalente a la sensibilidad (a), ni a la especificidad (b). Aunque el AUC no es directamente afectado por la prevalencia, la utilidad clínica de la prueba sí lo es (d y e).

6. El LLM sobreinterpreta un desenlace secundario como si fuera evidencia sólida de seguridad cardiovascular. Hay riesgo de sesgo de información por uso de registros administrativos, falta de poder estadístico para eventos cardiovasculares y confusión entre desenlace primario y secundario. La conclusión correcta sería: el fármaco redujo HbA1c, pero el estudio no permite afirmar seguridad cardiovascular definitiva.

7. Prompt sugerido: "Analiza la validez interna del metaanálisis: pregunta PICO, búsqueda, selección de estudios, riesgo de sesgo y métodos estadísticos."

"Evalúa la heterogeneidad clínica, metodológica y estadística: poblaciones, intervenciones, desenlaces, seguimiento e I²."

"Analiza la aplicabilidad a mi paciente considerando edad, comorbilidades, riesgo basal, tratamiento disponible, efectos adversos y preferencias."

8. La dependencia cognitiva ocurre cuando el residente acepta respuestas de IA sin razonamiento propio. Para reducirla, debe exigirse análisis humano previo, comparación con la IA, verificación en fuentes originales y discusión supervisada en sesiones clínicas. La IA debe funcionar como apoyo, no como sustituto del pensamiento médico.

9. HR 0.78 indica una reducción relativa del 22% en el riesgo instantáneo de muerte durante el seguimiento. El IC95% 0.65–0.94 no cruza 1 y p=0.008 indica significancia estadística. Para calcular NNT se necesita el riesgo absoluto en control e intervención, o la reducción absoluta del riesgo a un tiempo definido.

10. No debería prohibirse la IA, sino usarse con regulación, supervisión y verificación. Puede mejorar eficiencia y educación, pero también generar errores, sesgos y alucinaciones. Éticamente, debe respetar beneficencia, no maleficencia, autonomía, justicia y confidencialidad. La decisión final y la responsabilidad siguen siendo del médico.

REFERENCIAS

- Gartlehner G, Kugley S, Crotty K, Viswanathan M, Dobrescu A, Nussbaumer-Streit B, et al. Artificial intelligence-assisted data extraction with a large language model: A study within reviews. *Ann Intern Med*. 2025 Nov 4. Available at: <https://doi.org/10.7326/ANNALS-25-00739>.
- Peatman J, Lang S, Hardy E. Performance of large language model clinical data extraction by data domain: A rapid systematic review. *Value Health*. 2025;28(Suppl 2):SA75. Available at: <https://doi.org/10.1016/j.jval.2025.05.075>.
- Martin A, Roussi K, Rice H, Bertuzzi A. Evaluation of the use of ChatGPT in data extraction for systematic literature reviews: Friend, foe, or fiction? *Value Health*. 2025;28(Suppl 2):MSR100. Presented at: ISPOR Europe 2025; 2025 Nov; Glasgow, Scotland.
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–40. Available at: <https://doi.org/10.1038/s41591-023-02448-8>.
- Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6(1):195. Available at: <https://doi.org/10.1038/s41746-023-00939-z>.
- Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv*. 2023. Available at: <https://doi.org/10.48550/arXiv.2303.13375>.
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–80. Available at: <https://doi.org/10.1038/s41586-023-06291-2>.
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. Available at: <https://doi.org/10.1371/journal.pdig.0000198>.
- Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–8. Available at: <https://doi.org/10.7326/M18-1376>.
- Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928. Available at: <https://doi.org/10.1136/bmj.d5928>.
- Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):1–38. Available at: <https://doi.org/10.1145/3571730>.
- Guo Z, Schlichtkrull M, Vlachos A. A survey on automated fact-checking. *Trans Assoc Comput Linguist*. 2022;10:178–206. Available at: https://doi.org/10.1162/tacl_a_00454.
- Saraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from an artificial intelligence chatbot. *JAMA*. 2023;329(22):1985–7. Available at: <https://doi.org/10.1001/jama.2023.6969>.
- Harrer S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*. 2023;90:104512. Available at: <https://doi.org/10.1016/j.ebiom.2023.104512>.
- Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2):e230163. Available at: <https://doi.org/10.1148/radiol.230163>.
- Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucinations in scientific writing through ChatGPT references. *Cureus*. 2023;15(4):e37432. Available at: <https://doi.org/10.7759/cureus.37432>.
- Cahan P, Bhatt DL, Bhatt NM, Bhatt DL. Physicians and artificial intelligence. *JAMA*. 2023;330(6):509–10. Available at: <https://doi.org/10.1001/jama.2023.13030>.
- Topol EJ. Preparing the healthcare workforce to deliver the digital future. *Lancet*. 2019;394(10197):563–76. Available at: [https://doi.org/10.1016/S0140-6736\(19\)31804-4](https://doi.org/10.1016/S0140-6736(19)31804-4).
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924–6. Available at: <https://doi.org/10.1136/bmj.39489.470347.AD>.
- Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;6(7):e1000097. Available at: <https://doi.org/10.1371/journal.pmed.1000097>.
- Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332. Available at: <https://doi.org/10.1136/bmj.c332>.
- von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet*. 2007;370(9596):1453–7. Available at: [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X).
- Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. Available at: <https://doi.org/10.1038/s41746-023-00873-0>.
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021. Available at: <https://doi.org/10.1145/3442188.3445922>.
- Paranjape K, Schinkel M, Hammer RD, Schouten L, van Weert HC. Introducing AI to the clinical laboratory: Impact on the diagnosis and management of patients with acute chest pain. *Clin Chem*. 2019;65(8):1002–11.

Available at: <https://doi.org/10.1373/clinchem.2019.302364>.

Strömgren M, Gundersen T. Using AI-based large language models in medical education: A systematic review and meta-analysis. *Med Educ*. 2024;58(3):220–34. Available at: <https://doi.org/10.1111/medu.15289>.

World Health Organization. Ethics and governance of artificial intelligence for health. WHO guidance. Geneva: World Health Organization; 2021. Available at: <https://www.who.int/publications/i/item/9789240029200>.

Lujan HL, DiCarlo SE. The paradox of knowledge: Why medical students know more but understand less. *Med Sci Educ*. 2025;35(3):1761–6. Available at: <https://doi.org/10.1007/s40670-025-02379-8>.

Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233–9. Available at: <https://doi.org/10.1056/NEJMSr2214184>.