LLMs en Educación Médica: Impacto en la Integración de Saberes y el Razonamiento Clínico - Una Revisión Sistemática

ANDRÉS FELIPE BORRERO GONZÁLEZ

MAESTRÍA EN EDUCACIÓN

JORGE ALBERTO QUESADA HURTADO, Mg

MAESTRÍA EN EDUCACIÓN

ESCUELA DE EDUCACIÓN, CREACIÓN Y CULTURA

FACULTAD DE CIENCIAS HUMANAS

UNIVERSIDAD ICESI

28 DE MAYO DE 2025

TABLA DE CONTENIDO

ABSTRACT

Introducción: La efectiva integración de conocimientos de ciencias básicas y clínicas, junto con el desarrollo robusto del razonamiento clínico, son pilares fundamentales en la formación médica. Sin embargo, lograr esta sinergia representa un desafío pedagógico constante. Los Modelos de Lenguaje a Gran Escala (LLMs) han emergido como herramientas con potencial para transformar las estrategias educativas. Esta revisión sistemática tuvo como objetivo evaluar el impacto de las experiencias educativas asistidas por LLMs en la integración conceptual interdisciplinar, el desarrollo y la transferencia del razonamiento clínico en estudiantes de medicina, así como caracterizar las percepciones de estudiantes y docentes sobre su uso.

Metodología: Se realizó una revisión sistemática de la literatura. Se realizaron búsquedas en bases de datos electrónicas relevantes para identificar estudios publicados que evaluaran intervenciones educativas asistidas por LLMs dirigidas a estudiantes de medicina y enfocadas en la integración de conocimientos, el razonamiento clínico y las percepciones sobre su uso. Se incluyeron estudios con diversos diseños metodológicos. Los datos fueron extraídos de forma estandarizada y se realizó una síntesis narrativa descriptiva de los hallazgos.

Resultados: De 1442 registros iniciales, se incluyeron 17 estudios que cumplieron con los criterios de elegibilidad. Los hallazgos sugieren que las intervenciones asistidas por LLMs pueden tener un impacto positivo en el desarrollo de competencias de razonamiento clínico, incluyendo mejoras en la precisión diagnóstica y la eficiencia en la resolución de casos, aunque la evidencia sobre los procesos cognitivos específicos y la metacognición es aún incipiente. Se identificó que los LLMs pueden facilitar la integración conceptual entre ciencias básicas y clínicas, principalmente mediante el diseño de prompts que sirvan de andamiaje en la construcción de puentes cognitivos y esquemas integrados. La transferencia del razonamiento clínico desarrollado con LLMs a la práctica real es un área prometedora, con percepciones positivas de confianza y preparación por parte de los estudiantes, aunque los indicadores objetivos de desempeño en entornos reales son limitados. Las percepciones de estudiantes y docentes sobre los LLMs son mayoritariamente positivas, destacando beneficios en eficiencia, motivación y retroalimentación, pero también se identificaron limitaciones técnicas y pedagógicas significativas, barreras de infraestructura y formación, junto con preocupaciones éticas relevantes como el plagio, los sesgos algorítmicos y la equidad en el acceso.

Conclusiones: Las experiencias educativas asistidas por LLMs muestran un potencial considerable para enriquecer la formación médica, particularmente en la mejora del razonamiento clínico y la facilitación de la integración conceptual interdisciplinar. Sin embargo, su implementación efectiva requiere un diseño pedagógico cuidadoso, la superación de barreras tecnológicas y formativas, y una atención continua a las consideraciones éticas. Se necesita más investigación rigurosa para consolidar la evidencia sobre la transferencia a la práctica y los efectos a largo plazo.

INTRODUCCIÓN

La integración de la Inteligencia Artificial (IA), en particular de Modelos de Lenguaje a Gran Escala (Large Language Models, LLMs), en la educación médica está impulsando transformaciones significativas en las estrategias de enseñanza y aprendizaje. Este avance es especialmente relevante al abordar la compleja tarea de armonizar e integrar las ciencias biomédicas básicas con las ciencias clínicas, un pilar fundamental de la formación médica. Históricamente, el desarrollo del razonamiento clínico, componente crítico de esta formación, ha enfrentado brechas y dificultades tanto en su conceptualización como en su aplicación práctica. La proliferación de tecnologías de IA y LLMs abre un panorama prometedor para superar estos desafíos, pero también exige un análisis riguroso y sistemático de su impacto, implementación y evaluación en contextos educativos de alta complejidad.

La creciente presencia de los LLMs en el ámbito educativo y, concretamente, en la formación de profesionales de la salud, subraya la urgencia de comprender a fondo cómo estas herramientas pueden facilitar el aprendizaje, la integración efectiva de saberes básicos y clínicos, y el desarrollo de competencias esenciales en razonamiento clínico. Ante esta necesidad, **la presente tesis se embarca en una revisión sistemática de la literatura** con el objetivo de consolidar la evidencia actual sobre estas cuestiones. A medida que la bibliografía disponible se expande y presenta hallazgos diversos, a menudo con metodologías y enfoques variados, resulta imprescindible un estudio como el presente para articular los aportes de distintas experiencias, evitando una lectura fragmentada o reduccionista de la información (Lucas et al., 2024) y proporcionando una base sólida para la toma de decisiones informadas (Lawson McLean, 2024).

Este estudio, al abarcar una variedad de diseños metodológicos, enfoques pedagógicos y contextos institucionales reportados en la literatura, busca ofrecer un panorama organizado y articulado de la investigación actual. Para guiar este esfuerzo y asegurar que los datos se interpreten de forma consistente, se ha construido un marco teórico robusto (detallado en el capítulo correspondiente) que se fundamenta en principios teóricos bien establecidos en la educación médica y las ciencias cognitivas. Asimismo, la existencia de desafíos prácticos y éticos inherentes a la adopción de LLMs en la educación resalta la importancia de esta investigación para dar sentido a dichos hallazgos y promover recomendaciones fundamentadas (Yan et al., 2023).

La literatura reciente destaca el considerable potencial que los LLMs pueden desempeñar en la mejora de procesos educativos, la creación de contenidos de aprendizaje adaptativos y la optimización del razonamiento clínico, consolidando su pertinencia en el presente y futuro de la formación médica (Omiye et al., 2024). Sin embargo, este potencial no puede comprenderse ni aprovecharse plenamente sin un análisis sistemático que examine la evidencia disponible sobre los aspectos cognitivos, pedagógicos y contextuales que influyen en la adopción e impacto de estas herramientas. La complejidad del razonamiento clínico y las dificultades arraigadas en la transición desde las ciencias biomédicas básicas a las ciencias clínicas demandan una

investigación que integre estas dimensiones. En este sentido, esta revisión sistemática busca no solo sintetizar el conocimiento actual, sino también convertirse en un recurso valioso para el debate académico, la formulación de políticas y el diseño curricular, garantizando así su relevancia y aplicabilidad en el panorama educativo contemporáneo (Telenti et al., 2024).

Adicionalmente, se reconoce la imperativa necesidad de adaptar los programas formativos para que los futuros profesionales y docentes estén preparados ante las implicaciones de la IA y los LLMs en la medicina y su enseñanza. Tal como señalan (McCoy et al., 2024), la incorporación de estas herramientas exige una evolución en las competencias requeridas y en el diseño curricular, a fin de desarrollar habilidades específicas, como la ingeniería de prompts, la evaluación crítica de la información generada por IA y la gestión de sus sesgos inherentes, que contribuyan a una adopción efectiva y ética de los LLMs tanto en la educación como en la práctica clínica futura. Este trabajo de tesis, mediante la revisión y síntesis de la evidencia, aspira a informar este proceso de adaptación, contextualizando y delimitando el uso de LLMs en la educación médica, un escenario donde la integración de saberes básicos y clínicos es esencial para formar profesionales competentes, reflexivos y autónomos, capaces de enfrentar los desafíos de la medicina del siglo XXI.

PROBLEMA DE INVESTIGACIÓN

1 Problema de investigación

La educación médica contemporánea enfrenta el desafío constante de asegurar que los futuros profesionales de la salud no solo adquieran un vasto conocimiento en ciencias básicas y clínicas, sino que también desarrollen la capacidad de integrar estos saberes de manera efectiva para un razonamiento clínico competente. Tradicionalmente, existen brechas en la articulación de estos dominios del conocimiento y en el desarrollo óptimo de las habilidades de razonamiento práctico. La reciente y rápida proliferación de Modelos de Lenguaje a Gran Escala (LLMs) ofrece nuevas posibilidades y herramientas que podrían transformar las estrategias pedagógicas. Sin embargo, el impacto real, los beneficios, las limitaciones y las implicaciones éticas y prácticas de la incorporación de los LLMs en este contexto específico de la educación médica aún no están completamente comprendidos ni sistemáticamente evaluados. Existe una necesidad apremiante de analizar y categorizar el estado actual del conocimiento sobre cómo estas tecnologías están siendo utilizadas y qué efectos tienen en la integración de saberes y en el desarrollo del razonamiento clínico, para guiar su adopción informada y efectiva.

2 Predunta de investigación

¿De qué manera la incorporación de experiencias educativas asistidas por LLMs, en comparación con métodos pedagógicos tradicionales, impacta en la integración conceptual interdisciplinar entre ciencias básicas y clínicas, así como en el desarrollo y en la transferencia práctica del razonamiento clínico, en estudiantes de medicina?

3 Objetivos

3.1 Objetivo general

Analizar y sintetizar la evidencia científica disponible sobre el impacto de la incorporación de experiencias educativas asistidas por Modelos de Lenguaje a Gran Escala (LLMs) en la integración conceptual interdisciplinar entre ciencias básicas y clínicas, el desarrollo y la transferencia práctica del razonamiento clínico en estudiantes de medicina, así como en las percepciones de estudiantes y docentes sobre su uso.

3.2 Objetivos específicos

- Cuantificar el impacto sobre las competencias en razonamiento clínico de los estudiantes, tras haber participado en experiencias educativas asistidas con LLMs.
- Analizar la integración conceptual interdisciplinar facilitada por LLMs entre ciencias básicas y clínicas.
- Evaluar la transferencia percibida del razonamiento clínico desarrollado con la asistencia de LLMs a situaciones de práctica real.
- Caracterizar las percepciones de los estudiantes y docentes sobre los beneficios, limitaciones y barreras asociadas al uso de LLMs como herramienta educativa en la formación médica.

4 Justificación

La presente revisión sistemática se justifica por la creciente relevancia y rápida implementación de los Modelos de Lenguaje a Gran Escala (LLMs) en múltiples sectores, incluyendo la educación superior y, específicamente, la formación médica. Existe un interés significativo en el potencial de estas tecnologías para abordar desafíos persistentes en la educación médica, tales como la efectiva integración de los conocimientos de ciencias básicas con las aplicaciones clínicas y el desarrollo de un razonamiento clínico sólido y eficiente en los estudiantes.

A medida que la literatura científica sobre el uso de LLMs en este campo se expande, se hace necesario recopilar, evaluar críticamente y sintetizar los hallazgos de los estudios

primarios para obtener una comprensión clara del estado actual del conocimiento. Esta síntesis es crucial para identificar no solo los beneficios potenciales y las áreas de impacto positivo, sino también las limitaciones, los desafíos (pedagógicos, técnicos, éticos) y las barreras para la implementación efectiva de estas herramientas.

Los resultados de esta revisión sistemática podrán ofrecer una base de evidencia que oriente a educadores médicos, diseñadores curriculares, desarrolladores de tecnología educativa y responsables de políticas institucionales en la toma de decisiones informadas sobre la incorporación de LLMs en los programas de formación médica. Además, al identificar lagunas en la investigación actual, este estudio contribuirá a dirigir futuras líneas de investigación que exploren de manera más profunda y rigurosa el rol y el valor de los LLMs en la preparación de los futuros profesionales de la salud. En última instancia, una comprensión más matizada del impacto de los LLMs permitirá aprovechar sus ventajas mientras se mitigan sus riesgos, con el fin de mejorar la calidad y efectividad de la educación médica.

MARCO TEÓRICO

1 El razonamiento clínico en la educación médica

El razonamiento clínico implica la capacidad de analizar información compleja, identificar patrones relevantes y adoptar decisiones informadas frente a la incertidumbre y la variabilidad de las presentaciones clínicas de las enfermedades (Guzmán-Valdivia-Gómez et al., 2022; Kulasegaram et al., 2015). A diferencia de la mera acumulación de datos, el razonamiento clínico requiere comprender cómo se entrelazan los conocimientos biomédicos básicos (bioquímica, fisiología, anatomía, histología, microbiología, etc.) con las características epidemiológicas y semiológicas de pacientes individuales y todo esto, a su vez, con el contexto psicosocial en que se desenvuelve su situación de salud. Sin embargo, este proceso no es innato, ni surge de manera espontánea. Exige estrategias pedagógicas claras y un abordaje teórico-práctico que oriente su desarrollo a lo largo de la carrera médica (Parodis et al., 2021).

1.1 Definiciones y modelos del razonamiento clínico: enfoques cognitivos, metacognitivos y contextuales

El razonamiento clínico puede abordarse desde distintos marcos teóricos. Desde la perspectiva de la teoría del sistema dual del pensamiento de Kahneman (2013), por ejemplo, el razonamiento clínico puede requerir tanto procesos cognitivos conscientes y deliberados (pensamiento inductivo-reflexivo o sistema 2), como acciones intuitivas ancladas en el reconocimiento de patrones (razonamiento intuitivo o sistema 1). Bajo esta perspectiva dual, el clínico experto no solo recuerda datos, sino que construye significados y relaciones entre ellos, utilizando la experiencia previa para interpretar

señales clínicas de forma integral (Guzmán-Valdivia-Gómez et al., 2022). Además, la comprensión profunda del razonamiento clínico implica asumir una dimensión metacognitiva, es decir, la capacidad del estudiante para reflexionar sobre su propio proceso de pensamiento, identificar sesgos, cuestionar sus suposiciones y ajustar sus estrategias a medida que surgen nuevos datos o se modifica el contexto (Ambrose et al., 2017; Pears & Konstantinidis, 2024).

En el ámbito educativo, la taxonomía de Bloom, aplicada al razonamiento clínico, sugiere que el estudiante debe avanzar más allá de la simple memorización para llegar a niveles superiores de análisis, evaluación y creación, lo que requiere una interacción constante entre teoría y práctica (Ambrose et al., 2017; Pears & Konstantinidis, 2024). Así, la construcción del razonamiento clínico no se limita a la adquisición de información, sino que exige un andamiaje pedagógico que motive la comparación de casos, la aplicación contextualizada de conceptos científicos y el ejercicio sistemático de la reflexión sobre las decisiones tomadas. Estas dinámicas se vuelven indispensables para forjar un pensamiento clínico flexible, adaptativo y contextual, capaz de reconocer las limitaciones inherentes a la práctica médica real (Delavari et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Schwartzstein, 2024).

1.2 Adquisición y desarrollo progresivo del razonamiento clínico en el estudiante de medicina

El razonamiento clínico no se adquiere de forma inmediata, sino que se desarrolla a lo largo del tiempo, a medida que el estudiante enfrenta progresivamente casos más complejos y desafiantes. Inicialmente, el aprendiz se apoya principalmente en un razonamiento más analítico y estructurado (pensamiento inductivo-reflexivo), utilizando pautas y algoritmos que le permitan orientarse en un mar de datos a menudo desorganizados. Con la experiencia, las exposiciones repetidas a situaciones clínicas diversas y la reflexión guiada por docentes y sus propias vivencias, el estudiante integra gradualmente estos algoritmos en su acervo mental (pensamiento intuitivo), empezando a reconocer patrones, anticipar desenlaces y movilizar el conocimiento con mayor fluidez (Delavari et al., 2024; Guzmán-Valdivia-Gómez et al., 2022).

El proceso no es lineal ni uniforme, pues depende del entorno pedagógico, la variedad de experiencias clínicas, la calidad de la retroalimentación y las oportunidades de práctica deliberada (Durning et al., 2024; Kassirer, 2010; Parodis et al., 2021). Intervenciones educativas que combinan la exposición a casos clínicos realistas, la discusión estructurada de las decisiones, y la incorporación de herramientas complementarias (como recursos tecnológicos diversos, incluyendo a la IA) pueden acelerar esta maduración cognitiva(Choi & Abdirayimov, 2024; Dong et al., 2024; Lucas et al., 2024). Esta evolución paulatina del razonamiento, que transita desde la rigidez de la teoría hacia la flexibilidad del juicio clínico experto, es uno de los objetivos centrales de una formación médica de calidad.

1.3 Factores que influyen en la calidad del razonamiento: sesgos cognitivos, experiencia, retroalimentación y práctica guiada

La calidad del razonamiento clínico está determinada por múltiples factores. Por un lado, intervienen la experiencia acumulada, el bagaje conceptual y la capacidad de pensar de manera analítica e intuitiva a la vez. Por otro lado, factores del entorno académico y social, las emociones, la presión del tiempo, el tipo de pacientes atendidos y las condiciones institucionales de los escenarios de práctica, también influyen extensamente en la forma en que el estudiante procesa la información. Además, existen diversos sesgos cognitivos que pueden distorsionar el juicio y conducir a errores diagnósticos o terapéuticos, como el anclaje prematuro, la confirmación selectiva o la tendencia a sobredimensionar hipótesis familiares, entre otros (Ambrose et al., 2017; Guzmán-Valdivia-Gómez et al., 2022; Kulasegaram et al., 2015).

La retroalimentación oportuna, específica y constructiva, así como la práctica guiada en entornos controlados, resultan esenciales para mitigar estos sesgos y mejorar la calidad de los procesos cognitivos que enriquecen el razonamiento clínico (Ambrose et al., 2017; Durning et al., 2024). La exposición sistemática a casos clínicos bien diseñados, el uso de métodos que estimulen la metacognición y el acompañamiento de mentores o tutores experimentados, tanto en aspectos clínicos como pedagógicos, permiten al estudiante reconocer sus errores, ajustar su estrategia y refinar su juicio clínico. Asimismo, es fundamental que el entorno educativo promueva la reflexión sobre el error sin penalizaciones desproporcionadas, pues ello favorece el aprendizaje profundo y significativo. Asimismo, comprender que el razonamiento clínico no es un mero atributo cognitivo, sino un proceso sensible al contexto ayuda a diseñar estrategias más efectivas para su enseñanza (Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Schwartzstein, 2024).

En este marco, las herramientas de IA y, en particular, los LLMs, pueden complementar las estrategias didácticas tradicionales. Su rol potencial no se limita a ofrecer información, sino a simular situaciones complejas, adaptar la dificultad de los casos, señalar inconsistencias en el razonamiento y proporcionar apoyo personalizado. Sin embargo, su utilidad dependerá de la integración coherente con un diseño educativo sólido y un acompañamiento humano que garantice una formación balanceada, ética y orientada a fortalecer las competencias clínicas de los futuros médicos(Choi & Abdirayimov, 2024; Dong et al., 2024; Lucas et al., 2024; Pears & Konstantinidis, 2024; Yan et al., 2023). Adicionalmente, las iniciativas de formación docente enfocadas en la comprensión e integración de LLMs pueden potenciar su adopción segura y efectiva. Por ejemplo, se ha propuesto incluir competencias específicas en IA y razonamiento automatizado en los programas formativos, de modo que estudiantes y docentes manejen herramientas básicas de ingeniería de prompts y evaluación crítica de las respuestas generadas por estas tecnologías (McCoy et al., 2024). De esta manera, se promueve una apropiada

combinación de recursos digitales y enfoques pedagógicos tradicionales, ampliando el margen de beneficio en la construcción del razonamiento clínico.

2 Desafíos en la educación médica y la integración de saberes

Con todo lo antes mencionado, la formación de un médico competente exige que el estudiante asimile una amplia diversidad de conocimientos a lo largo de su trayectoria académica. Durante las etapas tempranas de la formación, las ciencias biomédicas básicas suelen enseñarse de forma parcelada, por asignaturas, dividiéndose en áreas como la anatomía, la fisiología, la bioquímica y otras disciplinas orientadas a la comprensión de los fundamentos biológicos del organismo. Conforme avanza la carrera, el alumno se introduce en las ciencias clínicas, también habitualmente segmentadas en áreas o especialidades individuales, en las que debe aplicar conceptos previamente aprendidos para analizar problemas complejos, vincular síntomas y manifestaciones físicas entre sí y con hallazgos de laboratorio, considerar el contexto social y emocional del paciente y, finalmente, ejercer un juicio clínico fundamentado, para tomar decisiones respecto a los diagnósticos y problemas de un paciente y definir las conductas a seguir.

No obstante, esta transición desde las ciencias básicas hacia las clínicas rara vez es fluida. Por lo general, tanto las ciencias biomédicas básicas como las clínicas se enseñan de manera fragmentada, dificultando la integración entre sus contenidos. Esta falta de coherencia interna, sumada a la dificultad para conectar más tarde los saberes básicos con las exigencias clínicas, limita la capacidad del estudiante para movilizar y aplicar el conocimiento de forma integral. La literatura identifica esta ausencia de integración como uno de los desafíos más persistentes de la educación médica contemporánea (Kulasegaram et al., 2015), un problema que impacta negativamente el razonamiento clínico dificulta la aplicación práctica de la teoría y, en última instancia, puede comprometer la calidad de la atención al paciente, ya durante el ejercicio profesional médico (Delavari et al., 2024; Kassirer, 2010; Parodis et al., 2021).

2.1 Transición de las ciencias básicas a las ciencias clínicas: dificultades y barreras

La brecha entre las ciencias biomédicas básicas y las ciencias clínicas se ha descrito como un problema estructural: por un lado, las ciencias básicas suelen presentarse en contextos aislados, muchas veces incurriendo en exigencias y estrategias que demandan la memorización de datos y la comprensión de procesos fisiológicos o patológicos de forma abstracta, con una limitada representación y aplicación práctica. Por otro lado, las ciencias clínicas también tienden a enfatizar la memorización de criterios diagnósticos, clasificaciones de severidad, guías, protocolos, dosis, interacciones farmacológicas y

efectos adversos, sin fomentar la construcción de puentes conceptuales entre distintas especialidades ni con los fundamentos biomédicos que el estudiante, a estas alturas, con frecuencia ya ha olvidado.

Esta desconexión conceptual y teórico-práctica perjudica la verdadera comprensión profunda y apropiación de saberes, el desarrollo efectivo de habilidades y competencias, y la satisfacción y motivación del alumno. Sin embargo, y para complejizar aún más esta problemática, a ello se suma que muchos profesionales médicos, que son a la vez docentes, aunque a veces sin una formación pedagógica adecuada, señalan que los estudiantes no aplican esos conocimientos básicos de forma suficiente, a la vez que se desempeñan de forma insatisfactoria en las actividades teóricas y prácticas clínicas, afectándose así la relación entre docentes y alumnos y la motivación del profesorado para buscar nuevas y mejores estrategias, para las actividades formativas que dirigen(Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kulasegaram et al., 2015).

Aunado a esto, en ambos momentos de la formación, las asignaturas suelen evaluarse con cuestionarios de opción múltiple o exámenes netamente teóricos o solamente prácticos, que no necesariamente exigen al estudiante desarrollar un razonamiento crítico y argumental profundo e integrador. En consecuencia, desde el inicio se promueve y perpetúa una visión segmentada del conocimiento, que luego es difícil reconfigurar en las etapas finales de la formación, cuando más se requiere un abordaje holístico del paciente. Esta falta de entrenamiento para vincular conceptos científicos con situaciones reales se reconoce como uno de los principales retos a superar para mejorar la formación médica integral (Delavari et al., 2024; Kassirer, 2010; Parodis et al., 2021).

2.2 Necesidad de una integración conceptual sólida en la formación del médico

La superación de estas barreras no pasa únicamente por “sumar” más contenido al currículo, sino por introducir estrategias pedagógicas que fomenten una integración conceptual sólida. Una formación verdaderamente efectiva debe evitar que las ciencias básicas y las ciencias clínicas operen de forma desarticulada, y en su lugar, debe alentar a los estudiantes a establecer puentes cognitivos entre ambas. Para ello, se propone un enfoque más dinámico y constructivista, en el que los alumnos se involucren en tareas que requieran aplicar la base científica a la resolución de problemas clínicos, promoviendo así la construcción de un andamiaje cognitivo que fortalezca su razonamiento (Delavari et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kassirer, 2010; Parodis et al., 2021). De igual manera, fomentar la metacognición y la autorreflexión resulta determinante para que el estudiante comprenda cómo integrar las ciencias básicas en la práctica clínica. Una vez que el alumno identifica sus propias dificultades y sesgos, puede modular sus estrategias de estudio y razonamiento, consolidando una

visión más cohesionada del conocimiento biomédico y sus aplicaciones (Ambrose et al., 2017).

Diversos autores han sugerido el uso de metodologías activas, como el aprendizaje basado en problemas o el uso de casos clínicos progresivos, para favorecer esta integración. Estas estrategias exigen del estudiante no solo recordar información, sino también relacionarla, evaluarla críticamente y adaptarla a contextos cambiantes. Además, resulta crucial incorporar actividades que estimulen la reflexión metacognitiva, ayudando a los alumnos a identificar sus propias limitaciones, reconocer sesgos cognitivos e implementar mejoras continuas. Las revisiones sobre la enseñanza del razonamiento clínico señalan que la integración conceptual no solo permite un mejor desempeño en exámenes y evaluaciones, sino que también contribuye a desarrollar una competencia más duradera y significativa para el ejercicio profesional (Delavari et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kassirer, 2010; Kulasegaram et al., 2015; Parodis et al., 2021).

2.3 Brechas formativas identificadas en la literatura: razonamiento clínico y su relación con el rendimiento profesional

El razonamiento clínico es, en concordancia con lo anterior, una habilidad compleja que va más allá de la mera aplicación mecánica del conocimiento. Involucra un proceso iterativo en el cual el médico genera hipótesis diagnósticas, las contrasta con la información disponible, reevalúa sus supuestos y ajusta su estrategia, siempre bajo condiciones de incertidumbre, obligándole a mantenerse abierto a rehacer sus interpretaciones de forma constante(Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kulasegaram et al., 2015). Las dificultades de nuestros sistemas formativos para impulsar el desarrollo de esta competencia se traducen en que los egresados, especialmente poco tiempo tras finalizar su formación, a menudo enfrentan dificultades frecuentes para tomar decisiones informadas en escenarios reales, incluso independientemente de su complejidad (Delavari et al., 2024; Kassirer, 2010; Parodis et al., 2021).

En muchos casos, la ausencia de un entrenamiento específico en razonamiento clínico lleva a los estudiantes a caer en errores cognitivos comunes, como el anclaje prematuro en una hipótesis errónea, a menudo impulsada por un sesgo de confirmación de la primera suposición, lo que reduce su capacidad de explorar otras posibilidades diagnósticas y, con ello, errar también en su plan de manejo. Es decir que, a raíz de estas limitaciones, que empiezan a percibirse desde las primeras etapas de la formación médica, no solo se afecta el desempeño académico del estudiante, sino que se compromete la calidad de la atención brindada al paciente, tanto durante los escenarios formativos, como ya en la práctica médica del profesional graduado. Por ende, la habilidad para razonar clínicamente no debería nunca verse como una habilidad que se desarrolla de forma pasiva, sino que es uno de los elementos centrales del

profesionalismo y la eficacia médica y que debe fomentarse de forma explícita en cualquier programa de formación de profesionales sanitarios (Delavari et al., 2024; Kassirer, 2010; Parodis et al., 2021).

La literatura propone múltiples estrategias para mejorar el razonamiento clínico en la educación médica. Una de las mejor respaldadas es el Aprendizaje Basado en Casos (ABC), especialmente haciendo uso de casos clínicos de complejidad progresiva, y el uso de simulaciones, cuyo nivel de fidelidad con la realidad también se puede ajustar al contexto específico en que se implementan, que son estrategias en las que el estudiante pone en práctica sus conocimientos y recibe retroalimentación oportuna (Delavari et al., 2024; Durning et al., 2024; Kassirer, 2010). Además, se enfatiza la necesidad de un entrenamiento que haga explícitos los pasos del razonamiento, la identificación de sesgos cognitivos y la práctica deliberada de diferentes situaciones imprevistas que aparezcan en escenarios clínicos diversos. En este sentido, el uso de herramientas de IA, incluyendo LLMs, podría mejorar estos procesos al ofrecer un entorno de práctica segura, retroalimentación continua y adaptación flexible a las necesidades individuales del aprendiz (Choi & Abdirayimov, 2024; Lawson McLean, 2024; Lucas et al., 2024; Suresh & Misra, 2024).

En suma, los desafíos identificados en la transición entre las ciencias básicas y clínicas, la necesidad de una integración conceptual sólida y las brechas en la enseñanza del razonamiento clínico convergen en un punto común: la imperante necesidad de encontrar e implementar estrategias pedagógicas y didácticas innovadoras en la formación médica. Es en este contexto que las herramientas basadas en IA, en especial los LLMs, podrían desempeñar un papel clave, siempre que se integren con un criterio pedagógico adecuado y una base teórica firme que oriente su implementación.

3 Definición, características y evolución de la Inteligencia Artificial y los Modelos de Lenguaje de Gran Escala

Aunque el término Inteligencia Artificial pueda parecer novedoso, en realidad se remonta a la década de 1950, con las primeras investigaciones en este campo, y tuvo su primera fase de rápido desarrollo hasta alrededor de la década de 1970, mientras que en las décadas de 1980 y 1990 se registraron fases de avance más lentas y periodos de relativo estancamiento. Sin embargo, es en los últimos 10 años, y especialmente en los últimos 2 a 3 años, que se han producido los avances tecnológicos más importantes en este campo, tras el desarrollo de los sistemas de redes neuronales profundas y la arquitectura de transformadores, además de fuertes avances en hardware especializado, que han impulsado fuertemente el desarrollo de tecnologías de IA, con los LLMs siendo una de sus implementaciones más notables; ha sido especialmente relevante la introducción de ChatGPT, una herramienta basada en un LLM llamado GPT-3.5, lanzada al público por OpenAI en noviembre de 2022, que demostró a un amplio rango de usuarios no expertos la capacidad de una IA para generar texto coherente, responder preguntas complejas y

simular conversaciones de forma contextualmente relevante. A partir de entonces, la IA ha comenzado a formar parte del vocabulario cotidiano, con aplicaciones tangibles para distintos tipos de IA, especialmente LLMs y otras formas de IA generativas, y en una amplia gama de sectores, entre ellos la educación, incluida la de las ciencias de la salud (Craig et al., 2024; Hashemi-Pour, 2023).

3.1 Fundamentos de la Inteligencia Artificial y del Aprendizaje Automático

La Inteligencia Artificial puede definirse, de manera general y simplificada, como un área de la informática dedicada al desarrollo de sistemas capaces de realizar tareas complejas que, normalmente, requieren de la inteligencia humana y no pueden delegarse enteramente a maquinas o sistemas informáticos menos sofisticados, tales como reconocer patrones, aprender de la experiencia, razonar, resolver problemas y comprender o generar lenguaje. Una de las subáreas más dinámicas de la IA es el aprendizaje automático (en inglés Machine Learning, ML), en la que se diseñan programas llamados modelos, que son conjuntos de algoritmos y ecuaciones matemáticas entrenadas con series de datos, de modo que pueda identificar patrones y luego aplicarlos para hacer predicciones o producir contenido nuevo. Asimismo, existen distintos tipos de modelos de ML, dedicados para desarrollar tareas particulares, con los LLM siendo un ejemplo de modelos de ML (Broshar, 2024; Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

Los LLMs están programados para el Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés), y se enfocan en la “comprensión” y generación de lenguaje, de forma similar como lo hace un humano. Su meta es “aprender” las estructuras, patrones, significados y matices del lenguaje para luego, al recibir una tarea o “entrada” (por ejemplo, una pregunta), generar una respuesta o “salida” textual o hablada coherente, contextualizada y adaptada a la tarea solicitada y para ello se entrenan con cantidades enormes de datos textuales, a menudo cientos de miles de millones de palabras, provenientes de fuentes diversas: libros, artículos científicos, páginas web, repositorios de código, foros, redes sociales, etc. (Broshar, 2024; Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

Sin embargo, es importante destacar que dichas capacidades de los modelos de ML para la “comprensión” y “aprendizaje” y, en última instancia, de “razonamiento”, no son ni funcionan de la misma forma que ocurre en los humanos, sino que emergen de un proceso estadístico: el modelo calcula, en cada paso, la probabilidad con que una palabra puede aparecer antes o después de otra en un mensaje, según esos patrones que ha logrado identificar, a partir de esas enormes cantidades de datos con que fue entrenado, sin poseer una verdadera comprensión del contenido de ese mensaje, pero tampoco a partir de una serie de reglas o definiciones preprogramadas, dado que esto limitaría mucho la versatilidad, adaptabilidad y alcance del modelo; no es una selección

determinista de qué palabra se sigue a otra, sino un proceso netamente probabilístico, basado en correlaciones estadísticas aprendidas del cuerpo de datos con que fue entrenado. Es decir que esto es más semejante, aunque no igual, a ese proceso de pensamiento predictivo-asociativo, rápido, pero a veces impreciso, de lo que se asemeja al proceso de pensamiento inductivo antes descrito, más concienzudo, creativo y capaz de resolver problemas complejos o excepcionales. Al depender sustancialmente de distribuciones de probabilidad basadas en datos de entrenamiento, si el modelo afronta una situación para la cual los patrones estadísticos son limitados o están sesgados, pueden surgir las llamadas “alucinaciones”, es decir, respuestas aparentemente coherentes, pero factualmente incorrectas, uno de los errores más frecuentemente cometidos por los LLMs modernos (Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

Ahora bien, el calificativo “a gran escala” que se le confiere a los LLMs proviene tanto de la magnitud de los datos utilizados en su entrenamiento, como de la complejidad interna del modelo, medida en términos del número de parámetros y pesos, y que suele correlacionarse con la capacidad que tiene para generar respuestas más complejas, sofisticadas y completas, para la tarea asignada (Broshar, 2024; Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

3.2 Arquitectura y elementos clave de los LLMs

Considerando que este tipo de tecnologías sólo se han empezado a popularizar ampliamente hace tan poco tiempo, y pese a lo antes descrito sobre la creciente inclusión de la IA y los LLMs en distintos campos de la vida cotidiana, es prudente abordar acá una serie de conceptos clave sobre el diseño y funcionamiento de este tipo de herramientas, con el objetivo de lograr una perspectiva más clara de la discusión posterior sobre sus usos, beneficios, limitaciones y problemas, en cuanto a su implementación para la educación médica:

i. Redes neuronales artificiales, redes neuronales profundas y el procesamiento de lenguaje natural:

Los LLMs se basan en sistemas de Redes Neuronales Artificiales (ANNs, por sus siglas en inglés), los cuales se inspiran en el funcionamiento del cerebro humano y emulan algunas de sus características, para el procesamiento de datos. En términos muy simples, estas redes se organizan con una serie de capas: la capa de neuronas entrada, que capta los datos proporcionados para el desarrollo de la tarea, una capa de neuronas ocultas, que realiza una serie de cálculos y procesamientos intermedios de esos datos, y una capa de neuronas de salida, que son las que organizan y presentan la respuesta o producto final. Específicamente, los LLMs modernos utilizan lo que se ha llamado Redes Neuronales Profundas (DNNs, por sus siglas en inglés), que son redes múltiples e interconectadas, que poseen varias capas ocultas, que trabajan en paralelo para procesar y transformar la información de entrada en la salida, de modo que se incrementa

y optimiza su capacidad de procesamiento. Cuantas más capas de neuronas ocultas y mayor dimensionalidad tengan estas redes, más complejas son las relaciones funciones que pueden desarrollar, permitiendo así llevar a cabo esas tareas que antes solo podían desarrollarse por humanos (Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

En el caso del lenguaje, estas redes se aplican dentro del campo del NLP, cuyo objetivo es que las máquinas entiendan, analicen y generen lenguaje humano. Los LLMs permitido un avance muy significativo en el NLP, dado que, gracias a su masividad y potencia, han dejado atrás enfoques más simples de procesamiento de lenguaje, que solo se basaban en conteos de palabras o reglas de asociación de palabras mucho más limitadas (Craig et al., 2024; Hashemi-Pour, 2023).

ii. Transformadores y mecanismos de atención:

Un hito crucial en el desarrollo de LLMs fue la invención de la arquitectura de "transformadores" (en inglés, transformers). Un transformador es un tipo de DNN que procesa secuencias de texto de forma paralela, no secuencial, utilizando un componente clave llamado "mecanismo de atención". La atención permite que el modelo evalúe qué partes del texto son más relevantes, en función de su contexto global; dicho de otro modo, mientras lee una frase, el modelo puede "mirar" hacia adelante y hacia atrás en el texto, para entender las relaciones entre las palabras contenidas en ella, con el texto en su totalidad. Esta capacidad para manejar dependencias lejanas resulta esencial, pues el significado de una palabra en un párrafo puede depender de conceptos introducidos muchas frases antes o ampliarse posteriormente. Además, los transformadores emplean "codificaciones posicionales" (en inglés, positional encodings) que le permiten al modelo incorporar y considerar, durante el procesamiento del lenguaje, la información sobre el orden de las palabras en las frases y párrafos, ya que el orden también es fundamental para el significado y finalidad de una oración o mensaje, lo que amplía la capacidad de los LLMs para generar distintos tipos de textos, según las necesidades e instrucciones específicas de cada contexto (Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

iii. Tokens y tokenización:

Para procesar el lenguaje, los LLMs lo separan el mensaje textual o verbal en "tokens", que en este contexto se traduciría como "unidad indivisible" o "elemento básico", y que pueden ser las palabras completas, fragmentos de palabras o caracteres individuales mínimos en los que se puede descomponer el mensaje. El modelo asigna un número a cada token, y aprende representaciones matemáticas (llamadas "embeddings", cuya mejor traducción al español, en este contexto particular, sería "relaciones instruccionales") que capturan su significado y "traducen" el mensaje del lenguaje humano natural a un código que el modelo puede procesar y viceversa. Este proceso se ha denominado "tokenización" (en inglés, tokenisation) (Broshar, 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

iv. Parámetros, pesos, tamaño y rendimiento de los modelos:

Los parámetros de un LLM son los valores numéricos internos que la red ajusta durante el entrenamiento del modelo. Entre ellos, los “pesos” (en inglés, weights) definen la fuerza de la conexión entre las capas de la red y determinan cómo se transforman las representaciones tokenizadas del mensaje, a medida que fluyen a través del modelo. Por lo general, cuantos más parámetros tenga un LLM, mayor su capacidad potencial para captar relaciones complejas en el lenguaje. En general, se consideran pequeños a los modelos con menos de 1x10^9 parámetros* (por ejemplo, TinyLlama, un modelo de código abierto derivado de Llama2, de Meta), medianos a los que tienen entre 1x10^9 y 10x10^9 parámetros (por ejemplo, GPT-2 de OpenAI), mientras que los modelos de mayor escala tienen más de 10x10^9 parámetros poderosos (por ejemplo, GPT-3.5 y GPT-4 de OpenAI, Llama2 y Llama 3 de Meta o Gemini 1.0 Pro y Gemini 1.0 Ultra de Google). Los modelos pequeños, por su parte, aunque no resalten por su capacidad para desempeñar tareas complejas, sobresalen por su eficiencia y capacidad para desarrollar tareas específicas (a veces incluso igualando a modelos más grandes, pero más generalistas) y por la mayor facilidad y menor costo en su desarrollo. La mayoría de los modelos actuales más avanzados entran en la categoría de los modelos grandes, y en la actualidad alcanzan ya cifras en el orden de 10^12 parámetros. A modo de ejemplo, aunque no cuenten con cifras oficiales publicadas, los expertos estiman que GPT-4, uno de los LLM de acceso público de OpenAI más avanzados a la fecha, tienen un tamaño de alrededor de 1.8x10^12 parámetros (Broshar, 2024; Craig et al., 2024; Hashemi-Pour, 2023; Nalpas, 2024; Nerella et al., 2024; Raiaan et al., 2023).

No obstante, debe anotarse que alcanzar una mayor complejidad en los modelos también trae dificultades y es una de las mayores limitantes en el avance de estas tecnologías: implica mayor dificultad para entrenar el modelo, más recursos computacionales, mayor riesgo de errores por sobreajuste y posibles sesgos no detectados en los datos de entrenamiento (Broshar, 2024; Craig et al., 2024; Hashemi-Pour, 2023; Nalpas, 2024; Nerella et al., 2024; Raiaan et al., 2023).

3.3 Entrenamiento, inferencia y tipos de aprendizaje:

Si bien una revisión detallada de este aspecto está fuera del alcance de este trabajo, es importante entender también algunos aspectos básicos sobre el entrenamiento de un

* Con el objetivo de evitar confusiones por las diferencias en las nomenclaturas de los sistemas de numeración corta, frecuentemente usada en países de habla inglesa, y sistemas de numeración larga, más habituales en los países de habla hispana, en este trabajo se expresarán las cifras del número de parámetros en notación exponencial. Sin embargo, la mayoría de los textos que tratan estos temas suelen expresar estas cifras representándolas de forma escrita y usando el sistema de numeración corto (ej. “billions” para 10^9 o “trillions” para 10^12).

LLM, pues resulta bastante útil para luego comprender las fortalezas y limitaciones de este tipo de herramientas.

Como lo describen varios autores, como Broshar (2024), Craig et al. (2024), Hashemi-Pour (2023), Nerella et al. (2024) y Raiaan et al. (2023), entrenar LLMs es una tarea que demanda recursos computacionales enormes, típicamente empleando un gran número de Unidades de Procesamiento Gráfico o Unidades de procesamiento tensorial (GPUs y TPUs, por sus siglas en inglés, respectivamente), durante días o semanas, y el acceso a conjuntos de datos masivos que pueden abarcar desde datos web, libros digitales, artículos científicos, contenidos educativos, repositorios de código y muchas otras fuentes, en buena parte dependiendo del propósito de cada modelo. Además, de la diversidad y calidad de esos datos dependerá la versatilidad y precisión del modelo en diferentes dominios, de modo que algunos modelos pueden tener propósitos más generalistas, como podrían serlo algunos chatbots, y otros ser más especializados, como ocurre con modelos entrenados para desarrollar tareas específicas en algunos campos científicos o educativos.

Así pues, y como lo señalan estos mismos autores, existen dos etapas clave en el "ciclo de vida" de los LLMs:

- Entrenamiento: el modelo "aprende" a partir de grandes conjuntos de datos. Durante este proceso, el modelo se expone a secuencias de texto y se le pide que anticipe el token siguiente. Inicialmente comete errores, tras lo que recibe una corrección que luego le permite ajustar los pesos y parámetros con que operan sus neuronas, mediante algoritmos de retro-propagación y optimización, para reducir la diferencia entre sus predicciones y la respuesta realmente correcta. Tras millones de iteraciones, el LLM desarrolla una representación interna que "asocia" la probabilidad de aparición de un token con el contexto dado y este enfoque estadístico permite al modelo capturar patrones sutiles del lenguaje, consolidando así la probabilidad de predecir la palabra correcta la próxima vez que encuentre un contexto parecido.

  Hay distintos enfoques en el entrenamiento:

  - Aprendizaje supervisado: se le proporcionan al modelo datos de entrada y las respuestas correctas. Por ejemplo, una pregunta particular y esto acompañado de su respuesta. A esa clasificación de los datos según sean o no correctos para una tarea y contexto específico, se le llama etiquetado. El modelo aprende a relacionar las etiquetas y mapear la salida correcta a la entrada adecuada.
  - Aprendizaje no supervisado: el modelo se entrena a partir de datos sin etiquetas explícitas, simplemente analizando grandes cantidades de texto, para identificar patrones y predecir las salidas correctas, como las palabras ocultas o la siguiente palabra en una secuencia. Esto es fundamental en LLMs, ya que la mayor parte de su conocimiento se adquiere de esta forma.

    - Aprendizaje auto supervisado: similar al no supervisado, pero utilizando tareas generadas automáticamente a partir de los mismos datos. Por ejemplo, enmascarar una palabra y pedir al modelo que la prediga. Esta técnica se usa masivamente en LLMs, pues no requiere etiquetado manual.
    - Aprendizaje por pocos disparos (en inglés, “few-shot learning”) o cero disparos (en inglés, “zero-shot learning”): una vez entrenado, el modelo puede adaptar su conocimiento a nuevas tareas con muy pocos ejemplos (pocos disparos) o incluso sin ejemplos (cero disparos) de respuestas correctas, gracias a la riqueza de patrones que ya internalizó durante el entrenamiento a gran escala.

- Inferencia: una vez entrenado, el modelo puede utilizarse para obtener resultados a partir de nuevos datos, sin requerir más ajustes internos en los pesos y parámetros de operabilidad, sino que, con cada nueva entrada, genera la distribución de probabilidad de tokens probables y selecciona aquellos que, según lo aprendido, hacen más sentido en el mensaje y su contexto específico. Por ejemplo, un LLM entrenado en millones de textos podrá, durante la inferencia, responder preguntas o generar resúmenes coherentes.

3.4 Prompts e ingeniería de prompts

Cuando se utiliza un LLM para resolver una tarea específica, se le proporciona una instrucción de entrada llamada “prompt”. Estos pueden ser una pregunta, una indicación u orden o cualquier otro tipo de información, que proporciona un contexto que el modelo debe tomar en cuenta para generar una respuesta acorde. Por lo anterior, resulta fundamental construir cada prompt de modo que oriente al modelo a dar las respuestas que realmente se espera obtener de él, de forma eficiente, efectiva y suficiente: dado que el LLM genera su respuesta en función de la distribución probabilística de tokens que encajan en el contexto, la forma en que se elabore el prompt influye drásticamente en esa distribución. Si este es ambiguo, llevará al modelo a contemplar múltiples rutas de respuesta posibles, “repartiendo” sus probabilidades entre distintas opciones, lo que puede derivar en respuestas vagas, imprecisas o desconectadas. Por el contrario, un prompt bien estructurado disminuye la incertidumbre y canaliza las probabilidades hacia la respuesta que buscamos (Broshar, 2024; Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

Ante esta necesidad, ha surgido la “ingeniería de prompts” (en inglés, prompt engineering), que se refiere a la práctica de diseñar cuidadosamente estas instrucciones para optimizar la calidad, relevancia y precisión de las respuestas del LLM, y que se ha convertido en un campo con un cuerpo teórico cada vez más extenso e incluso han surgido ya algunos cursos (muchos de ellos como diplomados en línea y otros como parte de programas de formación técnica o profesional) específicamente orientados al entrenamiento en esta práctica (Broshar, 2024; Craig et al., 2024; Hashemi-Pour, 2023; Nerella et al., 2024; Raiaan et al., 2023).

4 Utilidad y Usos de los LLMs en la Educación y en la Educación Médica

Como se ha venido esbozando en apartados previos, los LLMs han demostrado un potencial notable para optimizar distintos procesos educativos, no solo en el ámbito universitario en general, sino también de manera específica en la formación médica. A lo largo de esta sección, primero se describirán las ventajas que presentan los LLMs en entornos académicos diversos y, posteriormente, se examinará con mayor detenimiento su rol en la educación médica, con referencia explícita a la integración de ciencias básicas y clínicas, el razonamiento clínico y el desarrollo de competencias avanzadas y metacognitivas. Finalmente, se incluirá una breve revisión de la percepción que estudiantes, docentes y otros actores de la educación médica han expresado sobre la introducción de estas tecnologías.

4.1 LLMs en la educación general

4.1.1 Acceso a información y personalización del aprendizaje

Los LLMs pueden cumplir la función de asistentes virtuales que proporcionen de forma inmediata resúmenes, explicaciones o referencias bibliográficas, sirviendo como una suerte de biblioteca virtual interactiva, lista para filtrar, resumir y explicar grandes volúmenes de información, además de posibilitar el acceso a la misma, potencialmente, en cualquier momento que el alumno lo requiera, eliminando barreras como el horario de clases o de funcionamiento de bibliotecas formales o la disponibilidad de docentes o personal asistencial en general (Choi & Abdirayimov, 2024; Lawson McLean, 2024; Mannekote et al., 2024; Mondal, De, et al., 2024). Esta inmediatez y versatilidad mejoran la autonomía del alumno, favoreciendo la autogestión de su estudio y brindando la posibilidad de profundizar en tópicos concretos, de forma ágil y directa, desde una plataforma única, en lugar de tener que hacer búsquedas en bases de datos o bibliotecas dispersas (Choi & Abdirayimov, 2024; Dong et al., 2024).

Además, en entornos con recursos limitados (por ejemplo, regiones rurales o instituciones con poca infraestructura bibliotecaria), estos modelos pueden servir como fuente de conocimiento médico y científico relativamente accesible, reduciendo la brecha formativa y equilibrando oportunidades de aprendizaje cuando no es factible disponer de expertos en todas las áreas. En tales contextos, un LLM con un entrenamiento adecuado o un ajuste fino enfocado en bibliografía esencial, puede desempeñar el rol de tutor sustituto que oriente la adquisición de conocimientos clave (J. Liu et al., 2024; Rashid et al., 2024; Suresh & Misra, 2024).

Además, la capacidad de estos modelos para ajustar la complejidad del discurso, explicando un mismo tema bien sea a nivel muy básico o a nivel avanzando, o incluso

recurriendo a formas alternativas de abordar el contenido de materiales académicos estáticos y rígidos, adaptándolo según las características y necesidades individuales del usuario, puede resultar de gran utilidad en entornos con grupos heterogéneos, donde no todos los estudiantes poseen la misma base conceptual o el mismo ritmo de aprendizaje (Choi & Abdirayimov, 2024; Mannekote et al., 2024).

4.1.2 Generación de escenarios de aprendizaje y tutorías virtuales

La habilidad de los LLMs para generar contenido textual coherente les permite desempeñar el rol de "tutores virtuales", diseñando preguntas o ejercicios académicos, o bien proporcionando la valoración y retroalimentación inmediata a las respuestas, o bien proporcionando resúmenes o esquemas útiles para el repaso. Por otra parte, el hecho de poder graduar la dificultad de la información o el estilo de redacción según las necesidades de cada usuario les otorga un valor adicional, sobre todo en clases numerosas y heterogéneas, donde cada alumno progresa a un ritmo distinto. Todo esto enriquece la experiencia del alumno y lo mantiene en un intercambio constante con la herramienta, favoreciendo la motivación por el contenido y el proceso de aprendizaje potencialmente ayudando a desarrollar su autonomía en el aprendizaje (Choi & Abdirayimov, 2024; Dong et al., 2024; Mannekote et al., 2024; Mondal, De, et al., 2024).

Por ejemplo, el estudiante puede formular una duda sobre un tema complejo específico (ej. el rol de los linfocitos T CD4+ en la respuesta antiviral), recibiendo no solo una respuesta directa, sino también sugerencias para reflexionar sobre problemas análogos (ej. el impacto de los inmunosupresores e inmunomoduladores en la actividad de los linfocitos T CD4+ en la respuesta antiviral, en personas trasplantadas), complementando así la información y permitiendo al estudiante establecer puentes conceptuales con otros temas, con lo que se propicia un entorno de aprendizaje más interactivo, profundo, auténtico y reflexivo (Lawson McLean, 2024; Mondal, De, et al., 2024).

Adicionalmente, estas herramientas representan una ayuda importante a los docentes, al permitir optimizar varias de sus labores con lo que se logra disminuir el tiempo dedicado a la creación de materiales de académicos, planeación de actividades o a la revisión de trabajos, de modo que pueda dedicar más tiempo al seguimiento, supervisión y acompañamiento personalizado de sus estudiantes (Mannekote et al., 2024; Ragolane et al., 2024).

4.1.3 Simulación de casos y aprendizaje contextualizado

La simulación de escenarios es otro de los aportes centrales de los LLMs en la educación general, pues permiten presentar problemas progresivos, a menudo inspirados en situaciones reales y sin la limitación de seguir necesariamente guiones rígidos, por lo que pueden plantearse muchos escenarios diversos, donde el estudiante va recibiendo

información en forma escalonada y debe proponer hipótesis o soluciones en cada paso (Choi & Abdirayimov, 2024; Mannekote et al., 2024).

Estas simulaciones pueden aplicarse a múltiples disciplinas y se han vuelto especialmente atractivas en asignaturas que requieren razonamiento crítico y resolución de problemas, dado que el LLM no se limita a una mera calificación, sino que contribuye a la comprensión del proceso. Por ejemplo, el LLM puede tomar el papel de “paciente estandarizado” o “paciente virtual”, que presentará un cuadro clínico sugestivo de una condición o situación concreta (ej. una enfermedad aguda grave, como un infarto miocárdico; dificultades con la adherencia o tolerancia a un medicamento, etc.), permitiendo al estudiante practicar la entrevista médica, la toma de decisiones, e incluso la exploración de actitudes y habilidades comunicativas en un entorno seguro y repetible (Wang et al., 2025). El modelo, por su parte, responde con cambios en la situación simulada, progresando hacia distintos tipos de desenlaces, que pueden plantear nuevos retos que el estudiante debe enfrentar “en tiempo real”, y todo esto acompañado con momentos dedicados a la retroalimentación oportuna (durante o al final de la simulación), que pueden señalarle posibles omisiones o incluso introducir un componente de complejidad adicional para alentar la reflexión.

De esta forma, se amplía la variedad de casos clínicos que el alumno puede enfrentar y se propicia una experiencia de aprendizaje altamente personalizable. (ej. simuladores de consultas médicas, en las que el LLM actúa como paciente y responde a las preguntas y decisiones del estudiante) (Choi & Abdirayimov, 2024; Mannekote et al., 2024).

4.2 Contribuciones y aplicaciones de los LLMs en la educación médica

Aunque con las secciones anteriores ya se pueden entrever algunos potenciales beneficios de la incorporación de los LLMs a la enseñanza de las ciencias de la salud, este es uno de los ámbitos en los que estas herramientas cobran especial relevancia y, desde su popularización y posterior acelerado desarrollo de estas tecnologías en los últimos años, diversas investigaciones se han llevado a cabo para valorar a fondo los posibles usos e impactos de estas herramientas en la formación en áreas de la salud, especialmente en la educación médica.

4.2.1 Facilitación de la integración de ciencias básicas y clínicas

Uno de los momentos más retadores durante la formación médica es el paso de la etapa de formación en fundamentos biomédicos básicos, generalmente impartidos en los primeros años, y la etapa de formación clínica. La transición de una a otra fase requiere que el estudiante pueda integrar todos sus conocimientos biomédicos entre sí y con los nuevos saberes teóricos clínicos, para luego poder aplicar todo a situaciones prácticas.

Sin embargo, a menudo, más que una transición fluida, lo que ocurre es un salto brusco, en el que los estudiantes enfrentan grandes dificultades para conseguir esa integración armoniosa de saberes, lo cual no solo repercute en su desempeño académico inmediato, sino en la percepción que tiene del mismo proceso formativo y en aspectos de su vida personal (Delavari et al., 2024; Durning et al., 2024; Kassirer, 2010; Kulasegaram et al., 2015; Parodis et al., 2021). Este es uno de los puntos en que se han centrado varias investigaciones sobre los potenciales usos y beneficios de los LLMs en la educación médica, resaltando en algunos aspectos como:

- Articulación de fundamentos biomédicos con la práctica clínica: los LLMs son capaces de enlazar información de distintas disciplinas, presentando explicaciones o ejemplos aplicados que van desde la descripción de mecanismos bioquímicos hasta su relevancia en la detección de signos clínicos (Jiang et al., 2024; Lucas et al., 2024; Pears & Konstantinidis, 2024). Por ejemplo, para un alumno que está estudiando "shock séptico", el modelo puede integrar los fundamentos fisiológicos, como la activación de mediadores inflamatorios, con aspectos semiológicos y farmacológicos, como los mecanismos conducen a la presentación clínica de esa condición y al visibilizar el modo en que la acción de cada medicamento impacta en los desenlaces del paciente. Además, al funcionar a modo de un espacio conversacional, el estudiante puede continuar profundizando en el tema a partir de nuevas preguntas, como valorando los motivos por los que se hace seguimiento con estudios de laboratorio específico, a partir de lo que ellos reflejan de la fisiopatología de la enfermedad, con lo que va "uniendo puntos" entre la ciencia biomédica y la práctica clínica. Esta combinación de lo básico con lo aplicado otorga un enfoque más holístico al proceso de aprendizaje (Choi & Abdirayimov, 2024; Dong et al., 2024).
- Generación de escenarios clínicos basados en fundamentos científicos: en contextos de aprendizaje activo, el LLM puede generar casos basados en información biomédica robusta (Delavari et al., 2024; Kassirer, 2010; Parodis et al., 2021). Por ejemplo, se pueden generan casos clínicos concretos (esto es, una aplicación específica del Aprendizaje Basado en Casos, ABP), con manifestaciones clínicas y hallazgos de laboratorio derivados de una alteración en alguna función fisiológica particular, como la contractilidad cardíaca, invitando al alumno a reconocer el vínculo entre el sustrato biomédico básico y la presentación de la enfermedad. Esta conexión enfatiza la relevancia de la "causalidad" en la medicina: a cada cambio fisiopatológico le corresponde un conjunto de signos o síntomas, y los LLMs, al articular la información de manera contextual y adaptada al estudiante, refuerzan la comprensión holística de los procesos (Dong et al., 2024; Jiang et al., 2024; Lucas et al., 2024).
- Refuerzo en entornos con limitados recursos: en algunos entornos con escasa disponibilidad de docentes o de plataformas de simulación costosas, los LLMs pueden servir como un nodo central de referencia para aclarar dudas, proponer ejercicios prácticos y guiar el razonamiento en problemas clínicos que involucren conceptos básicos, así como permitir desempeñar funciones de simulación de algunas situaciones clínicas. En este sentido, si se dispone de acceso a un LLM

entrenado en literatura biomédica (o ajustado con fines didácticos) podría mitigar la desconexión entre la teoría y la práctica, en algunso escenarios con limitaciones en infraestructura u otros recursos (Dong et al., 2024; J. Liu et al., 2024; Lucas et al., 2024; Shah et al., 2024).

4.2.2 Refuerzo del razonamiento clínico y diagnósticos diferenciales

El razonamiento clínico es una de las habilidades transversales más exigentes de la carrera médica, exigiendo no solo la combinación de conocimientos, sino la capacidad de generar, contrastar y refinar hipótesis bajo incertidumbre (Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kassirer, 2010; Kulasegaram et al., 2015).

- Tutorías virtuales: diversas experiencias documentan que el LLM puede hacer las veces de tutor que revisa las propuestas diagnósticas del estudiante, cuestiona las bases de su hipótesis y sugiere diagnósticos alternativos (Kämmer et al., 2024; Kanjee et al., 2023). Por ejemplo, ante un caso ambiguo de dolor torácico, el alumno plantea su sospecha y el LLM reacciona confrontando la hipótesis con hallazgos clínicos adicionales, invitando a explorar otras vías como un tromboembolismo pulmonar o una disección aórtica. Este contraste permite al estudiante desarrollar una mayor amplitud diagnóstica y examinar la coherencia de su razonamiento (Guzmán-Valdivia-Gómez et al., 2022; Kämmer et al., 2024).
- Reconocimiento de sesgos cognitivos: un aspecto frecuentemente subrayado es la importancia de que el estudiante aprenda a reconocer sus sesgos cognitivos, como el de anclaje, confirmación o cierre prematuro (Durning et al., 2024; Shusterman et al., 2024). Al comparar la respuesta inicial del alumno con la propuesta del LLM, se estimula la reflexión sobre los criterios empleados para proponer una hipótesis específica. El estudiante puede notar que omitió datos relevantes o se cerró a un solo diagnóstico, sin considerar otras posibilidades. Este tipo de ejercicios conducen a un proceso reflexivo en el alumno, al hacerlo consciente de las fortalezas y debilidades de sus propias estrategias diagnósticas y procesos inductivos y deductivos, lo que fomenta la metacognición, reduciendo la repetición de errores en el futuro (Delavari et al., 2024; Kassirer, 2010; Parodis et al., 2021).
- Situaciones de entrenamiento simuladas: además de las tutorías personalizadas, durante el estudio o repaso independiente, como antes se mencionaba, los LLMs pueden asumir el rol de pacientes, presentando signos, síntomas y datos de laboratorio, a medida que el estudiante los va solicitando, durante una atención médica simulada (Choi & Abdirayimov, 2024; Jiang et al., 2024; J. Liu et al., 2024; Wang et al., 2025). En estos escenarios, la IA puede calibrar la dificultad (por ejemplo, añadiendo o suprimiendo factores de confusión) e incluso reproducir diferentes variaciones en la evolución del caso, según las distintas decisiones o formas en que el estudiante, correcta o incorrectamente, vaya tomando (Choi & Abdirayimov, 2024; Lucas et al., 2024; Nerella et al., 2024; Wang et al., 2025).

4.2.3 Desarrollo de competencias médicas

En la formación médica, y como se discute ampliamente en los documentos técnicos del Ministerio de Salud y Protección Social de Colombia (Ortiz Monsalve et al., 2014) sobre el enfoque de educación por competencias para las profesiones de la salud, el paradigma educativo actual busca articular conocimientos, habilidades, actitudes y valores en un desempeño práctico y contextualizado. Se reconocen, en términos generales, tres categorías principales de competencias:

- Competencias básicas: aquellas que usualmente se desarrollan durante la formación básica y secundaria y se consolidan a lo largo de la formación profesional (ej. Lecto-escritura, fundamentos de razonamiento matemático, uso básico de Tecnologías de la Información y Comunicación – TIC), sirviendo de cimiento para el desarrollo posterior de otras competencias.
- Competencias genéricas o transversales: aplicables a múltiples campos profesionales y que incluyen elementos como la comunicación, la capacidad de investigación, el liderazgo, el razonamiento crítico, la metacognición y la habilidad de trabajo colaborativo, entre otras. Se corresponden, en buena medida, con la clasificación de Tuning (instrumentales, interpersonales y sistémicas).
- Competencias específicas o disciplinares: las que orientan el desempeño propio de la profesión médica, como el diagnóstico clínico, la formulación de planes terapéuticos, la seguridad del paciente y la gestión de la atención, entre otras.

Aunque es útil esta categorización para fines del planeamiento curricular y la evaluación del desempeño de los estudiantes, en la vida y la práctica profesional, estas categorías no funcionan de forma aislada, sino que se interrelacionan para dar lugar a un médico con una preparación integral, capaz de afrontar escenarios diversos y cambiantes (Ortiz Monsalve et al., 2014). A continuación, se discuten algunos aspectos centrales de cómo las herramientas de IA, especialmente los LLMs, pueden insertarse en la formación médica por competencias.

i. Competencias Específicas en la Formación del Médico y el Aporte de LLMs

Las competencias específicas en la carrera de medicina son diversas. Todas ellas exigen la integración de aspectos cognitivos (conocimientos biomédicos), procedimentales (habilidades prácticas) y actitudinales (valores éticos, empatía y respeto por la diversidad sociocultural) (Delavari et al., 2024; Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kassirer, 2010; Ortiz Monsalve et al., 2014; Parodis et al., 2021).

A. Diagnóstico clínico

Esta competencia implica recabar información mediante la anamnesis y la exploración física, articularla con la evidencia científica y el contexto del paciente, y hacer uso de un razonamiento clínico que conjuga procesos intuitivos y analíticos (Ortiz Monsalve et al., 2014). Los LLMs pueden reforzar la toma de decisiones diagnósticas, al ofrecer un entorno conversacional que promueva el contraste, argumentación y generación de hipótesis alternas. En otras palabras, la IA puede funcionar como un “co-razonador”, capaz de sugerir pruebas complementarias o llamar la atención sobre datos omitidos, potenciando la competencia diagnóstica (Choi & Abdirayimov, 2024; Jiang et al., 2024; Kämmer et al., 2024; Kanjee et al., 2023; Lawson McLean, 2024; J. Liu et al., 2024; Lucas et al., 2024; Shah et al., 2024; Suresh & Misra, 2024; Togunwa et al., 2024).

B. Intervención terapéutica y seguridad del paciente

El médico, además de diagnosticar, debe plantear planes de manejo (farmacológicos, de control de hábitos, quirúrgicos o de rehabilitación) apropiados a cada situación y que además eviten exponer al paciente de forma innecesaria a riesgos de efectos adversos o complicaciones ante esos mismos enfoques terapéuticos. Esto requiere aplicar la medicina basada en la evidencia y optimizar recursos para salvaguardar la seguridad del paciente (Ortiz Monsalve et al., 2014). Los LLMs, de forma similar a lo que pueden hacer con el soporte en los planteamientos diagnósticos, también pueden apoyar la formulación de planes basados en guías clínicas, cuestionando aspectos de seguridad del paciente (interacciones farmacológicas, identificación de factores de riesgo para efectos adversos, etc.) en los escenarios formativos, ofreciendo un ambientes de práctica simulada o espacios conversacionales con un “tutor virtual”, facilitando la integración de los conocimientos teóricos sobre enfoques terapéuticos con el desarrollo de una conciencia por la seguridad del paciente y el uso racional de los recursos (Choi & Abdirayimov, 2024; Jiang et al., 2024; Kämmer et al., 2024; Kanjee et al., 2023; Lawson McLean, 2024; J. Liu et al., 2024; Lucas et al., 2024; Shah et al., 2024; Suresh & Misra, 2024; Togunwa et al., 2024).

C. Promoción de la salud y prevención de la enfermedad

Los médicos también tienen una importante función en la educación a la comunidad en aspectos para la promoción de la salud y la prevención de enfermedades, a través del control de factores de riesgo, el fortalecimiento de factores protectores, la detección temprana de potenciales problemas de salud y la participación en programas preventivos (Ortiz Monsalve et al., 2014). En este sentido, un LLM puede simular conversaciones con pacientes, familias y comunidades, donde el estudiante practica distintos enfoques para hacer este tipo de intervenciones de promoción y prevención, pudiendo incluso ajustar la complejidad de estos escenarios, a partir de la inclusión de aspectos socioculturales distintos en esas interacciones (Choi & Abdirayimov, 2024; Jiang et al., 2024; Kämmer et al., 2024; Kanjee et al., 2023; Lawson McLean, 2024; J. Liu et al., 2024; Lucas et al., 2024; Shah et al., 2024; Suresh & Misra, 2024; Togunwa et al., 2024).

D. Competencias específicas de subespecialidades: un ejemplo para la formación general

Existen, además, competencias específicas aplicadas a especialidades concretas de la medicina, que no son solo esenciales para el médico especialista en ese campo específico, sino que el médico general debe desarrollar, al menos hasta nivel inicial (Delavari et al., 2024; Kassirer, 2010; Ortiz Monsalve et al., 2014; Parodis et al., 2021). LLMs contribuyen a integrar dichos saberes disciplinares.

Por ejemplo, algunos estudios han demostrado que en especialidades clínicas, como pediatría, medicina interna y psiquiatría, el residente† puede ensayar la elaboración de planes diagnósticos o guías de procedimientos y estrategias de comunicación con sus pacientes (Chien & Tai, 2024; Kämmer et al., 2024; Kanjee et al., 2023; Shusterman et al., 2024; Suresh & Misra, 2024); en radiología, otra especialidad clínica pero con unas características asistenciales diferentes, el LLM ayuda a sistematizar la búsqueda de alteraciones y el reporte de hallazgos y verificar la concordancia con protocolos de seguridad de paciente (Togunwa et al., 2024); de forma análoga, en especialidades quirúrgicas, como la neurocirugía, se han documentado usos de LLMs para guiar la comprensión de técnicas operatorias y el manejo perioperatorio de patologías complejas, así como para entrenar a los residentes en la identificación de urgencias neurológicas (Arfaie et al., 2024); también, en oftalmología, se explora la asistencia de estos modelos al analizar correlaciones entre hallazgos imagenológicos y la fisiopatología ocular, reforzando la competencia diagnóstica y el manejo de casos a menudo desatendidos en entornos con pocos oftalmólogos (Agnihotri et al., 2024).

Es razonable esperar que este tipo de efectos no sean aplicables sólo a estudiantes de nivel posgrado, sino que puedan ser extrapolables a la formación del médico general, en cuanto al desarrollo de los niveles fundamentales de competencias de esas y otras especialidades médicas.

ii. Competencias genéricas y la metacognición en la formación clínica

Las llamadas competencias genéricas son indispensables para la práctica médica y atraviesan todo el proceso de enseñanza (Ortiz Monsalve et al., 2014). Los LLMs también han demostrado utilidad en el fomento del desarrollo de estas competencias, en múltiples estudios:

A. Pensamiento crítico y metacognición

Como ya antes se describió, la autorreflexión, el reconocimiento de los propios sesgos cognitivos y la evaluación continua del propio desempeño son competencias fundamentales en todo programa de educación médica. Los LLMs pueden fomentar esa autorreflexión cuando el estudiante compara su razonamiento o su plan de manejo con la propuesta de la IA y descubre discrepancias, identifica sus propios sesgos y otras actitudes que puedan favorecer el error en su razonamiento y le ayuda tomar consciencia

† Un estudiante de una especialización médica, de nivel posgrado.

de cómo se enfrenta al material de estudio y organiza la información y otros aspectos clave de su proceso de aprendizaje, reforzando así la dimensión metacognitiva (Choi & Abdirayimov, 2024; Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kämmer et al., 2024; Kanjee et al., 2023; J. Liu et al., 2024; Ortiz Monsalve et al., 2014; Shusterman et al., 2024).

B. Comunicación y relación médico-paciente

La habilidad comunicativa y la empatía constituyen competencias genéricas clave en el acto médico (Choi & Abdirayimov, 2024; Omiye J. A. & 2024, 2024; Ortiz Monsalve et al., 2014; Suresh & Misra, 2024). Sin embargo, con frecuencia no se entrenan de manera sistemática. Un LLM puede simular un “paciente virtual” con dudas específicas, actitudes reticentes o rasgos socioculturales distintos y retadores, para que el estudiante practique su forma de indagar y empatizar, ajustando el discurso según el contexto, fortaleciendo las competencias del alumno en aspectos comunicacionales, empatía y construcción de una relación médico-paciente adecuada (Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kämmer et al., 2024; Kanjee et al., 2023; J. Liu et al., 2024; Mondal, De, et al., 2024; Ortiz Monsalve et al., 2014; Shusterman et al., 2024)..

C. Trabajo en equipo y colaboración interprofesional

El médico general debe colaborar con otros profesionales de la salud (enfermería, trabajo social, especialistas) en la atención del paciente (Ortiz Monsalve et al., 2014). Aunque un LLM no sustituye la dinámica real de un equipo humano, puede servir como recurso para ensayar la asignación de roles o la discusión de casos de manera virtual, siempre con supervisión docente. El objetivo es ejercitar la toma de decisiones compartidas y coordinar planes de cuidado interdisciplinario (Lawson McLean, 2024; Lucas et al., 2024; Mannekote et al., 2024; Suresh & Misra, 2024).

iii. Articulación con las Competencias Básicas

Tal como se señalaba, las competencias básicas (lecto-escritura, razonamiento matemático esencial, uso elemental de TIC, etc.) se consolidan antes o en los primeros semestres de la carrera. Sin embargo, por lo reciente de la popularización y uso extendido de LLMs (y otras formas de IA) en entornos educativos y profesionales, estas competencias deben ser repensadas para la inclusión de fundamentos del uso adecuado de este tipo de herramientas. Estas competencias representan la base fundamental para que el estudiante pueda interactuar eficientemente con un LLM y, en general, aprovecharlos al máximo para potenciar su formación, lo cual ciertamente aplica a cualquier campo del conocimiento, no solo a la medicina.

4.3 Percepciones de estudiantes, docentes y otros actores

Por supuesto, la creciente incorporación de LLMs en la educación médica, sea a modo de usos esporádicos por parte de estudiantes o docentes, o por estrategias formales de inclusión en los programas formativos, ha suscitado diversas reacciones y expectativas

por parte de los estudiantes, el profesorado y otros actores vinculados al proceso formativo (instituciones, cuerpos reguladores, grupos empresariales, etc.), que van desde visiones altamente positivas por la versatilidad y utilidad de estas herramientas de IA, hasta preocupaciones por aspectos tanto pedagógicos como éticos en torno a la forma en que estos modelos operan y la falta de directrices claras sobre cómo utilizarlos. A continuación, se detallan algunas de las percepciones identificadas por varios estudios, que han evaluado la forma en que estudiantes, docentes y otros actores educativos se relacionan con los LLMs y sus impresiones para el presente y futuro del uso de estas herramientas, en la educación médica.

4.3.1 Perspectiva estudiantil

Varios estudios describen un alto grado de aceptación de los LLMs por parte de los estudiantes de medicina, y de otras áreas de las ciencias de la salud, citando la inmediatez en la retroalimentación y la posibilidad de personalizar el aprendizaje como los beneficios más valorados (Buabbas et al., 2023; Busch et al., 2024; Cherrez-Ojeda et al., 2024; Ganjavi et al., 2024; Jackson et al., 2024; Mondal, Karri, et al., 2024; Pan & Ni, 2024; Safranek et al., 2023). Los alumnos también resaltan que la facilidad de accesos y la disponibilidad continua a estos modelos es valorada como un factor que fomenta la autonomía y la autogestión del estudio, al no depender de horarios de clase ni de la disponibilidad de tutores o docentes para resolver dudas (Ganjavi et al., 2024; Mondal, Karri, et al., 2024). Además, se aprecia un incremento en la motivación y la autoconfianza cuando el estudiante contrasta sus razonamientos clínicos con las sugerencias del modelo, y percibe, en muchos casos, una mejora en su proceso de razonamiento clínico y habilidades metacognitivas (Amiri et al., 2024; Zdravkova et al., 2023).

Diversos estudios han valorado, a través de encuestas en escuelas de medicina y otras ciencias de la salud, en varios países (ej. Estados Unidos, China, Alemania, Palestina, Tailandia y algunos países latinoamericanos), los estudiantes coinciden en que los LLMs ofrecen un entorno seguro y flexible para la práctica de competencias clínicas y la revisión de información compleja (Angkurawaranon et al., 2024; Ganjavi et al., 2024; Jackson et al., 2024; Jebreen et al., 2024; Laupichler et al., 2024; Pan & Ni, 2024). No obstante, también destacan la heterogeneidad en las destrezas digitales de los estudiantes, con niveles típicamente medios o bajos de familiaridad con los LLMs y otras formas de IA[‡], de modo que los estudiantes consideran indispensable que se implementen espacios capacitación complementaria para usar de manera efectiva estas

[‡] De forma llamativa, tanto en estudios en países desarrollados (Laupichler et al., 2024; Pan & Ni, 2024) como en desarrollo(Jebreen et al., 2024), existe una tendencia marcada a una mayor aceptación y familiaridad con los LLMs entre hombres que entre mujeres. Esto supone que pueden existir también otros factores sociales, más allá de los económicos, que afecten la equidad en la alfabetización digital entre estudiantes, lo que puede repercutir tanto en el acceso, como en el uso adecuado de IA en la formación médica.

herramientas (Amiri et al., 2024; Buabbas et al., 2023). Esta última necesidad hace evidente que la adopción exitosa de los LLMs no solo depende de su disponibilidad, tema del que se hablará adelante, sino también de una adecuada preparación institucional que promueva en los estudiantes una alfabetización digital robusta, unida a la comprensión de los alcances y limitaciones propios de la IA (Busch et al., 2024; Cherrez-Ojeda et al., 2024; Javed & Arooj, 2023; Jebreen et al., 2024; Li & Qin, 2023; D. S. Liu et al., 2022; Mondal, Karri, et al., 2024; Safranek et al., 2023; Waldock et al., 2024).

4.3.2 Perspectivas del profesorado y otros actores educativos

La valoración docente frente a los LLMs combina, por un lado, entusiasmo por el potencial de estas tecnologías para dinamizar la docencia, ampliar los recursos pedagógicos, potenciar los resultados de aprendizaje en los estudiantes y reducir algunas cargas laborales, y por otro, prudencia respecto al grado de fiabilidad y a la eventual superficialidad de las respuestas generadas, a problemas con la alfabetización digital tanto de docentes como de estudiantes y a la dependencia excesiva de estas herramientas (Javed & Arooj, 2023; Lucas et al., 2024; Shah et al., 2024). En varios estudios, los profesores han reconocido que los LLMs pueden agilizar la generación de cuestionarios, casos clínicos y actividades de evaluación, reduciendo el tiempo invertido en labores repetitivas y facilitando una retroalimentación inicial al estudiante (Javed & Arooj, 2023; Waldock et al., 2024). Asimismo, en contextos con escasa infraestructura o limitaciones de personal, se observa una recepción favorable, pues estos modelos podrían suplir en parte la carencia de expertos en áreas específicas o permitir una mayor inversión de tiempo en el seguimiento y retroalimentación de los estudiantes (Choi & Abdirayimov, 2024; Dong et al., 2024).

Sin embargo, la preocupación por la posible dependencia excesiva de los alumnos y docentes en estas tecnologías, especialmente bajo el riesgo de que afecte el desarrollo del pensamiento crítico y reflexivo, al facilitar excesivamente algunos procesos, y la carga adicional de trabajo por la supervisión requerida para garantizar la exactitud y pertinencia de las respuestas generadas por los LLMs, en cualquier tipo de interacción con los estudiantes en sus procesos formativos, son elementos casi invariablemente reportados en todos los estudios que valoran la posición de los docentes frente a estas herramientas (Li & Qin, 2023; Mondal, Karri, et al., 2024; Safranek et al., 2023).

Desde la perspectiva institucional, se plantea la necesidad de formalizar la integración de los LLMs en el currículo, asegurando que su rol sea complementario y no sustitutivo de las experiencias educativas teóricas y prácticas, tanto en las ciencias básicas como en las clínicas, ni de la interacción humana, con docentes, pacientes y otros profesionales de la salud, esencial para el desarrollo de competencias profesionales (Mannekote et al., 2024; Ragolane et al., 2024). Asimismo, autoridades académicas y diseñadores de planes de estudio, en consonancia con la literatura, insisten en la imperativa necesidad de establecer lineamientos claros para el uso responsable de la IA

y la incorporación progresiva y sistemática de estos modelos en asignaturas, seminarios o prácticas (Dong et al., 2024; Lawson McLean, 2024).

5 Consideraciones éticas, técnicas y regulatorias en la adopción de LLMs en la educación médica

Como ya anteriormente se empezaba a esbozar, la implementación de LLMs en la formación médica no solo representa oportunidades de innovación, sino que también conlleva una serie de desafíos que deben atenderse de manera integral. A continuación, se exponen los principales aspectos éticos, técnicos y regulatorios, así como posibles estrategias de mitigación para asegurar un uso responsable y eficaz de estas herramientas.

5.1 Fiabilidad de la información, consistencia y riesgo de "alucinaciones" en entornos formativos

Una de las inquietudes más citadas en la literatura es la fiabilidad de los contenidos generados por los LLMs, dadas sus bases probabilísticas. Estos modelos pueden generar respuestas que aparentan ser coherentes pero que, en realidad, contienen datos erróneos o incompletos, fenómeno frecuentemente denominado "alucinaciones" o fabricaciones de información (Gonzalez et al., 2024; Omiye J. A. & 2024, 2024; Raiaan et al., 2023; Shusterman et al., 2024), y que pueden ser particularmente difíciles de identificar por, precisamente, estudiantes que aún no son expertos en un campo específico, dada la prolijidad de estos modelos para redactar textos de forma convincente. En el ámbito de la educación médica, donde los estudiantes dependen de la exactitud de los conceptos para forjar sus conocimientos, habilidades y competencias, estos errores pueden comprometer seriamente la calidad del aprendizaje y, en el largo plazo, la seguridad del paciente.

Por ejemplo, se han documentado casos donde la IA propone diagnósticos o tratamientos que parecen plausibles, a primera vista, pero no se basan en evidencia médica actualizada (Gonzalez et al., 2024; J. Liu et al., 2024). Aunque dichas "alucinaciones" pueden mitigarse mediante supervisión humana y verificación de las fuentes y la construcción de mecanismos de retroalimentación (por ejemplo, a través de "revisores" expertos o rutinas de cotejo bibliográfico), subsiste la necesidad de que tanto docentes como estudiantes comprendan que la naturaleza estadística de un LLM no garantiza la veracidad del contenido y que, incluso desde el entrenamiento del modelo, se implementen estrategias para mejorar la fiabilidad de sus respuestas (Choi & Abdirayimov, 2024; Ganjavi et al., 2024; Shusterman et al., 2024).

De otro lado, incluso cuando los LLMs generan respuestas correctas, surgen dificultades de reproducibilidad. Con pequeños ajustes en los parámetros de inferencia o ante

preguntas con ligeras variaciones semánticas, el modelo puede proveer salidas distintas (Raiaan et al., 2023). Esto impacta en la posibilidad de replicar con fidelidad, cuando sea necesario, experimentos o simulaciones clínicas, dificultando la evaluación sistemática de su calidad (Jiao et al., 2024; Shusterman et al., 2024). En un contexto educativo, donde se esperan criterios de evaluación estables, esta variabilidad exige pautas claras para definir los prompts y la configuración del modelo, con el fin de asegurar consistencia en el proceso formativo.

5.2 Sesgos inherentes a los modelos y su impacto en la formación clínica

Una de las problemáticas más resaltadas por distintos estudios sobre el uso de LLMs en la educación y la práctica médica, es que los datos de entrenamiento de los LLMs suelen estar compuestos por volúmenes de texto que, aunque sean masivos, en muchos casos reflejan sesgos derivados de la publicación y selección de la información con que se entrena el modelo, lo que facilita que los datos tengan sesgos de género, étnicos, culturales y de otros tipos. Esto puede traducirse en la perpetuación de estereotipos o la infrarrepresentación de características poblacionales que, aún si son minoritarias, deberían ser consideradas en la formación del profesional de la salud, pues un modelo sesgado podría mostrar inclinaciones al sugerir manejos clínicos menos apropiados para determinados grupos poblacionales o enfatizar incorrectamente ciertas patologías sobre otras (J. Liu et al., 2024; Omiye J. A. & 2024, 2024; Yan et al., 2023).

Los riesgos se elevan cuando el estudiante, con poca experiencia, confía de manera acrítica en la respuesta del LLM. Dado que el razonamiento clínico debe considerar la diversidad de contextos, la formación médica necesita incluir, en su programa curricular, la enseñanza de estrategias sistemáticas que permitan controlar, prever, detectar y corregir dichos sesgos (Jebreen et al., 2024; Jiao et al., 2024; Shusterman et al., 2024; Yanagita et al., 2024). No obstante, también es fundamental la vigilancia docente para confirmar que los alumnos no asuman estas respuestas sesgadas como verdades clínicas, sino que aprendan a cuestionar, contrastar y reevaluar la información, así como medidas regulatorias para garantizar que los LLMs destinados a la educación médica cuenten con medidas para mitigar este tipo de problemáticas.

5.3 Restricciones por contenido sensible

Un aspecto que ha emergido recientemente en la literatura es la restricción que los filtros de contenido pueden imponer en ciertos escenarios educativos, sobre todo en temas delicados o que, en otros contextos, usualmente puedan aparecer censurados para proteger al usuario y público general, como aspectos relacionados con la sexualidad y reproducción, el uso de sustancias ilegales o la violencia. No obstante, en la labor de los profesionales de la salud, es frecuente que se deban atender casos precisamente

relacionados con estas y otras situaciones potencialmente sensibles, por lo que es necesario incluirlas en el entrenamiento de los estudiantes de estas ciencias. De este modo, por ejemplo, en situaciones en que se desea simular un caso con una historia clínica compleja relacionada con agresión sexual, comportamientos autolesivos o problemáticas sociales extremas (ej. sexualidad en menores de edad, abuso de sustancias psicoactivas, etc.), las políticas de uso y los filtros automáticos de los LLMs podrían bloquear o recortar las respuestas, impidiendo al alumno profundizar con realismo en la exploración o impedir la continuidad de la simulación (Chien & Tai, 2024).

Esta situación particular, identificada por algunos estudios, suscita la necesidad de configurar adecuadamente las herramientas de IA para que puedan adaptarse a las necesidades reales de cada contexto de uso específico, por ejemplo, habilitando funciones específicas como “modos académicos” (en inglés, usualmente llamados “sandbox modes”) cuando sea viable, y, sobre todo, de definir protocolos institucionales para la supervisión y justificación de la simulación, balanceando el respeto por los estándares éticos de los modelos y la legitimidad de formar a los futuros profesionales frente a problemáticas clínicamente reales (Choi & Abdirayimov, 2024; Lucas et al., 2024)

5.4 Privacidad, protección de datos y brecha digital

Otro de los puntos más polémicos en las revisiones sobre el uso de LLMs en temas relacionados con la salud, incluyendo la educación médica, es que en estos contextos se manejan datos sensibles, tanto de pacientes como de estudiantes (ej. Datos personales, historias clínicas de práctica, resultados de estudios diagnósticos, etc.). El uso de LLMs en interacciones de chat o en análisis de expedientes clínicos electrónicos puede exponer dicha información a riesgos de seguridad, especialmente en LLMs en línea o que almacenen datos en nubes o servidores no propios de la institución educativa o de salud (Jackson et al., 2024; Jiang et al., 2024).

Además, aunque uno de los potenciales beneficios del uso de LLMs en entornos educativos es que puede hacer asequibles materiales bibliográficos y recursos pedagógicos que no están disponibles en algunos escenarios de recursos limitados, también se debe resaltar que algunos estudios han evidenciado que en países o instituciones donde la infraestructura tecnológica es limitada, el acceso a estas tecnologías aún es reducido o incluso mínimo, evidenciando una brecha tecnológica y digital que agrava la inequidad formativa (Choi & Abdirayimov, 2024; Pan & Ni, 2024).

Las normativas de protección de datos de salud de distintos países e instituciones no siempre incluyen lineamientos que sean aplicables a las especificidades de las IAs. Por ello, se requiere crear o actualizar las dichas políticas para que indiquen cómo, cuándo y con qué finalidad se puede introducir datos reales en las interacciones con LLM u otro tipo de herramienta de IA (Javed & Arooj, 2023; Jebreen et al., 2024).

5.5 Integridad académica y riesgo de uso indebido en tareas y evaluaciones

Un reto adicional es la potencial utilización de los LLMs para elaborar respuestas automáticas en exámenes o trabajos académicos. Con la facilidad que ofrece un chatbot avanzado, algunos estudiantes podrían limitarse a “copiar y pegar” sin procesar la información, afectando la genuina adquisición de competencias clínicas. Asimismo, el modelo puede generar referencias falsas o poco verificables (Sallam, 2023; Shusterman et al., 2024; Yan et al., 2023), dificultando el rastreo de plagio y la detección de autoría.

Las instituciones, entonces, han de definir con precisión los límites de un uso aceptable: por ejemplo, autorizar el apoyo en LLMs para el repaso teórico, pero prohibirlo en la realización de evaluaciones sumativas (Javed & Arooj, 2023; Pan & Ni, 2024). El desarrollo de software que identifique señales de contenido generado por IA puede ser una respuesta, si bien se han reportado limitaciones en su precisión (#36, #40).

5.6 Falta de formación en “ingeniería de prompts” para docentes y estudiantes

Para que los LLMs funcionen como auxiliares verdaderamente efectivos, para cualquier tipo de tarea y en cualquier entorno, es crucial comprender cómo formular y validar prompts adecuados (Shah et al., 2024; Shusterman et al., 2024; Waldock et al., 2024). Una pregunta mal planteada puede llevar a respuestas superficiales, ambiguas, irrelevantes o francamente erradas, mientras que un prompt bien diseñado orienta la respuesta hacia la información requerida y limita la probabilidad de que contenga datos sesgados o cuestionables. Sin embargo, muchos programas de educación médica todavía no incluyen asignaturas o módulos dedicados a la ingeniería de prompts, en su programa curricular, dejando a docentes y estudiantes sin la preparación específica necesaria (Mannekote et al., 2024; Pears & Konstantinidis, 2024).

Impulsar talleres o seminarios que abarquen técnicas de ingeniería de prompts, verificación crítica de la respuesta y estrategias de retroalimentación y refinamiento de las consultas resultará determinante para reducir la brecha entre la capacidad potencial de la IA y su aplicación real en contextos formativos (Raiaan et al., 2023; Shusterman et al., 2024; Waldock et al., 2024).

5.7 Falta de un corpus validado y adaptado para la formación médica

La disponibilidad de fuentes biomédicas actualizadas, revisadas y culturalmente diversas para afinar (fine-tune) modelos es escasa (Jiao et al., 2024; J. Liu et al., 2024; Raiaan et al., 2023). Muchos LLMs se entrenan con datos de la web de forma general, lo cual no siempre cubre las necesidades curriculares de la educación médica, ni garantiza la

precisión de los contenidos. Sin un material cuidadosamente curado, es posible que el modelo caiga en sesgos u otro tipo de errores y que no contemple la información más actualizada, en algún dominio específico.

Adicionalmente, la construcción de bibliotecas médicas especializadas para el entrenamiento de modelos de IA (con datos de guías clínicas, consensos de sociedades científicas, estudios controlados, etc.) ofrece la posibilidad de entrenar LLMs que respondan con más rigor a las preguntas académicas (Jiang et al., 2024; Moreno & Bitterman, 2024). Esto implica procesos continuos y complejos de validación y filtrado de fuentes, pero constituye una vía segura para elevar la fiabilidad.

5.8 Marco regulatorio, lineamientos institucionales y responsabilidad profesional

El papel de las entidades reguladoras es clave para establecer criterios de uso que aseguren la calidad de la educación y, en última instancia, la seguridad del paciente (Jiao et al., 2024; Moreno & Bitterman, 2024; Yan et al., 2023). A nivel académico, se requiere que las universidades y facultades de medicina definan políticas claras sobre:

- Integración curricular: determinar en qué asignaturas, con qué objetivos de aprendizaje y bajo que metodologías se usará la IA.
- Responsabilidad profesional: especificar cuándo y cómo se requerirá de supervisión obligatoria de un docente o tutor, al usar LLMs en determinadas actividades formativas.
- Propiedad intelectual y ética: regular los criterios para uso de IA en la producción de material académico, teniendo en cuenta aspectos relacionados con la autoría de los trabajos y controlar el plagio.
- Protección de datos y confidencialidad: cumplir leyes locales e internacionales al introducir información sensible a una interacción con una IA (Jebreen et al., 2024; J. Liu et al., 2024; Wang et al., 2025).

La existencia de comités éticos o subcomisiones en las que participen expertos en IA, educadores, estudiantes y profesionales clínicos podría ayudar a actualizar y adaptar los lineamientos de manera continua (Jiao et al., 2024; Wang et al., 2025).

5.9 Estrategias de mitigación de riesgos y construcción de confianza en la adopción de LLMs

Finalmente, se proponen una serie de lineamientos prácticos para atenuar los riesgos descritos:

- Enfoque “human-in-the-loop”: asegurar la supervisión docente o de un clínico experto en cualquier decisión crítica generada por el LLM (J. Liu et al., 2024; Omiye J. A. & 2024, 2024; Shah et al., 2024).
- Capacitación sistemática: incluir módulos de formación en IA y LLMs para docentes y estudiantes, explicando su funcionamiento, aspectos de ingeniería de prompts, riesgo de sesgos, alucinaciones y otros errores y la necesidad validación de respuestas, y bases éticas del uso de IA en temas relacionados con la salud (Raiaan et al., 2023; Shusterman et al., 2024; Waldock et al., 2024).
- Protocolos de validación interna: someter las respuestas del LLM a comprobaciones aleatorias con fuentes oficiales o con expertos, especialmente cuando se trate de escenarios de diagnóstico o manejo terapéutico (Jiao et al., 2024; Moreno & Bitterman, 2024).
- Ajuste fino con material de referencia verificado: en la medida de lo posible, entrenar (o refinar) los modelos con bases de datos médicas de alta calidad, revisados por pares y contextualizados a la realidad local (Jiang et al., 2024; J. Liu et al., 2024; Raiaan et al., 2023).
- Regulación y transparencia: definir guías claras sobre su uso en evaluaciones, establecer normativas de integridad académica y diseñar mecanismos para reportar y corregir errores.
- Observatorios o comités de ética: crear instancias encargadas de monitorear la evolución de estas herramientas y emitir recomendaciones periódicas ante nuevos hallazgos o problemáticas (Jebreen et al., 2024; Jiao et al., 2024; Wang et al., 2025).

En conclusión, el uso de LLMs en la educación médica puede potenciar la adquisición de competencias clínicas, la personalización del aprendizaje y la formación metacognitiva. No obstante, para que su adopción sea segura y socialmente responsable, resulta imperativo abordar de manera integral los desafíos éticos (sesgos, integridad académica, privacidad), técnicos (fiabilidad, reproducibilidad, corpus validados) y regulatorios (marcos institucionales, lineamientos de calidad y protección de datos). Por encima de todo, la confianza en la adopción de LLMs vendrá dada por la combinación de transparencia en la metodología, rigor en la supervisión, y cultura de la autoevaluación por parte de todos los actores: estudiantes, docentes, directivos y reguladores del área de la salud. Solo así se logrará una implementación que complemente los métodos pedagógicos tradicionales, impulse la innovación, y forme profesionales de la salud críticos, reflexivos y preparados para enfrentar los requerimientos de la medicina contemporánea.

6 Síntesis Integrativa y perspectiva adoptada en este proyecto

Esta última sección ofrece una visión de conjunto de los aportes descritos a lo largo del marco teórico, resaltando cómo convergen en la necesidad de proponer y sustentar un

proyecto de investigación que aborde la adopción de LLMs en la educación médica, con miras a mejorar la integración de ciencias básicas y clínicas y fortalecer el razonamiento clínico.

6.1 Articulación de los hallazgos conceptuales y empíricos

Las secciones previas han evidenciado, en primer lugar, la importancia de construir un razonamiento clínico sólido, entendiendo que este no se limita a la mera asimilación de datos, sino que implica integrar información biomédica y clínica de modo flexible y contextual (Durning et al., 2024; Guzmán-Valdivia-Gómez et al., 2022; Kulasegaram et al., 2015). Al mismo tiempo, la literatura coincide en señalar las barreras persistentes para articular las ciencias básicas con las ciencias clínicas, lo que repercute negativamente en la formación de médicos capaces de vincular fundamentos fisiopatológicos con la práctica asistencial (Delavari et al., 2024; Kassirer, 2010; Parodis et al., 2021).

En paralelo, y gracias a los avances en la IA y, particularmente, a los LLMs, se presentan nuevas oportunidades para personalizar el aprendizaje, aumentar la disponibilidad de contenidos y reforzar el razonamiento clínico (Choi & Abdirayimov, 2024; Dong et al., 2024; Jiang et al., 2024; J. Liu et al., 2024; Lucas et al., 2024; Mannekote et al., 2024; McCoy et al., 2024; Nerella et al., 2024; Pears & Konstantinidis, 2024; Sallam, 2023; Zarei, Eftekhari Mamaghani, et al., 2024; Zarei, Zarei, et al., 2024). No obstante, la adopción de estas herramientas también ha generado nuevas problemáticas éticas y técnicas, como son los sesgos en los datos, la posibilidad de “alucinaciones” e información incorrecta, las implicaciones de la privacidad y la equidad en el acceso tecnológico (Choi & Abdirayimov, 2024; Jiao et al., 2024; Yan et al., 2023).

Con todo, se observa una sinergia potencial: la implementación de LLMs orientados pedagógicamente podría incidir, positivamente, en esa integración de saberes, posibilitando la simulación de casos complejos, la retroalimentación personalizada e, incluso, la enseñanza de competencias en resolución de problemas multidisciplinares. De allí la pertinencia de un enfoque investigativo que indague cómo diseñar, evaluar y regular el uso de LLMs en la enseñanza médica, con criterios pedagógicos claros.

6.2 Limitaciones en la literatura actual y vacíos de conocimiento

Pese a los estudios en aumento, persisten importantes brechas:

Escasez de evidencia sistemática: aunque existen investigaciones que describen el uso de LLMs en la educación médica, son pocas las que evalúan de manera sistemática su efecto sobre competencias específicas, como la habilidad para integrar ciencias básicas y clínicas (Jiao et al., 2024; J. Liu et al., 2024; Yan et al., 2023).

Insuficiente análisis sobre el razonamiento clínico: la mayoría de los estudios se centra en la utilidad inmediata de los LLMs para contestar preguntas o generar resúmenes, sin un enfoque profundo en medir cómo influyen en la construcción del razonamiento clínico y la toma de decisiones (Guzmán-Valdivia-Gómez et al., 2022; Kanjee et al., 2023; Lucas et al., 2024; Sallam, 2023; Zarei, Eftekhari Mamaghani, et al., 2024; Zarei, Zarei, et al., 2024).

Carencia de lineamientos para su implementación responsable: hay ausencia de guías que describan qué salvaguardas, mecanismos de supervisión y procesos formativos deben adoptarse para minimizar riesgos de sesgos y usos indebidos (Jiao et al., 2024; Moreno & Bitterman, 2024).

Escasa formación docente en IA: la literatura subraya la necesidad de que el personal académico domine, al menos, conceptos básicos de “ingeniería de prompts”, verificación de respuestas y supervisión del aprendizaje asistido por IA, algo aún poco sistematizado en los planes de estudio (Shah et al., 2024; Shusterman et al., 2024; Waldock et al., 2024).

6.3 Pertinencia y justificación del marco teórico para la presente investigación

El marco teórico expuesto (focalizado en la confluencia de razonamiento clínico, brecha entre ciencias básicas y clínicas, y el potencial y las limitaciones de los LLMs) resulta esencial para:

- Delimitar con claridad las dimensiones pedagógicas, cognitivas y éticas involucradas en la adopción de LLMs.
- Interpretar datos de modo coherente, reconociendo que los efectos en el razonamiento clínico no pueden analizarse al margen de factores contextuales (como la disponibilidad de recursos o la formación del docente).
- Diseñar estrategias metodológicas que contemplen tanto la evaluación del aprendizaje con LLMs, como la validez de la información, los sesgos y la viabilidad de la integración curricular.

Asimismo, el marco teórico proporciona el sustento conceptual para abordar las preguntas clave de la investigación:

- ¿De qué manera podría el uso de LLMs solventar o mitigar la brecha entre ciencias básicas y clínicas?
- ¿Cuáles son los riesgos o factores críticos a la hora de implementar LLMs como herramienta formativa?
- ¿Qué directrices y competencias específicas se requieren para garantizar un aprovechamiento ético y efectivo?

Al situar estas cuestiones dentro del entramado teórico ya expuesto, se reafirma la relevancia de explorar no solo el desempeño de los estudiantes ante LLMs, sino también los procesos de diseño curricular y las políticas institucionales.

En conclusión, la revisión conceptual y empírica sugiere un panorama complejo donde convergen oportunidades y desafíos. El presente proyecto, al apoyarse en las bases teóricas establecidas, queda justificado en su afán de investigar y proponer enfoques pedagógicos para la adopción responsable de LLMs en la educación médica. Con ello, se aspira a contribuir a la formación de médicos con un razonamiento clínico integral, respaldado por la solidez científica y la sensibilidad necesaria para enfrentar las demandas de la medicina actual.

MARCO METODOLÓGICO PARA LA REVISIÓN SISTEMÁTICA

1 Tipo de diseño y enfoque de investigación

1.1 Temática general

El presente estudio adopta un diseño no experimental de tipo documental, estructurado bajo los lineamientos de una revisión sistemática. Este diseño es adecuado para recopilar, evaluar y sintetizar información de estudios primarios previamente publicados, con el objetivo de responder a las preguntas de investigación y cumplir los objetivos planteados. El enfoque es descriptivo, ya que busca analizar y categorizar el estado actual del conocimiento sobre el impacto de los Modelos de Lenguaje a Gran Escala (LLMs) en la integración de saberes básicos y clínicos en la educación médica, con el objetivo de mejorar el desarrollo de competencias en y relacionadas con razonamiento clínico.

1.2 Objetivos

- Cuantificar el impacto sobre las competencias en razonamiento clínico de los estudiantes, tras haber participado en experiencias educativas asistidas con LLMs.
- Analizar la integración conceptual interdisciplinar facilitada por LLMs entre ciencias básicas y clínicas.
- Evaluar la transferencia percibida del razonamiento clínico desarrollado con la asistencia de LLMs a situaciones de práctica real.
- Caracterizar las percepciones de los estudiantes y docentes sobre los beneficios, limitaciones y barreras asociadas al uso de LLMs como herramienta educativa en la formación médica.

1.3 Pregunta de investigación

¿De qué manera la incorporación de experiencias educativas asistidas por LLMs, en comparación con métodos pedagógicos tradicionales, impacta en la integración conceptual interdisciplinar entre ciencias básicas y clínicas, así como en el desarrollo y en la transferencia práctica del razonamiento clínico, en estudiantes de medicina?

1.3.1 Estructuración según el enfoque PICO:

- Población (P):
  - Estudiantes de medicina, en etapas de formación en ciencias clínicas.
- Intervención (I):
  - Experiencias educativas diseñadas y asistidas con LLMs.
- Comparación (C):
  - Métodos pedagógicos y estrategias didácticas tradicionales (sin apoyo de LLMs).
- Resultados (O):
  - Integración conceptual interdisciplinar entre ciencias básicas y clínicas.
  - Desarrollo de competencias en razonamiento clínico.
  - Transferencia percibida del razonamiento clínico a la práctica real.
  - Percepciones de estudiantes y docentes sobre beneficios, limitaciones y barreras del uso de LLMs.

1.4 Apoyo metodológico de la IA

Durante la fase de planeación y diseño metodológico del presente estudio, se implementarán GPTs (en ChatGPT) o Gems (en Gemini) específicos, entrenados por el autor, para orientar algunos aspectos metodológicos, como la adherencia a los criterios de calidad AMSTAR-2 para revisiones sistemáticas durante la construcción del protocolo de estudio y el refinamiento de los objetivos y de los criterios de inclusión y exclusión para la selección de los estudios revisados.

2 Participantes o fuentes de información

2.1 Unidad de análisis

La unidad de análisis consiste en estudios primarios publicados en revistas científicas indexadas que aborden el uso de LLMs en la educación médica o contextos relacionados con el aprendizaje del razonamiento clínico. Se incluirán artículos que no hayan sido revisados por pares (*preprints*) si cumplen con un alto rigor metodológico y tienen bajo riesgo de sesgos, acorde a las evaluaciones de calidad mencionadas adelante. Se considerarán las revisiones sistemáticas únicamente para su uso en el marco teórico y contextual, pero no para la síntesis de datos primaria en este estudio.

2.1.1 Selección de artículos

i. Etapa 1: búsqueda y recolección inicial:

- Diseñar la estrategia de búsqueda con términos clave relacionados (ej. MeSH) y libres, así como operadores booleanos, representativa de la pregunta de investigación y los objetivos definidos en la Sección 1.
- Ejecutar una búsqueda preliminar para elegir las bases de datos relevantes.
- Ejecutar las búsquedas formales, en las bases de datos seleccionadas, como se describe más adelante.
- Extraer los resultados de búsqueda, organizarlos en una tabla (base de datos) que incluya códigos de identificación propios para cada uno, los datos básicos de cada trabajo (título, autores, fecha de publicación, identificador único, sitio de publicación, idioma), los enlaces de acceso al texto y sus resúmenes (*abstracts*), eliminando duplicados.

ii. Etapa 2: selección preliminar (cribado) por títulos y resúmenes:

- Revisar los títulos y resúmenes de todos los estudios identificados.
- Aplicar una lista de chequeo basada en los criterios de inclusión y exclusión, para filtrar los trabajos potencialmente relevantes (ver Anexos, Tabla suplementaria 1. Herramienta para la selección de artículos para la inclusión final en la revisión sistemática):
  - Criterios de inclusión:
    - Publicaciones entre enero de 2023 y marzo de 2025.
      - Se eligió enero de 2023 como punto de inicio, por el alto impacto que tuvo la publicación de ChatGPT 3.0 de OpenAI en noviembre de 2022.
    - Artículos de acceso libre o con acceso posible por convenio institucional.
    - Título, resumen y texto completo disponible en idiomas inglés o español.
    - Estudio primario con datos relevantes.
    - Artículos publicados en revistas indexadas o *preprints* que tengan un alto rigor metodológico y bajo riesgo de sesgos acorde a:
      - Una clara descripción metodológica, que permita identificar sus objetivos o propósitos, sus participantes, intervenciones, comparaciones, estrategias de recolección de datos y resultados y su metodología de análisis de estos, con suficiente detalle para entenderla y evaluar su rigor y posibles sesgos.
    - Enfoque en uso de LLMs en educación médica y que hagan una evaluación (cuantitativa o cualitativa) de al menos 2 de los siguientes criterios (representativos de los objetivos descritos en la Sección 1.1):
      - El impacto en el desempeño o en el desarrollo de competencias relacionadas con el razonamiento clínico
      - La integración conceptual interdisciplinar básica-clínica
      - La transferencia de saberes, habilidades y competencias adquiridas a escenarios prácticos reales.
      - Percepciones de los estudiantes o docentes sobre los beneficios, limitaciones o barreras asociadas al uso de LLMs en la educación médica.

- Criterios de exclusión:
  - Acceso restringido.
  - Publicaciones sin datos primarios.
    - Revisiones literarias, sistemáticas, metaanálisis u otros estudios sin datos primarios (los cuales, sin embargo, podrán ser tenidos en cuenta como referentes teóricos para contrastar sus resultados con los de la presente revisión).
  - Estudios fuera del ámbito educativo médico (aplicables a la formación de otro tipo de profesionales, enfocados en la implementación en la práctica profesional y no en la educación o enfocados en otros aspectos relacionados al uso de LLMs distintos a su impacto en la educación médica).
  - Estudios que solo evalúen el desempeño de los LLMs en pruebas o exámenes, pero no en el impacto sobre los estudiantes al usar los LLMs en su proceso educativo.
  - Estudios que valoren tecnologías de IA o de otro tipo, distintas a LLMs.
  - Estudios que solo traten de forma superficial o tangencial los temas relacionados con los objetivos de este estudio (ver Sección 1.1)

iii. Etapa 3: Evaluación de textos completos:

- Recuperar y leer el texto completo de los estudios preseleccionados en la Etapa 2.
- Aplicar nuevamente la Tabla suplementaria 1. Herramienta para la selección de artículos para la inclusión final en la revisión sistemática, para:
  - Reaplicar los criterios de inclusión y exclusión de la Etapa 2, para confirmar la relevancia y el enfoque educativo de los artículos.
  - Excluir aquellos estudios que, al analizar el texto completo, no cumplan satisfactoriamente con los criterios de inclusión o cumplan con alguno de los de exclusión.
  - Hacer una evaluación formal de calidad a todos los artículos hasta el momento no descartados, con herramientas específicas, según el tipo de estudio:
    - RoB2 para ensayos clínicos o estudios experimentales aleatorizados.
    - Herramientas específicas de JBI para los estudios de intervención no aleatorizados (cuasi experimentales).
    - Herramientas específicas de JBI para los distintos estudios observacionales.
    - Herramientas específicas de JBI para los distintos estudios cualitativos.
    - MMAT para los estudios mixtos.
  - Documentar y puntuar cada estudio según estos instrumentos, descartando aquellos que no alcancen un umbral de calidad que implique bajo riesgo de sesgo o de errores metodológicos.

iv. Etapa 4: inclusión final:

- Integrar en la revisión sistemática los estudios que hayan pasado satisfactoriamente las Etapas 1-3.

- Documentar todo el proceso utilizando el diagrama PRISMA, detallando el número de estudios en cada etapa y las razones de exclusión.

2.2 Procedimiento de selección

Se hará una selección de estudios que cumplan con los criterios mencionados, mediante una búsqueda estructurada en bases de datos científicas seleccionadas y procesos de selección minuciosos. El proceso seguirá el modelo PRISMA para garantizar transparencia y reproducibilidad.

3 Instrumentos y técnicas de recolección de información

3.1 Instrumentos

3.1.1 Bases de datos:

Se hizo una búsqueda preliminar exploratoria en bases de datos disponibles para acceso libre o a través de los convenios con la Universidad Icesi. Se documentó que las bases de datos accesibles con un alto número de artículos relevantes, y en las que se hará la búsqueda formal posterior, son:

- PubMed
- ArXiv (se incluirán preprints relevantes de ArXiv, sometidos a cribado de calidad metodológica idéntico al de artículos publicados)
- ERIC
- EBSCO

3.1.2 Herramientas de gestión:

- Mendeley: para la gestión de referencias bibliográficas.
- Plantillas PRISMA: se seguirá el modelo PRISMA (que incluye la identificación, selección, elegibilidad y finalmente la inclusión de los estudios en la revisión), documentando cada etapa con el diagrama PRISMA correspondiente, para garantizar transparencia y reproducibilidad.
- Lista de chequeo para evaluación de criterios de inclusión y exclusión: se construirá una lista de chequeo con todos los criterios de inclusión y exclusión definidos anteriormente, para la selección preliminar de los estudios.
- Criterios de la herramienta AMSTAR: sus lineamientos serán considerados durante el diseño y desarrollo de esta misma revisión sistemática.
- Herramientas de evaluación de calidad de los estudios: implementación de criterios según las herramientas RoB2, JBI y MMAT (según corresponda, por tipo de estudio), para hacer una evaluación objetiva y sistemática de los artículos a incluirse en la revisión sistemática.
- ChatGPT y Gemini: serán utilizados para optimizar procesos de gestión de información y contenidos, revisar y optimizar la calidad de un texto escrito, actuar como asistente general y consultor durante la producción del trabajo de investigación.

Se garantizará que todas las contribuciones sean revisadas críticamente por el autor del proyecto de investigación y que no comprometan la objetividad e integridad del análisis. Adicionalmente, se garantizará que toda IA que sea empleada, se usará como una herramienta que asista y complemente el trabajo del investigador, más no como un sustituto de este.

3.2 Técnicas

3.2.1 Construcción de estrategias de búsqueda:

Utilización de operadores booleanos (AND, OR, NOT) y términos clave relacionados con LLMs, razonamiento clínico e integración de saberes.

Ejemplo de estrategia de búsqueda en PubMed: ("large language models"[MeSH Terms] OR "generative AI") AND ("clinical reasoning" OR "medical education") AND ("integration of knowledge").

3.2.2 Prompt de entrenamiento de GPT y Gem asistente metodológico de revisión sistemática:

Para el uso de IA en el proceso de diseño y desarrollo de la presente revisión, se diseñó un *prompt* (ver Anexos, Prompt 1) para para transformar a la IA en un asistente y tutor académico experto, especializado en guiar el desarrollo de revisiones sistemáticas y metaanálisis para tesis de maestría o doctorado, mediante el entrenamiento de un GPT (en ChatGPT) y una Gem (en Gemini) llamado “Asistente y tutor en el desarrollo de una Revisión Sistemática/Metaanálisis para tesis de grado”. Se siguió una estructura de diseño adaptada de la metodología RITA (*Role, Instruction, Task, Action*), basada en la definición de un Rol, Contexto de uso, Instrucciones y tareas, Formato y estilo de salida, para el diseño del *prompt* y se realizaron múltiples ajustes durante pruebas piloto, para adaptar el tipo de respuestas, su especificidad y objetividad y minimizar la probabilidad de errores fácticos en lo que ella ha propuesto.

4 Procedimientos y condiciones para la recolección de información

4.1 Pasos detallados

4.1.1 Búsqueda y selección:

- Se harán búsquedas exhaustivas en las bases de datos, con las estrategias antes descritas.
- Registro del número total de resultados obtenidos y eliminación de duplicados, siguiendo el protocolo PRISMA.

4.1.2 Aplicación de criterios de inclusión y exclusión:

- Revisión inicial: evaluación de títulos y resúmenes para identificar estudios potencialmente relevantes y eliminar duplicados, como resultados de las búsquedas.
- Revisión a texto completo: a los artículos antes identificados, se les hará una evaluación con la lista de chequeo de criterios de inclusión y exclusión, como se describió antes.

4.1.3 Evaluación de calidad:

- A los artículos que pasen la preselección con la lista de chequeo de criterios de inclusión y exclusión, se les hará una nueva revisión empleando las herramientas RoB2, JBI o MMAT para filtrarlos e identificar los estudios de alta calidad y excluir aquellos con riesgos o problemas metodológicos significativos.
- Revisión de expertos: se dispondrá de asesoría y supervisión constante del tutor de tesis de grado, respaldado y asignado por la Universidad de Icesi, para asegurarse que todos los procesos desde el diseño y hasta el desarrollo y publicación de resultados se hagan de forma idónea.

5 Variables, operacionalización e instrumentos de análisis

5.1.1 Bloque A: Datos de Identificación y características del estudio

- **A.1. Código propio del artículo**

  Definición: Es el código que se asigna a cada artículo, una vez es aprobado para su inclusión en el estudio.

- **A.2. Título del artículo.**

  Definición: Es el título completo del estudio.

- **A.3. Autor/es.**

  Definición: La lista de autores principales (seguido de et al. si aplica) que desarrollaron el estudio.

- **A.4. Año de publicación.**

  Definición: Es el año en que aparece oficialmente publicado.

- **A.5. ID del artículo publicado.**

  Definición: Es el identificador único del artículo, que puede ser el DOI, PMID u otros.

- **A.6. Tipo de diseño del estudio.**

Definición: Hace referencia al diseño metodológico: experimental aleatorizado, cuasiexperimental, observacional de cohortes, observacional de casos y controles, observacional transversal, etc.).

Códigos/Etiquetas para el registro de datos:

- 0 = Otro.
- 1 = Experimental aleatorizado.
- 2 = Estudio de intervención no aleatorizado (cuasiexperimental).
- 3 = Observacional: cohortes.
- 4 = Observacional: casos-controles.
- 5 = Observacional: transversal.
- 6 = Observacional: prevalencia/descriptivo.

- **A.7. Enfoque principal del estudio.**

  Definición: Hace referencia al enfoque primario de tipo cuantitativo, cualitativo o mixto.

  Códigos/Etiquetas para el registro de datos:

  - 1 = Cuantitativo.
  - 2 = Cualitativo.
  - 3 = Mixto.

- **A.8. Objetivos de investigación.**

  Definición: Es el listado de objetivos definidos en el estudio (si aplica)

- **A.9. Contexto geográfico del grupo de investigación.**

  Definición: Hace referencia al país o países en que trabaja el grupo investigador.

- **A.10. Contexto institucional del grupo de investigación.**

  Definición: Hace referencia a la o las instituciones en que trabaja el grupo investigador.

- **A.11. Tamaño muestral del estudio.**

  Definición: El número total de participantes en el estudio.

- **A.12. Tamaño del grupo de intervención (si aplica).**

Definición: El número de participantes asignados al grupo de intervención que “es expuesto” al uso del LLM en evaluación.

- **A.13. Tamaño del grupo de control (si aplica).**

  Definición: El número de participantes asignados al grupo de control que “no es expuesto” al uso del LLM en evaluación.

- **A.14. Periodo (meses y/o años) en que se aplicó o hizo el estudio.**

  Definición: Hace referencia al periodo temporal durante el cual se hizo el estudio de intervención u observacional y la recolección de los datos.

5.1.2 Bloque B: Variables principales - Cualitativas

- **B.1. Percepción de los estudiantes sobre el uso de LLMs en la educación**

  Definición: opiniones generales de los estudiantes respecto al uso de herramientas de IA generativas tipo LLMs en su formación, a partir de sus experiencias directas o indirectas con estas herramientas.

  Subvariables (indicadores):

  - **B.1.1. Opinión general (positiva, negativa, neutra o mixta, no reportada)**

    Definición (e indicador): calificación global del sentir estudiantil hacia los LLMs en su educación.

  - **B.1.2. Usos de mayor utilidad**

    Definición (e indicador): aplicaciones específicas de los LLMs consideradas más provechosas por los estudiantes.

  - **B.1.3. Beneficios más destacados**

    Definición (e indicador): ventajas o aspectos positivos principales del uso de LLMs mencionados por estudiantes.

  - **B.1.4. Preocupaciones por interferencia en el desarrollo de saberes, habilidades o competencias**

    Definición (e indicador): inquietudes estudiantiles sobre cómo los LLMs podrían obstaculizar su aprendizaje o la adquisición de destrezas.

  - **B.1.5. Preocupaciones por aspectos éticos**

    Definición (e indicador): inquietudes estudiantiles vinculadas a dilemas morales o de integridad académica (plagio, sesgos, privacidad).

- **B.1.6. Preocupaciones por aspectos relacionados al acceso a la tecnología**

  Definición (e indicador): inquietudes estudiantiles sobre barreras para disponer o usar la tecnología (costo, equidad, conectividad).

- **B.1.7. Otras preocupaciones, problemas o limitaciones percibidas**

  Definición (e indicador): cualquier otra dificultad, desventaja o inquietud expresada por estudiantes no categorizada previamente.

Códigos/Etiquetas para el registro de datos:

- Para B.1.1 (Opinión general): registrar usando únicamente una de las siguientes etiquetas: "positiva"; "negativa"; "neutra o mixta"; "no reportada".
- Para B.1.2 a B.1.7 (Usos, Beneficios, Preocupaciones): valorar los temas específicos o listas de ítems mencionados en el texto y registrar tomando como etiqueta o código el propio tema/ítem encontrado. Si se identifican múltiples temas/ítems distintos para la misma subvariable, separarlos con punto y coma (";"). (Ejemplo para B.1.2: "generación de resúmenes; resolución de dudas"; Ejemplo para B.1.5: "plagio; sesgos del modelo; privacidad de datos").

- **B.2. Percepción de los docentes sobre el uso de LLMs en la educación**

Definición: opiniones de los docentes respecto al uso de LLMs como herramientas complementarias en la enseñanza, a partir de sus experiencias directas o indirectas con estas herramientas.

Subvariables (indicadores):

- **B.2.1. Opinión general (“positiva”, “negativa”, “neutra o mixta”, “no reportada”)**

  Definición (e indicador): valoración global (positiva/negativa/neutra/mixta) del docente sobre los LLMs en la enseñanza.

- **B.2.2. Usos para los que se percibe mayor utilidad**

  Definición (e indicador): tareas o funciones docentes donde los LLMs son vistos como más beneficiosos o aplicables.

- **B.2.3. Beneficios más destacados para esos usos**

  Definición (e indicador): ventajas o aspectos positivos específicos que los docentes identifican al aplicar LLMs en las áreas de mayor utilidad.

- **B.2.4. Preocupaciones por interferencia en el desarrollo de saberes, habilidades o competencias**

  Definición (e indicador): temores docentes sobre cómo los LLMs podrían obstaculizar el aprendizaje o la adquisición de competencias estudiantiles.

- **B.2.5. Preocupaciones por aspectos éticos**

  Definición (e indicador): inquietudes de los docentes vinculadas a la integridad académica, privacidad, sesgos u otros dilemas morales del uso de LLMs.

- **B.2.6. Preocupaciones por aspectos relacionados al acceso a la tecnología**

  Definición (e indicador): dificultades o barreras percibidas por los docentes sobre la disponibilidad, equidad o requerimientos técnicos para usar LLMs.

- **B.2.7. Preocupaciones asociadas a la estabilidad o seguridad laboral**

  Definición (e indicador): inquietudes de los docentes sobre cómo la adopción de LLMs podría afectar su rol profesional o estabilidad en el empleo.

- **B.2.8. Otras preocupaciones, problemas o limitaciones percibidas**

  Definición (e indicador): cualquier otra reserva, dificultad o desventaja mencionada por los docentes sobre el uso de LLMs no cubierta antes.

Códigos/Etiquetas para el registro de datos:

- Para B.2.1 (Opinión general): registrar usando únicamente una de las siguientes etiquetas: "positiva"; "negativa"; "neutra o mixta"; "no reportada".
- Para B.2.2 a B.2.8 (Usos, Beneficios, Preocupaciones): valorar los temas específicos o listas de ítems mencionados en el texto y registrar tomando como etiqueta o código el propio tema/ítem encontrado. Si se identifican múltiples temas/ítems distintos para la misma subvariable, separarlos con punto y coma (";"). (Ejemplo para B.2.2: "generación de material didáctico; calificación automatizada"; Ejemplo para B.2.5: "integridad académica; sobrecarga de información").

- **B.3. Evidencias de integración conceptual entre ciencias básicas y clínicas**

Definición: evidencia cualitativa de cómo los estudiantes asimilan y conectan conceptos de ciencias biomédicas básicas con los de la práctica clínica, después de utilizar o a través del uso de LLMs.

Subvariables (indicadores):

- **B.3.1. Descripciones del efecto en la coherencia conceptual entre la teoría y práctica clínica**

  Definición (e indicador): relatos o ejemplos cualitativos sobre cómo el uso de LLMs mejora (o afecta) la conexión lógica entre conocimientos básicos y clínicos.

- **B.3.2. Evidencias del impacto en razonamiento integrado en actividades educativas.**

  Definición (e indicador): manifestaciones concretas, observadas en tareas o evaluaciones, de la aplicación conjunta de saberes básicos y clínicos facilitada por LLMs.

Códigos/Etiquetas para el registro de datos:

- Para B.3.1 y B.3.2: valorar los temas específicos o listas de ítems mencionados en el texto y registrar tomando como etiqueta o código el propio tema/ítem encontrado. Si se identifican múltiples elementos distintos para la misma subvariable, separarlos con punto y coma (";"). (Ejemplo para B.3.1: "mejores argumentos para propuesta diagnóstica; mayor cuidado con interacciones farmacológicas"; Ejemplo para B.3.2: "uso adecuado de epidemiología para respaldar argumentos; previsión de efectos sistémicos en casos simulados").

- **B.4. Formas de uso de LLMs por parte de los docentes**

Definición: caracterización de los usos principales que le dan los docentes a las IA generativas tipo LLM.

Subvariables (indicadores):

- **B.4.1. Diseño macro o micro curricular**

  Definición (e indicador): empleo de LLMs para planificar o estructurar cursos, módulos, unidades temáticas o el currículo general.

- **B.4.2. Diseño de actividades académicas**

  Definición (e indicador): utilización de LLMs para crear o refinar tareas, ejercicios, casos u otras experiencias de aprendizaje específicas.

- **B.4.3. Creación de preguntas para cualquier actividad**

  Definición (e indicador): uso de LLMs para generar ítems, enunciados o interrogantes para evaluaciones, talleres, discusiones, etc.

- **B.4.4. Revisión de calidad de preguntas para cualquier actividad**

  Definición (e indicador): aplicación de LLMs para evaluar la claridad, pertinencia, dificultad u otros aspectos de calidad de ítems o preguntas ya existentes.

- **B.4.5. Calificación de actividades distintas a exámenes**

  Definición (e indicador): empleo de LLMs para evaluar o asignar puntuaciones a trabajos, tareas, ensayos, informes u otras actividades formativas (no exámenes formales).

- **B.4.6. Calificación de exámenes**

  Definición (e indicador): utilización de LLMs en el proceso de corrección o asignación de notas a pruebas o exámenes formales.

- **B.4.7. Retroalimentación general sobre el desempeño**

  Definición (e indicador): uso de LLMs para proporcionar comentarios globales o sumarios sobre el rendimiento de los estudiantes en una actividad o periodo.

- **B.4.8. Retroalimentación específica sobre el desempeño**

  Definición (e indicador): aplicación de LLMs para generar comentarios detallados, personalizados o puntuales sobre aspectos concretos del trabajo o desempeño estudiantil.

- **B.4.9. Generación de contenido o material para actividades específicas (presentaciones, resúmenes, imágenes, etc.)**

  Definición (e indicador): empleo de LLMs para crear recursos educativos como textos, diapositivas, resúmenes, gráficos u otros materiales de apoyo.

- **B.4.10. Otros (especificar)**

  Definición (e indicador): cualquier otra aplicación de LLMs por parte de los docentes en su labor educativa no listada previamente.

Códigos/Etiquetas para el registro de datos:

- Para B.4.1 a B.4.10: valorar los temas específicos o listas de ítems mencionados en el texto para cada categoría y registrar tomando como etiqueta o código el propio tema/ítem encontrado. Si se identifican múltiples elementos distintos para la misma subvariable, separarlos con punto y coma (";"). (Ejemplo para B.4.1: "selección de temáticas de estudio; diseño

de unidades temáticas; definición de formas de calificación"; Ejemplo para B.4.10: "redacción de comunicaciones; distribución de grupos de trabajo").

- **B.5. Formas de uso de LLMs por parte de los estudiantes**

  Definición: caracterización de los usos principales que le dan los estudiantes a las IA generativas tipo LLM.

  Subvariables (indicadores):

  - **B.5.1. Resolución de preguntas, tareas, casos clínicos o ejercicios**

    Definición (e indicador): empleo de LLMs por estudiantes para responder interrogantes, completar asignaciones, analizar escenarios clínicos o solucionar problemas propuestos.

  - **B.5.2. Resolución de cuestionarios de evaluación**

    Definición (e indicador): utilización de LLMs por estudiantes durante la realización de pruebas, quiz u otros instrumentos de evaluación formal.

  - **B.5.3. Planificación y gestión de tiempo**

    Definición (e indicador): uso de LLMs por estudiantes como herramienta para organizar horarios, administrar tareas académicas o planificar su estudio.

  - **B.5.4. Síntesis de anotaciones o apuntes de clase**

    Definición (e indicador): aplicación de LLMs por estudiantes para resumir, organizar o reestructurar sus notas personales tomadas durante clases o lecturas.

  - **B.5.5. Generación de contenido académico (presentaciones, resúmenes, imágenes, etc.)**

    Definición (e indicador): empleo de LLMs por estudiantes para crear materiales como borradores de textos, diapositivas, esquemas u otros trabajos académicos.

  - **B.5.6. Autoevaluación o simulaciones de evaluaciones**

    Definición (e indicador): utilización de LLMs por estudiantes para generar o realizar pruebas simuladas, obtener retroalimentación sobre su nivel o practicar para exámenes.

  - **B.5.7. Estudio autodirigido**

Definición (e indicador): uso independiente de LLMs por estudiantes para explorar temas, profundizar conocimientos o guiar su propio aprendizaje fuera de las tareas formales.

- **B.5.8. Otros (especificar)**

  Definición (e indicador): cualquier otra aplicación de LLMs por parte de los estudiantes en su proceso de aprendizaje no listada previamente.

Códigos/Etiquetas para el registro de datos:

- Para B.5.1 a B.5.8: valorar los temas específicos o listas de ítems mencionados en el texto para cada categoría y registrar tomando como etiqueta o código el propio tema/ítem encontrado. Si se identifican múltiples elementos distintos para la misma subvariable, separarlos con punto y coma (";"). (Ejemplo para B.5.3: "creación de cronograma de estudio diario; gestión de recordatorios o pendientes"; Ejemplo para B.5.8: "propuesta de actividades para clases; traducción de textos").

- **B.6. Tipo de actividad educativa específica**

Definición: naturaleza concreta de la actividad de aprendizaje donde se integró el LLM para fomentar la integración básico-clínica o el razonamiento clínico, por parte de los docentes o estudiantes.

Subvariable (indicador):

- **B.6.1. Actividad de los docentes**

  Definición (e indicador): descripción concreta de la tarea o contexto pedagógico específico en el que el docente integró el LLM.

- **B.6.2. Actividad de los estudiantes**

  Definición (e indicador): descripción concreta de la tarea o contexto de aprendizaje específico en el que el estudiante utilizó el LLM.

Códigos/Etiquetas para el registro de datos:

- Para cada caso (B.6.1 y B.6.2): cada actividad debe individualizarse y registrarse de forma clara y concisa, separando con punto y coma (";") cuando existan distintos elementos o aspectos por registrar para cada subvariable. (Ej. Actividades de los docentes: “revisión de historias clínicas; calificación de exámenes”).

5.1.3 Bloque C: Variables principales - Cuantitativas

- **C.1. Rendimiento académico de los estudiantes**

Definición: mediciones objetivas del rendimiento académico de los estudiantes, antes y/o después de la implementación de LLMs o en comparación con grupos control (uso de otro tipo de estrategias didácticas que no incluyen LLMs).

Subvariables (indicadores):

- **C.1.1. Resultados en actividades específicas (teóricas, prácticas, teórico-prácticas)**

  Definición (e indicador): calificaciones o puntajes obtenidos en tareas, exámenes o evaluaciones puntuales del curso/módulo.

- **C.1.2. Resultados en evaluaciones de ciencias biomédicas básicas**

  Definición (e indicador): calificaciones o puntajes en pruebas centradas específicamente en conocimientos de ciencias fundamentales (anatomía, fisio, etc.).

- **C.1.3. Resultados en evaluaciones de ciencias clínicas**

  Definición (e indicador): calificaciones o puntajes en pruebas enfocadas en conocimientos o habilidades de áreas clínicas (medicina interna, pediatría, etc.).

- **C.1.4. Resultados en evaluaciones integradoras básico-clínicas**

  Definición (e indicador): calificaciones o puntajes en pruebas diseñadas para evaluar la aplicación conjunta de ciencias básicas y clínicas.

- **C.1.5. Resultados en evaluaciones estandarizadas (certificación, graduación)**

  Definición (e indicador): puntajes obtenidos en exámenes formales y estandarizados requeridos para avanzar, graduarse o certificarse.

- **C.1.6. Resultados en evaluaciones nacionales o internacionales**

  Definición (e indicador): puntajes en exámenes externos de gran escala, de ámbito nacional o supranacional (ej. MIR, USMLE).

- **C.1.7. Resultados en evaluaciones para especialización**

  Definición (e indicador): puntajes en exámenes específicos utilizados para el ingreso o progreso en programas de posgrado/residencia.

- **C.1.8. Otro indicador objetivo del desempeño (especificar)**

  Definición (e indicador): cualquier otra medida cuantitativa del rendimiento académico no cubierta por las categorías anteriores.

Códigos/Etiquetas para el registro de datos:

- Para cada subvariable aplicable (C.1.1 a C.1.8):
  - Registrar en la casilla de “Valor” los datos numéricos específicos reportados (ej. Media ± DE; Mediana (RI); %; puntaje) para cada grupo (ej. Intervención vs Control) y/o momento (ej. Pre vs Post).
  - Añadir en “Valor” la información estadística comparativa si se reporta (ej. valor de p; tamaño del efecto).
  - Consignar en la casilla de “Valor_complementario/Comentarios” la escala de medición utilizada (ej. "escala 0-100"; "porcentaje correcto") y la naturaleza específica de la actividad o evaluación si se detalla (ej. "examen final teórico"; "evaluación ECOE estación X").
  - Si se reportan múltiples medidas distintas para una misma subvariable (ej. resultados teóricos y prácticos separados), registrarlas todas en “Valor”, separando cada conjunto de datos con punto y coma (";"). (Ejemplo: "Resultados Examen Teórico: Grupo LLM 85 ± 5 vs Control 80 ± 7, p=0.04; Resultados Examen Práctico: Grupo LLM 87 ± 3 vs Control 81 ± 4, p=0.03"). En el ejemplo, la escala "0-100" se consignaría en "Valor_complementario/Comentarios".
  - Si no se encuentran datos cuantitativos específicos para una subvariable, registrar "No reportado" en “Valor”.

- **C.2. Frecuencia de uso de LLMs por parte de los docentes**

  Definición: medición de la intensidad, frecuencia o cantidad de uso de herramientas LLM por parte de los docentes participantes en el estudio o en la intervención descrita.

  Subvariables (indicadores):

  - **C.2.1. Frecuencia de uso (categórica)**

    Definición (e indicador): nivel general de utilización del LLM por docentes reportado en categorías (ej. alta, media, baja).

  - **C.2.2. Periodicidad de uso (categórica)**

    Definición (e indicador): intervalo temporal con que los docentes usan el LLM (ej. diaria, semanal, mensual).

  - **C.2.3. No. Interacciones (numérica)**

Definición (e indicador): cantidad numérica de veces que los docentes interactuaron con el LLM en un periodo determinado.

- **C.2.4. Tiempo de uso promedio (numérica)**

  Definición (e indicador): duración media (en horas, minutos) de las sesiones de uso del LLM por parte de los docentes.

- **C.2.5. Otra medida de intensidad de uso (especificar)**

  Definición (e indicador): cualquier otra métrica cuantitativa sobre qué tan intensivamente usaron los docentes el LLM.

Códigos/Etiquetas para el registro de datos:

- Para C.2.1 y C.2.2 (categóricas):
  - Registrar en la casilla de “Valor” la categoría exacta reportada en el estudio (ej. "Frecuencia: frecuente"; "Periodicidad: semanal").
  - Si el estudio no usa categorías claras o se requiere comparar entre estudios, anotar en la casilla de “Valor_complementario/Comentarios” la categoría estandarizada sugerida (ej. "Frecuencia: alta/media/baja"; "Periodicidad: diaria/semanal/mensual/ocasional") que mejor corresponda, indicando "(Original: '[categoría original del estudio]')".
- Para C.2.3 y C.2.4 (numéricas):
  - Registrar en la casilla de “Valor” el valor numérico reportado, incluyendo siempre la unidad (ej. "No. interacciones: 5.2 / día"; "Tiempo uso: 30 min / semana").
- Para C.2.5 (otra medida):
  - Describir brevemente la medida en “Valor_complementario/Comentarios” y registrar el valor numérico con su unidad en “Valor”.
- Si no se reporta información cuantitativa o categórica para una subvariable específica, registrar "No reportado" en “Valor”.

- **C.3. Frecuencia de uso de LLMs por parte de los estudiantes**

  Definición: medición de la intensidad, frecuencia o cantidad de uso de herramientas LLM por parte de los estudiantes participantes en el estudio o en la intervención descrita.

  Subvariables (indicadores):

  - **C.3.1. Frecuencia de uso (categórica)**

Definición (e indicador): nivel general de utilización del LLM por estudiantes reportado en categorías (ej. alta, media, baja).

- **C.3.2. Periodicidad de uso (categórica)**

  Definición (e indicador): intervalo temporal con que los estudiantes usan el LLM (ej. diaria, semanal, mensual).

- **C.3.3. No. Interacciones (numérica)**

  Definición (e indicador): cantidad numérica de veces que los estudiantes interactuaron con el LLM en un periodo determinado.

- **C.3.4. Tiempo de uso promedio (numérica)**

  Definición (e indicador): duración media (en horas, minutos) de las sesiones de uso del LLM por parte de los estudiantes.

- **C.3.5. Otra medida de intensidad de uso (especificar)**

  Definición (e indicador): cualquier otra métrica cuantitativa sobre qué tan intensivamente usaron los estudiantes el LLM.

Códigos/Etiquetas para el registro de datos:

- Para C.3.1 y C.3.2 (categóricas):
  - Registrar en la casilla de “Valor” la categoría exacta reportada en el estudio (ej. "Frecuencia: frecuente"; "Periodicidad: semanal").
  - Si el estudio no usa categorías claras o se requiere comparar entre estudios, anotar en la casilla de “Valor_complementario/Comentarios” la categoría estandarizada sugerida (ej. "Frecuencia: alta/media/baja"; "Periodicidad: diaria/semanal/mensual/ocasional") que mejor corresponda, indicando "(Original: '[categoría original del estudio]')".
- Para C.3.3 y C.3.4 (numéricas):
  - Registrar en la casilla de “Valor” el valor numérico reportado, incluyendo siempre la unidad (ej. "No. interacciones: 10.5 / semana"; "Tiempo uso: 45 min / día").
- Para C.3.5 (otra medida):
  - Describir brevemente la medida en “Valor_complementario/Comentarios” y registrar el valor numérico con su unidad en “Valor”.

- Si no se reporta información cuantitativa o categórica para una subvariable específica, registrar "No reportado" en “Valor”.

- **C.4. Impactos objetivos en el razonamiento clínico**

  Definición: mediciones cuantitativas específicas del desarrollo o desempeño de saberes, habilidades o competencias de razonamiento clínico en estudiantes, comparando antes y después del uso de LLMs o entre grupos.

  Subvariables (indicadores):

  - **C.4.1. Precisión diagnóstica (ej. % correcto, puntaje)**

    Definición (e indicador): medida cuantitativa de la exactitud al establecer diagnósticos (porcentaje de aciertos, puntaje de calidad).

  - **C.4.2. Tiempo hasta diagnóstico/resolución**

    Definición (e indicador): medida del tiempo (minutos, segundos) requerido para llegar a una conclusión diagnóstica o resolver un caso.

  - **C.4.3. Proporción/Tipo de errores en razonamiento clínico**

    Definición (e indicador): cuantificación o categorización de fallos específicos cometidos durante el proceso de razonamiento clínico.

  - **C.4.4. Puntajes en pruebas específicas de razonamiento clínico (ej. ECOE...)**

    Definición (e indicador): calificaciones obtenidas en evaluaciones diseñadas explícitamente para medir habilidades de razonamiento clínico (como ECOE).

  - **C.4.5. Otra medida objetiva de razonamiento clínico (especificar)**

    Definición (e indicador): cualquier otra métrica cuantitativa del desempeño en razonamiento clínico no cubierta previamente.

  Códigos/Etiquetas para el registro de datos:

  - Para cada subvariable aplicable (C.4.1 a C.4.5):
    - Registrar en la casilla de “Valor” los datos numéricos específicos reportados (ej. %; puntaje; tiempo en minutos) para cada grupo (ej. Intervención vs Control) y/o momento (ej. Pre vs Post).
    - Añadir en “Valor” la información estadística comparativa si se reporta (ej. valor de p; tamaño del efecto).

- Consignar en la casilla de “Valor_complementario/Comentarios” la prueba o método específico utilizado para la medición (ej. "ECOE estación 3"; "Script Concordance Test"; "examen de casos clínicos simulados") y la escala si aplica (ej. "escala 0-100").
- Si se reportan múltiples medidas distintas para una misma subvariable (ej. resultados de precisión y tiempo separados), registrarlas todas en “Valor”, separando cada conjunto de datos con punto y coma (";"). (Ejemplo C.4.4: "ECOE Global: Grupo LLM 75 ± 8 vs Control 70 ± 10, p=0.06; ECOE Estación Diagnóstico: Grupo LLM 82 ± 5 vs Control 75 ± 7, p=0.04).
- Si no se encuentran datos cuantitativos objetivos para una subvariable, registrar "No reportado" en “Valor”.

5.1.4 Bloque D: Variables secundarias - Cualitativas

- **D.1. Facilidad de implementación de los LLMs por estudiantes o docentes**

  Definición: grado percibido por los estudiantes o docentes respecto a la simplicidad o complejidad de integrar y operar los LLMs en la experiencia educativa específica, considerando aspectos técnicos, recursos y procesos organizacionales.

  Subvariables (indicadores):

  - **D.1.1. Percepción sobre complejidad técnica (hardware, software, interoperabilidad, etc.)**

    Definición (e indicador): juicio sobre la dificultad relacionada con hardware, software o integración tecnológica al implementar el LLM.

  - **D.1.2. Percepción del consumo de recursos (financieros, temporales, humanos, técnicos, etc.) para la implementación**

    Definición (e indicador): juicio sobre la magnitud de la inversión (tiempo, dinero, personal) necesaria para implementar el LLM.

  - **D.1.3. Percepción sobre la complejidad en la adaptabilidad a flujos de trabajo**

    Definición (e indicador): juicio sobre la dificultad para integrar el uso del LLM en los procesos y rutinas académicas existentes.

  - **D.1.4. Identificación de barreras y facilitadores organizacionales o administrativos**

    Definición (e indicador): mención específica de factores institucionales o de gestión que obstaculizan o ayudan a la implementación del LLM.

Códigos/Etiquetas para el registro de datos:

- Para D.1.1, D.1.2, D.1.3 (Percepciones de complejidad/consumo):
  - Identificar el grado de dificultad/costo reportado y el motivo o aspecto asociado.
  - Intentar estandarizar el grado usando las escalas provistas: (muy difícil, difícil, moderado, fácil, muy fácil) para D.1.1/D.1.3; (muy altos, altos, medios, bajos, muy bajos) para D.1.2.
  - Registrar en la casilla “Valor” el término estandarizado (ej. "Moderado"; "Altos"). Si la estandarización no es posible según el estudio original, registrar el término literal principal usado en el estudio (ej. "Desafío considerable").
  - Registrar en la casilla “Valor_complementario/Comentarios” el motivo o aspecto específico asociado (ej. "por interoperabilidad limitada"; "debido a necesidad de licencias"). Si se aplicó estandarización, añadir: "(Original: “[término original del estudio]”)". Si no fue posible estandarizar, añadir: "(No estandarizable, anotación: [describir brevemente por qué o cómo se reporta en el estudio])".
- Para D.1.4 (Barreras y facilitadores):
  - Extraer la lista concisa de barreras y facilitadores mencionados. Procurar estandarizar las categorías usadas para facilitar la comparación entre estudios.
  - Registrar en la casilla “Valor” los elementos listados, separados por punto y coma (";") si son múltiples (ej. "Barrera: falta de política institucional; Facilitador: apoyo del departamento TI").
  - Usar la casilla “Valor_complementario/Comentarios” si se requiere alguna aclaración contextual sobre los ítems, o si la forma en que es reportada no es realmente compatible con el estándar definido, usar el sistema propio del estudio y hacer la anotación al respecto. (Ej.: "Categorías originales: [listar]; Estandarización intentada: [describir]").

- **D.2. Aceptación cultural, organizacional o ética hacia los LLMs**

  Definición: actitudes, creencias y juicios de valor de la comunidad (estudiantes, docentes, personal) sobre la pertinencia, adecuación y consecuencias del uso de LLMs en su contexto, incluyendo consideraciones sobre normas sociales y principios éticos.

Subvariables (indicadores):

- **D.2.1. Percepciones de usuarios sobre utilidad o confiabilidad**

  Definición (e indicador): juicio de estudiantes o docentes sobre qué tan provechoso o fiable consideran el LLM en su contexto.

- **D.2.2. Posturas u opiniones del docentes, directivos o autoridades supervisoras sobre el uso de LLMs**

  Definición (e indicador): puntos de vista expresados por personal con roles de supervisión o liderazgo sobre el uso de LLMs en educación médica.

- **D.2.3. Preocupaciones éticas identificadas**

  Definición (e indicador): mención específica de dilemas o problemas morales (plagio, sesgo, privacidad, etc.) asociados al uso de LLMs.

- **D.2.4. Percepción sobre impacto en la interacción humana**

  Definición (e indicador): opiniones sobre cómo el uso de LLMs afecta las relaciones entre personas (docente-alumno, alumno-paciente, etc.) en el ámbito educativo.

Códigos/Etiquetas para el registro de datos:

- Para D.2.1 (Utilidad/Confiabilidad) y D.2.4 (Impacto en interacción):
  - Identificar el grado de utilidad/confiabilidad/impacto reportado y el aspecto asociado.
  - Intentar estandarizar usando escalas sugeridas: (muy alta, alta, media, baja, muy baja) para D.2.1; (muy alto, alto, medio, bajo, muy bajo) para D.2.4.
  - Registrar en la casilla “Valor” el término estandarizado (ej. "Alta"; "Bajo"). Si no es posible según el estudio original, registrar el término literal principal (ej. "Generalmente útil").
  - Registrar en la casilla “Valor_complementario/Comentarios” el aspecto específico asociado (ej. "para síntesis de información"; "en relación estudiante-paciente"). Si se aplicó estandarización, añadir: "(Original: “[término original]”)". Si no fue posible estandarizar, añadir: "(No estandarizable, anotación: [describir brevemente])".
- Para D.2.2 (Posturas/Opiniones) y D.2.3 (Preocupaciones éticas):

- Extraer la lista concisa de posturas, opiniones o preocupaciones mencionadas. Procurar estandarizar con categorías temáticas (para D.2.2) o categorías de preocupaciones éticas (para D.2.3).
- Registrar en la casilla “Valor” los elementos listados o las categorías estandarizadas, separados por punto y coma (";") si son múltiples (ej. D.2.2: "facilita retroalimentación; riesgo de dependencia; transforma rol docente"; D.2.3: "plagio; sesgo algorítmico; privacidad").
- Usar la casilla “Valor_complementario/Comentarios” para notas contextuales, o si la forma en que es reportada no es realmente compatible con el estándar definido, usar el sistema definido en el estudio y hacer la anotación al respecto (ej. "Original: '[término/descripción original]'; Estandarización intentada: [categoría]").

- **D.3. Entrenamiento o soporte en el uso de los LLMs para estudiantes o docentes**

  Definición: valoración de la adecuación, suficiencia y efectividad de las acciones de entrenamiento y soporte proporcionadas a los estudiantes o docentes para el manejo competente del LLM, así como la identificación de necesidades adicionales.

  Subvariables (indicadores):

  - **D.3.1. Descripción del tipo y modalidad del entrenamiento o soporte recibido por los docentes**

    Definición (e indicador): detalles sobre la capacitación o ayuda formal/informal proporcionada a los profesores para usar LLMs.

  - **D.3.2. Descripción del tipo y modalidad del entrenamiento o soporte recibido por los estudiantes**

    Definición (e indicador): detalles sobre la capacitación o ayuda formal/informal proporcionada a los alumnos para usar LLMs.

  - **D.3.3. Percepción sobre la suficiencia y efectividad del entrenamiento o soporte recibido por los docentes**

    Definición (e indicador): valoración de los profesores sobre si la capacitación/ayuda recibida fue adecuada y útil.

  - **D.3.4. Percepción sobre la suficiencia y efectividad del entrenamiento o soporte recibido por los estudiantes**

Definición (e indicador): valoración de los alumnos sobre si la capacitación/ayuda recibida fue adecuada y útil.

- **D.3.5. Identificación de necesidades no cubiertas por los docentes**

  Definición (e indicador): áreas o temas específicos donde los profesores sienten que requieren más entrenamiento o soporte para usar LLMs.

- **D.3.6. Identificación de necesidades no cubiertas por los estudiantes**

  Definición (e indicador): áreas o temas específicos donde los alumnos sienten que requieren más entrenamiento o soporte para usar LLMs.

- **D.3.7. Disponibilidad percibida de recursos de ayuda por los docentes**

  Definición (e indicador): valoración de los profesores sobre la facilidad para acceder a guías, soporte técnico u otros recursos de ayuda sobre LLMs.

- **D.3.8. Disponibilidad percibida de recursos de ayuda por los estudiantes**

  Definición (e indicador): valoración de los alumnos sobre la facilidad para acceder a guías, soporte técnico u otros recursos de ayuda sobre LLMs.

Códigos/Etiquetas para el registro de datos:

- Para D.3.1 y D.3.2 (Descripción entrenamiento/soporte):
  - Extraer descripción concisa de tipo, modalidad, contenido, fuente. Intentar usar categorías estándar (formal/informal, grupal/individual, contenido general/específico, duración, fuente interna/externa) si aplica, según lo definido en el estudio o procurando estandarizar.
  - Registrar en la casilla “Valor” la descripción principal (ej. "Taller presencial grupal sobre prompts"; "Tutorial en línea sobre herramienta X"; "Soporte técnico institucional"). Si son múltiples, separar con ";".
  - Usar la casilla “Valor_complementario/Comentarios” para detalles adicionales (ej. "Duración: 2 horas"; "Fuente: proveedor externo"; "Enfoque: generalidades") o para anotar si la estandarización no fue compatible, usando el sistema del estudio.
- Para D.3.3 y D.3.4 (Percepción suficiencia/efectividad):
  - Identificar la valoración reportada (ej. "suficiente", "insuficiente") y el aspecto asociado.

        - Intentar estandarizar usando: "Suficiente"; "Insuficiente"; "Deficiente".
        - Registrar en la casilla “Valor” el término estandarizado. Si no es posible, usar término literal.
        - Registrar en la casilla “Valor_complementario/Comentarios” el aspecto específico (ej. "para evaluación de sesgos"; "para mejorar rigor"). Si se estandarizó, añadir: "(Original: “[término original]”)". Si no, añadir: "(No estandarizable, anotación: [describir])".
    - Para D.3.5 y D.3.6 (Necesidades no cubiertas):
        - Extraer lista concisa de áreas de carencia o necesidad de apoyo adicional. Procurar estandarizar las categorías con que se describen los tipos de dificultades o necesidades, definiendo categorías temáticas (ej. “manejo de sesgos”; “profundidad de la respuesta”; “veracidad de la información”, etc.).
        - Registrar en la casilla “Valor” los temas listados o categorías estandarizadas, separados por ";" si son múltiples (ej. "manejo de sesgos; profundidad de respuesta; veracidad").
        - Usar la casilla “Valor_complementario/Comentarios” para contexto si es necesario, o si la forma en que es reportada no es realmente compatible con el estándar, usar el sistema del estudio y hacer la anotación.
    - Para D.3.7 y D.3.8 (Disponibilidad recursos ayuda):
        - Identificar la valoración sobre facilidad de acceso (ej. "fácil", "difícil") y el recurso/motivo asociado.
        - Intentar estandarizar usando: (muy difícil, difícil, moderado, fácil, muy fácil).
        - Registrar en la casilla “Valor” el término estandarizado. Si no es posible, usar término literal.
        - Registrar en la casilla “Valor_complementario/Comentarios” el recurso o motivo específico (ej. "por tutoriales en línea"; "porque no hay soporte específico"). Si se estandarizó, añadir: "(Original: “[término original]”)". Si no, añadir: "(No estandarizable, anotación: [describir])". También se pueden estandarizar los aspectos con que se relaciona esa facilidad/dificultad (ej. guías, soporte técnico).

- **D.4. Descripción de las características del grupo de comparación (control)**

  Definición: caracterización detallada de las características que definen al grupo de comparación o control (nivel académico, entorno de clase, herramientas

tecnológicas disponibles, actividades específicas, etc.), que no recibe la intervención con LLM, para establecer una línea base comparable.

Subvariables (indicadores):

- **D.4.1. Descripción de herramientas tecnológicas y recursos estándar disponibles o utilizados**

  Definición (e indicador): listado de tecnología (software, BD, plataformas) y materiales (libros, guías) usados habitualmente por el grupo control.

- **D.4.2. Descripción de métodos o procedimientos estándar utilizados**

  Definición (e indicador): detalle de las técnicas o metodologías pedagógicas (ABP, clase magistral, etc.) empleadas por el grupo control.

- **D.4.3. Descripción del currículo, actividades formativas y nivel académico**

  Definición (e indicador): información sobre los temas, estructura de clases/prácticas y evaluaciones realizadas por el grupo control.

- **D.4.4. Exposición o uso auto reportado de LLMs fuera de la intervención específica del estudio**

  Definición (e indicador): reporte sobre si los participantes control usan LLMs por su cuenta, con qué frecuencia y para qué fines.

- **D.4.5. Características relevantes del entorno de clase o aprendizaje**

  Definición (e indicador): detalles sobre el ambiente físico, tecnológico general o social/organizacional que caracteriza al grupo control.

Códigos/Etiquetas para el registro de datos:

- Para D.4.1, D.4.2, D.4.3, D.4.5 (Descripciones de recursos, métodos, currículo, entorno):
  - Extraer la descripción concisa de los elementos relevantes para cada categoría.
  - Registrar en la casilla “Valor” la lista o descripción principal, usando ";" para separar ítems distintos dentro de la misma subvariable (ej. D.4.1: "Acceso PubMed; Licencia software X; Libro guía Y"; D.4.2: "ABP con tutor; Discusión plenaria casos").
  - Intentar usar categorías estándar sugeridas si aplica (ej. D.4.1: "Bases de datos bibliográficas; Software especializado"; D.4.2: "ABP; Clase magistral"). Anotar la categoría estándar en “Valor_complementario/Comentarios” si ayuda a la comparación y

no reemplaza la información específica en “Valor”. Si la forma en que es reportada no es realmente compatible con el estándar, usar el sistema definido en el estudio y hacer la anotación al respecto.

- Para D.4.3, incluir nivel académico y fuente si se reporta, preferiblemente en “Valor_complementario/Comentarios”.
- Usar la casilla “Valor_complementario/Comentarios” para cualquier nota contextual importante.

- Para D.4.4 (Uso externo de LLMs en grupo control):
  - Registrar en la casilla “Valor” la respuesta categórica principal: "Sí" o "No".
  - Si es "Sí", añadir en “Valor” la frecuencia estandarizada si es posible (usando categorías como "Nunca"; "Rara vez (<1/mes)"; "Ocasionalmente (1-3/mes)"; "Frecuentemente (Semanal)"; "Muy frecuentemente (Diario)"). Si no es posible estandarizar frecuencia, registrar solo "Sí".
  - Registrar en la casilla “Valor_complementario/Comentarios” los detalles: propósitos principales (intentar usar categorías estándar: "Búsqueda información general"; "Redacción/corrección textos"; "Traducción"; etc.), frecuencia original si no se pudo estandarizar (ej. "Original: “un par de veces por semana”"), y cualquier otra nota relevante sobre este uso externo (clave para evaluar contaminación). Si la forma en que es reportada no es compatible con el estándar, usar el sistema del estudio y anotar.
  - Si la respuesta es "No", el campo Valor_complementario/Comentarios puede quedar vacío o indicar "No aplica".

5.1.5 Bloque E: Variables secundarias - Cuantitativas

- **E.1. Características sociodemográficas y académicas de los participantes**

  Definición: atributos individuales de los participantes (tanto del grupo intervención como del grupo control) que permiten describir la muestra total y los subgrupos, controlar posibles variables de confusión y analizar diferencias en los resultados según estas características.

  Subvariables (indicadores) y Códigos/Etiquetas para el registro de datos:

  - **E.1.1. Edad**

    Definición (e indicador): años cumplidos por el participante al momento del estudio.

Códigos/Etiquetas para el registro de datos:

- Registrar en la casilla “Valor” el número entero reportado (ej. "24"; "35") o la medida de tendencia central y dispersión (ej. "Media 23.5 ± 2.1 años").
- Si no se especifica, registrar "No reportado".

- **E.1.2. Sexo / Género**

  Definición (e indicador): identificación de sexo o género reportada por el participante.

  Códigos/Etiquetas para el registro de datos:

  - Registrar en la casilla “Valor” la categoría estandarizada que mejor corresponda: "Hombres"; "Mujeres"; "Otro / no definido". Usar ";" para separar si se reportan datos para más de una categoría (ej. "% Hombres; % Mujeres").
  - Si la categoría del estudio no coincide exactamente o usa terminología diferente (procurar estandarizar y adaptarse), registrar la categoría original en “Valor_complementario/Comentarios” (ej. "Original: Varones") y anotar el comentario si no fue posible la adaptación.
  - Si no se especifica, registrar "No reportado".

- **E.1.3. Nivel académico del participante**

  Definición (e indicador): categoría que describe la etapa formativa o rol profesional del participante dentro del ámbito médico/académico.

  Códigos/Etiquetas para el registro de datos:

  - Categorías estándar (Procurar estandarizar y adaptarse a estas categorías. Si no es posible adaptarse, registrarlo como está definido en el estudio y anotar el comentario):
  - 0 = Médico especialista o con posgrado, docente, sin formación pedagógica.
  - 1 = Médico especialista o con posgrado, docente, con formación pedagógica.
  - 2 = Médico general o sin posgrado, docente, sin formación pedagógica.
  - 3 = Médico general o sin posgrado, docente, con formación pedagógica.

- 4 = Otro profesional (no médico) docente, sin formación pedagógica.
- 5 = Otro profesional (no médico) docente, con formación pedagógica.
- 6 = Médico estudiante (posgrado), de especialidad médico-quirúrgica.
- 7 = Médico estudiante (posgrado), de maestría o doctorado.
- 8 = Médico interno (estudiante de medicina de último año).
- 9 = Estudiante de medicina (pregrado), anterior al último año o sin año especificado.
- 10 = Otro médico asistencial, sin otra especificación.
- 11 = Otro médico estudiante de posgrado, sin otra especificación.
- 12 = Otro profesional no médico, sin otra especificación. Registrar en la casilla “Valor” el código numérico (0-12) de la categoría estandarizada que mejor describa a los participantes. Si aplica a múltiples categorías, separarlas con ";".
- Registrar en "Valor" el código numérico (0-12) de la que mejor describa a los participantes. Si aplica a múltiples categorías, separarlas con ";".
- Registrar en "Valor_complementario/Comentarios" la descripción original usada en el estudio si difiere o aporta contexto relevante.
- Si no está especificado, registrar "No reportado".

- **E.1.4. Experiencia previa con LLMs**

  Definición (e indicador): nivel auto reportado de familiaridad o uso de LLMs por el participante antes de iniciar el estudio.

  Códigos/Etiquetas para el registro de datos:

  - Escala estándar (Procurar estandarizar y adaptarse a estas categorías. Si no es posible adaptarse, registrarlo como está definido en el estudio y anotar el comentario):
    - 0 = Ninguna / Ni conoce, ni usa.
    - 1 = Muy baja o mínima o la ha usado en contadas ocasiones.
    - 2 = Baja o discreta o la usa esporádicamente o para tareas elementales.

- 3 = Moderada o media o la usa ocasionalmente o para tareas simples.
- 4 = Alta o la usa a menudo o para tareas complejas.
- 5 = Muy alta o experto o la usa constantemente o para tareas avanzadas.

- Registrar en la casilla “Valor” el código numérico (0-5) de la escala estandarizada que mejor refleje la experiencia reportada.
- Si el estudio usa una escala diferente, registrar la descripción/categoría original en “Valor_complementario/Comentarios” (ej. "Original: “Usuario ocasional”").
- Si no se especifica, registrar "No reportado".

- **E.1.5. Área de conocimiento / Especialidad principal**

  Definición (e indicador): campo disciplinar o especialidad médica/académica principal del participante.

  Códigos/Etiquetas para el registro de datos:

  - Registrar en la casilla “Valor” la categoría nominal estandarizada que aplique (ej. "Ciencias Básicas Biomédicas"; "Medicina Interna"; "Cirugía General"; "Pediatría"; "Ginecología y Obstetricia"; "Salud Pública"; "Educación Médica"). Usar ";" si aplica a múltiples áreas.
  - Si el área reportada no está en la lista estándar o requiere aclaración (procurar estandarizar y adaptarse), registrarla en “Valor_complementario/Comentarios” y anotar el comentario si no fue posible la adaptación.
  - Si no aplica o no se especifica, registrar "No reportado".

- **E.1.6. Rendimiento académico previo**

  Definición (e indicador): medida cuantitativa (promedio, rango) del desempeño académico del participante antes del estudio.

  Códigos/Etiquetas para el registro de datos:

  - Si se reporta valor numérico (promedio): Registrarlo en “Valor” (ej. "4.2"; "85"). Registrar la escala en “Valor_complementario/Comentarios” (ej. "Escala 1-5"; "Escala 0-100").
  - Si se reporta categoría ordinal (rangos): Registrar la categoría en “Valor” (ej. "Tercil superior"; "Promedio > [Y]"). Registrar la definición de los rangos/cortes en “Valor_complementario/Comentarios”.

- Si no se especifica, registrar "No reportado".

- **E.2. Características técnicas del (los) LLM(s) empleado(s) en la intervención**

  Definición: especificaciones técnicas todos los LLMs particulares utilizados como parte de la intervención en el grupo experimental, relevantes para entender su funcionamiento, el alcance o las limitaciones de los resultados y replicar el estudio.

  Subvariables (indicadores) y Códigos/Etiquetas para el registro de datos:

  - **E.2.1. Nombre del modelo LLM**

    Definición (e indicador): identificador comercial o técnico del LLM principal utilizado en la intervención.

    Códigos/Etiquetas para el registro de datos:

    - Registrar en la casilla “Valor” el nombre comercial/técnico reportado (ej. "GPT-4"; "Claude 3 Opus"; "Gemini 1.5 Pro"; "Llama 3 70B Instruct").

  - **E.2.2. Versión del modelo**

    Definición (e indicador): especificador de la versión o fecha de corte del conocimiento del LLM utilizado.

    Códigos/Etiquetas para el registro de datos:

    - Registrar en la casilla “Valor” la versión específica reportada (ej. "gpt-4-1106-preview"; "claude-3-opus-20240229"; "Modelo con fecha de corte de conocimiento: septiembre 2023").
    - Si no es claro o no se reporta, registrar "No reportado".

  - **E.2.3. Plataforma o Modo de acceso**

    Definición (e indicador): descripción de la interfaz o medio a través del cual los participantes interactuaron con el LLM.

    Códigos/Etiquetas para el registro de datos:

    - Registrar en la casilla “Valor” la categoría que mejor describa el modo de acceso, usando los ejemplos como guía (ej. "Interfaz web oficial (chat.openai.com)"; "API integrada en software [Nombre del Software]"; "Aplicación móvil oficial"; "Plataforma institucional [Nombre de la Plataforma]"; "Acceso vía [Proveedor Tercero]").
    - Si el modo de acceso es diferente o requiere detalle (procurar estandarizar y adaptarse), registrar la descripción original en “Valor_complementario/Comentarios” y anotar el comentario si no fue posible la adaptación.

- Si no se especifica, registrar "No reportado".

- **E.2.4. Existencia de ajuste fino (“fine-tuning”) o entrenamiento específico del (los) LLM(s)**

  Definición (e indicador): indicación sobre si el LLM fue modificado o entrenado con datos específicos para el estudio.

  Códigos/Etiquetas para el registro de datos:

  - Registrar en la casilla “Valor” una de las categorías: "Sí"; "No"; "No reportado / desconocido".
  - Si Valor es "Sí", registrar la descripción del tipo de ajuste realizado en “Valor_complementario/Comentarios” si se conoce (ej. "Ajustado con datos de casos clínicos anonimizados de la institución"; "Entrenado para responder preguntas sobre guías de práctica clínica específicas").

- **E.2.5. Configuración de parámetros clave (si aplica y se controló)**

  Definición (e indicador): detalle de los ajustes específicos (temperatura, top_p, etc.) aplicados al LLM durante la intervención.

  Códigos/Etiquetas para el registro de datos:

  - Registrar en la casilla “Valor” los parámetros y sus valores si fueron fijados (ej. "temperature=0.7; top_p=1.0; max_tokens=1000"). Usar ";" para separar múltiples parámetros. Listar solo los parámetros relevantes modificados.
  - Si se usaron valores por defecto o no se controlaron, registrar "Default" o "No controlado".
  - Si no se reporta información, registrar "No reportado".

- **E.2.6. Costo asociado al uso (si aplica)**

  Definición (e indicador): descripción del modelo de pago o gratuidad para el uso del LLM en el estudio.

  Códigos/Etiquetas para el registro de datos:

  - Registrar en la casilla “Valor” la categoría de costo que aplique (ej. "Gratuito (acceso libre)"; "Gratuito (cubierto por la institución)"; "Basado en suscripción personal"; "Costo por uso (API)"; "No aplica").
  - Si se reporta costo promedio específico o costo variable medible, añadirlo en “Valor_complementario/Comentarios”.
  - Si no se reporta, registrar "No reportado".

5.2 Operacionalización de las variables

Con el propósito de homogeneizar la captura y codificación de la información de todos los estudios primarios incluidos, se definieron pautas comunes para llevar cada variable de su formulación teórica a su expresión empírica.

- Definición de niveles de medición
  - Cualitativas (ej., percepciones, formas de uso): se trataron como variables nominales u ordinales. Cada etiqueta (por ejemplo “positiva”, “negativa”, “neutra”) se codificó en un campo específico, utilizando textos estandarizados y separando múltiples elementos con punto y coma.
  - Cuantitativas (ej. calificaciones, frecuencia de uso, tiempo de interacción): se manejaron como variables continuas o discretas, siempre expresadas en una unidad única (%, puntaje 0–100, número de veces por periodo).
- Fuente de datos e instrumento de registro
  - Todos los valores se extrajeron directamente de tablas, figuras o anexos de los artículos.
  - Se diseñó una plantilla de extracción en hoja de cálculo con columnas fijas: nombre de la variable, subvariable, valor reportado, unidad de medida y grupo (intervención/control).
- Codificación y estandarización
  - Para las variables cualitativas, se aplicó codificación temática previa (etiquetas acordadas) y se permitió la inclusión de códigos emergentes, siempre documentados en un glosario interno.
  - En las variables cuantitativas, se convirtieron todos los valores a la misma escala (ej. transformar mediana y rango intercuartílico a media ± DE cuando fue posible) y se anotó el método de conversión.

6 Técnicas de análisis

6.1 Análisis cualitativo

6.1.1 Codificación temática:

Se aplicará un proceso de codificación basado en las categorías definidas en las secciones 5.1 a 5.5, con el fin de transformar las citas textuales y descripciones narrativas en códigos estandarizados.

A. Identificación de unidades de análisis: frases, párrafos o pasajes relevantes a cada variable o subvariable cualitativa.
B. Asignación de códigos predefinidos (ej., “opinión positiva”, “preocupación ética”, “mejoría conceptual”) y registro de nuevos códigos emergentes cuando aparezcan temas no previstos inicialmente.

6.1.2 Análisis temático:

A partir de la matriz de códigos, se organizarán temas y subtemas para descubrir patrones, convergencias y divergencias entre los estudios incluidos.

A. Agrupación de códigos en categorías superiores (ej., “ventajas percibidas”, “barreras organizacionales”).
B. Exploración de relaciones internas entre categorías (ej., vínculo entre “facilidad de implementación” y “aceptación cultural”).

6.1.3 Resultados esperados:

A. Mapa de temas con frecuencia y representación en la muestra de estudios.
B. Descripción narrativa de los hallazgos cualitativos más relevantes, ilustrada con citas textuales.

6.2 Análisis cuantitativo

6.2.1 Estadística descriptiva:

Se presentarán los datos cuantitativos mediante:

A. Tablas de frecuencias y porcentajes para variables categóricas (ej. frecuencia de uso alta/media/baja).
B. Medidas de tendencia central (media, mediana) y dispersión (desviación estándar, rango intercuartílico) para variables continuas (ej. puntajes, número de interacciones).

6.2.2 Estadística inferencial (si aplica):

Para comparaciones puntuales o análisis de relaciones fuera del metaanálisis, según la necesidad:

A. t-test para comparar medias entre dos grupos (intervención vs. control).
B. ANOVA para comparar tres o más grupos o condiciones.
C. Modelos de regresión lineal o logística para explorar asociaciones predictivas entre variables (p. ej., rendimiento académico y número de interacciones con LLM).

6.2.3 Herramientas de análisis estadístico:

Se emplearán programas de libre acceso y con amplia documentación, según la necesidad:

A. R + RStudio (paquetes metafor, meta)
B. JASP (interfaz gráfica, módulo de metaanálisis)
C. OpenMeta[Analyst] (aplicación independiente en Java)
D. RevMan (Cochrane Review Manager)

6.3 Validación del análisis

6.3.1 Triangulación:

Contraste sistemático de hallazgos cualitativos y cuantitativos para reforzar la coherencia interna de las conclusiones; por ejemplo, comparar percepciones documentadas con resultados de desempeño académico.

6.3.2 Control de errores estadísticos:

A. Ajuste del nivel de significancia ($p < 0,05$) para minimizar errores tipo I.
B. Cálculo de tamaños de efecto (d de Cohen, OR) para valorar la relevancia práctica y reducir riesgos de errores tipo II.

7 Procedimientos logísticos

7.1 Cronograma

Se establece un cronograma con fases claramente definidas:

- Diseño de estrategias de búsqueda: semana 1.
- Recolección de información: semana 2-3.
- Selección y evaluación de calidad: semana 4-8.
- Análisis de datos: semana 9.
- Redacción de resultados y discusión: semana 10-12.

7.2 Selección y contacto de participantes

- Por la naturaleza de este trabajo de investigación, los participantes no serán individuos, sino estudios primarios de investigación.
- Selección: acorde a los criterios descritos previamente.
- Método de contacto: no aplica en esta revisión sistemática.

7.3 Recolección de datos

7.3.1 Método de aplicación de instrumentos:

La búsqueda en bases de datos bibliográficas se hará con las estrategias específicas antes descritas. Tras ella, se procede a hacer la selección de artículos secuencialmente, en 3 etapas:

A. Selección inicial: relación de títulos y resúmenes con los objetivos de esta revisión sistemática
B. Selección preliminar: valoración de cada artículo con los criterios de inclusión y exclusión y una evaluación preliminar (breve) de la calidad metodológica, por medio de una lista de chequeo.
C. Selección final: a los artículos preseleccionados, se los evaluará según su calidad metodológica, con las herramientas antes descritas, para tomas solo los que tengan un bajo riesgo de errores o problemas metodológicos.

7.4 Organización de datos

7.4.1 Creación de bases de datos:

Ingreso de datos en hojas de cálculo de programas como Excel o Google Sheets, estructurando variables e indicadores relevantes.

7.4.2 Codificación y limpieza de datos:

- Asignación de códigos únicos para identificar fácilmente cada uno de los estudios incluidos y prevenir errores durante el análisis de datos.
- Revisión de los datos para corregir errores de ingreso y asegurar consistencia.

7.5 Validación y control de calidad

7.5.1 Prueba piloto:

Realizar una prueba piloto de la estrategia de búsqueda y selección, con una muestra pequeña de estudios, para identificar y corregir posibles problemas con la estrategia empleada.

7.5.2 Monitoreo:

Supervisar el cumplimiento de los protocolos durante todo el proceso de revisión, asegurando que los criterios de inclusión y exclusión se apliquen de manera consistente.

7.5.3 Revisión de expertos:

Se dispondrá de asesoría y supervisión constante del tutor de tesis de grado, respaldado y asignado por la Universidad de Icesi, para asegurarse que todos los procesos desde el diseño y hasta el desarrollo y publicación de resultados se hagan de forma idónea.

7.6 Gestión postrecolección

7.6.1 Transcripción de datos cualitativos:

Se extraen datos cualitativos, verificando hallazgos comunes y específicos de los artículos, que se caracterizarán según las variables de estudio, y se clasificarán de forma precisa en la base de datos.

7.6.2 Almacenamiento seguro:

Guardar los datos dispositivos seguros, con acceso solo habilitado para el autor de este estudio y realizando copias de seguridad periódicas para evitar pérdidas de información.

8 Consideraciones éticas

8.1 Principios éticos fundamentales

- Respeto por las personas: dado que no se tendrá contacto directo con individuos participantes en los estudios revisados o con sus equipos de trabajo, no es directamente aplicable en esta revisión sistemática, pero se garantizará el respeto a los autores y grupos de trabajo y a los individuos participantes en los estudios revisados y seleccionados, con el manejo, análisis y síntesis de resultados de forma profesional, objetiva y respetuosa a los resultados obtenidos por cada uno.
- Beneficencia: los datos serán analizados de forma cuidadosa, sistemática y objetiva, para que los resultados de esta revisión sistemática sean precisos y sus conclusiones

sean coherentes y objetivas, con el fin de minimizar cualquier riesgo y maximizar los beneficios para la educación médica derivados de este estudio.

- Justicia: el protocolo de investigación construido acorde con criterios de calidad AMSTAR-2 y el reporte de resultados y conclusiones de esta revisión sistemática se hará siguiendo los criterios del protocolo PRISMA, todo lo cual permite que esta investigación se lleve a cabo de forma transparente.
- Transparencia y responsabilidad en el uso de IA: se declara el empleo de herramientas de GenAI (ChatGPT, Gemini; ver Anexos: Prompt 1. Asistente y tutor en el desarrollo de una Revisión Sistemática/Metaanálisis para tesis de grado.) para la generación y edición de contenido, con constante proceso de revisión y reedición del material generado con esta herramienta, asumiendo la autoría responsable de todas las salidas producidas. Este enfoque promueve la honestidad, la responsabilidad y la equidad, alineándose con los principios éticos de la AIAS al combinar tecnologías emergentes con la evaluación y supervisión crítica del investigador.
- Respeto a los derechos de autor: se garantizará citando adecuadamente todas las fuentes utilizadas, respetando la propiedad intelectual de los autores de cada trabajo incluido en cualquier momento del desarrollo de la investigación.

8.2 Consentimiento informado

No aplica directamente en una revisión sistemática, ya que solo se utiliza información ya publicada en revistas indexadas, de acceso público.

8.3 Privacidad y confidencialidad

No aplica directamente en una revisión sistemática, ya que solo se utiliza información ya publicada en revistas indexadas, de acceso público.

8.4 Aprobación por el comité de ética

No aplica directamente una revisión y aprobación formal por un comité de ética, dada la naturaleza de este estudio.

Sin embargo, todo el protocolo y la ejecución de esta revisión sistemática será previamente revisado por el tutor del proyecto de investigación de grado, quien cuenta con respaldo de la Universidad Icesi para gestionar y avalar el desarrollo del trabajo.

RESULTADOS

1 Selección de estudios

El resumen de los resultados de las búsquedas bibliográficas se presenta en la Figura 1. Diagrama PRISMA de la selección de estudios. La búsqueda inicial arrojó 1442 registros identificados en bases de datos electrónicas. Tras eliminar 139 duplicados, 1303 títulos

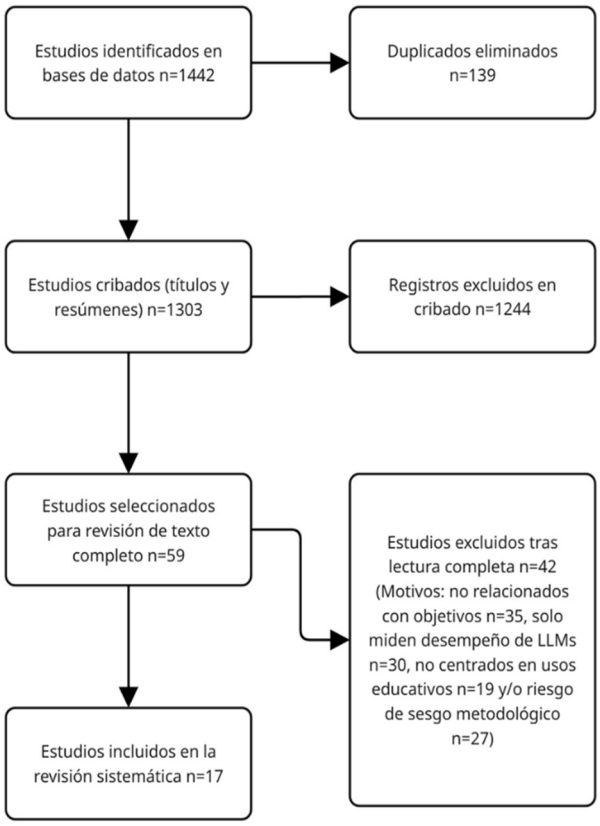
y resúmenes fueron sometidos a cribado; de éstos, 1244 fueron excluidos por no cumplir criterios de elegibilidad predefinidos.

Los 59 estudios restantes fueron revisados en detalle, con la lectura completa de sus textos. De ellos, 42 se descartaron por motivos principales (no relacionados con los objetivos n = 35; solo miden desempeño de LLMs n = 30; no centrados en usos educativos n = 19; riesgo de sesgo metodológico n = 27; el detalle de los motivos de rechazo puede consultarse en la

Tabla suplementaria 2. Motivos detallados de exclusión de estudios tras la lectura del texto completo.

Finalmente, 17 estudios cumplieron los criterios para su inclusión en la revisión sistemática

Figura 1. Diagrama PRISMA de la selección de estudios



2 Características generales de los estudios incluidos

Las principales características de los estudios incluidos en esta revisión se presentan en la Tabla 1. Caracterización de los estudios incluidos. La caracterización detallada de las poblaciones de estudio se incluye en la Tabla suplementaria 3. Caracterización detallada de las poblaciones de los estudios incluidos.

Los 17 estudios proceden de 8 países, con predominio de Estados Unidos (29.4%) y China (23.5%), y abarcan publicaciones entre 2023 y 2025. En conjunto, los diseños metodológicos incluyen ensayos aleatorizados (23.5%), estudios cuasi experimentales no aleatorizados (29.4%) y observacionales transversales (35.3%) y descriptivos (11.8%). El enfoque mixto fue el más frecuente (64.7%), seguido del puramente cuantitativo (23.5%) y cualitativo (11.8%).

El tamaño muestral total entre estudios fue bastante heterogéneo y osciló entre 15 y 2027 participantes (mediana = 64); para los estudios controlados, los grupos de exposición y control mostraron medianas de 30 y 31 sujetos, respectivamente.

Tabla 1. Caracterización de los estudios incluidos

No. Original	Título	Autores	Muestra total	Muestra grupo exposición	Muestra grupo control	Metodología	Enfoque	País
1	Virtual Patient Simulations Using Social Robotics Combined With Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study	Borg et al. (2025a)	15	15	15	Cuasi experimental	Mixto	Suecia
3	AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians	Öncü et al. (2025)	21	21	N/A	Cualitativo	Mixto	Turquía
32	Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial	Brügge et al. (2024)	21	10	11	Cuasi experimental	Cuantitativo	Alemania
36	A Pilot Study of Medical Student Opinions on Large Language Models	A. Y. Xu et al. (2024)	102	N/A	N/A	Cualitativo	Cuantitativo	Estados Unidos
45	Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology	Borg et al. (2024)	23	N/A	N/A	Comparativo cualitativo	Cualitativo	Suecia
59	Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education	Zheng et al. (2024)	66	34	32	Cuasi experimental	Mixto	China
81	Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy	Bharatha et al. (2024)	96	N/A	96	Cualitativo	Cuantitativo	Barbados; Bangladesh
177	Exploring medical student's outlook on use of artificial intelligence in medical education: a multicentric cross-sectional study in Jharkhand, India	Ahmad et al. (2024)	417	N/A	N/A	Cualitativo	Mixto	India
347	Advancing Problem-Based Learning with Clinical Reasoning for Improved Differential Diagnosis in Medical Education	Y. Xu et al. (2025)	18	9	9	Descriptivo	Mixto	China
350	MedSimAI: Simulation and Formative Feedback Generation to Enhance Deliberate Practice in Medical Education	Hicke et al. (2025)	104	104	N/A	Cuasi experimental	Mixto	Estados Unidos
414	Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis	Shakur et al. (2024)	2027	N/A	N/A	Comparativo cualitativo	Mixto	Estados Unidos

500	Assessing the Usability of GutGPT: A Simulation Study of an AI Clinical Decision Support System for Gastrointestinal Bleeding Risk	C. Chan et al. (2023)	55	Fase 1: 24; Fase 2: 28	Fase 1: 31; Fase 2: 23	Descriptivo	Mixto	Estados Unidos
943	Using ChatGPT for medical education: the technical perspective	K. Y. Chan et al. (2025)	45	45	N/A	Cualitativo	Mixto	Hong Kong; Singapur; Australia
959	Large language models in internal medicine residency: current use and attitudes among internal medicine residents	Fried et al. (2024)l	152	N/A	N/A	Cualitativo	Cuantitativo	Estados Unidos
1127	ChatGPT as a Virtual Patient: Written Empathic Expressions During Medical History Taking	Aster et al. (2025)	35	35	N/A	Cuasi experimental	Mixto	Alemania
1138	Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods	Ba et al. (2024)	77	39	38	Descriptivo	Mixto	China
1155	Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT	Huang et al. (2024)	64	32	32	Descriptivo	Cuantitativo	China

3 Calidad metodológica de los estudios

Para garantizar la transparencia y la solidez de los hallazgos, cada grupo de diseño fue evaluado con la herramienta de calidad más adecuada. Los resultados resumidos de riesgo de sesgo y conformidad metodológica se presentan a continuación.

3.1 Ensayos aleatorizados

Los dos ensayos controlados aleatorizados fueron valorados con la herramienta RoB-2 de la Cochrane Collaboration (Ver Anexos, Tabla suplementaria 4. Evaluación de calidad: estudios experimentales aleatorizados - RoB-2. En los dos ensayos controlados aleatorizados, el dominio de desviaciones de la intervención alcanzó un cumplimiento pleno (100%), lo que evidencia alta fidelidad al protocolo original. Sin embargo, los procesos de generación de la secuencia aleatoria, el ocultamiento de la asignación y la comparabilidad de grupos al inicio no se documentaron adecuadamente (0%), lo cual incrementa el riesgo de sesgo de selección. El manejo de datos incompletos y el reporte selectivo de resultados mostraron un cumplimiento moderado (50%), sugiriendo preocupación por sesgos de atrición y reporte.

3.2 Estudios transversales o descriptivos

Los cinco estudios observacionales obtuvieron de corte transversal o descriptivos fueron evaluados con la herramienta JBI-ACSS de la JBI Collaboartion (ver Anexos, Tabla suplementaria 5. Evaluación de calidad: estudios transversales o descriptivos - JBI-ACSS). En la valoración, los estudios obtuvieron un cumplimiento total (100%) en criterios como definición clara de la muestra, descripción detallada de sujetos y contexto, uso de criterios de medición objetivos y análisis estadístico descriptivo apropiado. En contraste, la validez y fiabilidad de la medición de la exposición y la identificación de factores de confusión alcanzaron un cumplimiento del 75%, lo que indica riesgos moderados de sesgo de medición y confusión.

3.3 Estudios cualitativos

Solo un estudio cualitativo fue incluido en esta revisión sistemática y fue evaluado con la herramienta JBI-QS de la JBI Collaboartion (ver Anexos, Tabla suplementaria 6. Evaluación de calidad: estudios cualitativos - JBI-QS). En la valoración, mostró cumplimiento íntegro (100%) en todos los criterios de congruencia metodológica, profundidad analítica y reflexividad investigadora. Esta consistencia sugiere un rigor metodológico destacado en la fase cualitativa.

3.4 Estudios de métodos o enfoque mixto

Los diez estudios con componentes metodológicos y de enfoques analíticos mixtos fueron valorados con la herramienta MMAT en su versión 2018 (ver Anexos, Tabla suplementaria 7. Evaluación de calidad: estudios de métodos o enfoque mixto - MMAT). Estos estudios presentaron altos niveles de cumplimiento en los componentes cualitativo y cuantitativo descriptivo (80 – 100%), reflejando solidez en la justificación del enfoque y en la calidad del reporte estadístico. El control de variables de confusión en los cuantitativos no aleatorizados alcanzó 60%, mientras que la justificación del diseño de métodos mixtos y la integración efectiva de componentes lograron 70%. Sólo el 40% de los estudios abordó de forma adecuada las discrepancias entre hallazgos cualitativos y cuantitativos, lo que señala una oportunidad de mejora en la fase de integración de métodos.

4 Hallazgos de la revisión sistemática

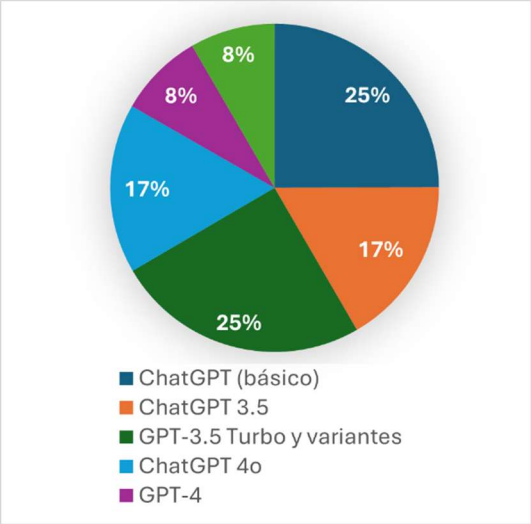
El propósito central de esta revisión es responder a la pregunta de investigación: ¿De qué manera la incorporación de experiencias educativas asistidas por Modelos de Lenguaje a Gran Escala (LLMs), en comparación con métodos pedagógicos tradicionales, impacta en la integración conceptual interdisciplinar entre ciencias básicas y clínicas, así como en el desarrollo y en la transferencia práctica del razonamiento clínico, en estudiantes de medicina?

Los hallazgos que se expondrán a continuación se fundamentan en la información extraída de los 17 estudios incorporados a la revisión, y que se organizó en Bloques temáticos de variables A, B, C, D y E. El detalle de los resultados se presenta en la

Tabla 2. Tabla de resultados. La exposición de estos resultados se estructurará en función de los objetivos específicos planteados para esta investigación, abordando secuencialmente el impacto de los LLMs en las competencias en razonamiento clínico (Objetivo O1), la facilitación de la integración conceptual interdisciplinar (Objetivo O2), las percepciones sobre el uso de estas herramientas (Objetivo O4), y la evidencia disponible respecto a la transferencia percibida del razonamiento clínico a la práctica real (Objetivo O3).

4.1 Impacto de las experiencias educativas asistidas por LLMs en las competencias en razonamiento clínico (objetivo O1)

1.1 Figura 2. LLMs implementados en los estudios



Para comprender mejor el entorno en el que se evaluó el razonamiento clínico, en la **Error! Reference source not found.** se describen los modelos utilizados en los estudios que aportan datos a este objetivo. Solo 3 de 17 estudios (17.6%) detallaron cambios en configuración de parámetros del modelo (temperatura y top_p), mientras que la mayoría aplicó valores por defecto o no lo especificó. Además, en cuanto a las plataformas de acceso a estos modelos, 3 (17.6%) estudios reportan haber implementado una API institucional o de terceros, 2 (11.8%) usaron la interfaz web oficial del modelo y otros 2 (11.8%) usaron plataformas robóticas sociales, 1 (5.9%) utilizo aplicaciones móviles, 4 (23.5%) emplearon entornos HPC o software especializado y en el resto no fue reportado (5 = 29.4%). Resulta llamativo que apenas un tercio de los estudios especifique parámetros de configuración, un aspecto clave cuya variabilidad exploraremos en la discusión.

El desarrollo de un razonamiento clínico eficaz es fundamental en la formación médica, pues combina la capacidad de identificar patrones con un análisis reflexivo y fundamentado. En los 17 estudios incluidos, siete (41.2%) reportaron datos numéricos sobre desempeño académico en actividades centradas en razonamiento clínico, comparando grupos expuestos a LLMs con controles tradicionales.

En términos de actividades específicas, los resultados muestran en varios casos una ventaja para los grupos asistidos por IA. Por ejemplo, en un estudio cuasiexperimental con internos de medicina que participaban en un módulo de educación en enfermedades

cardiovasculares, Zheng et al. (2024) documentaron que quienes utilizaron un modo de simulación potenciado por IA obtuvieron una media de 92.94 ± 2.13 puntos en una escala de pensamiento crítico clínico, frente a 89.31 ± 2.53 en el grupo control ($p < 0.001$). De forma similar, en el contexto de interacciones con pacientes estandarizados mediante IA, Hicke et al. (2025) se enfocaron en la generación de retroalimentación formativa, validando la escala MIRS para evaluar competencias comunicativas, aunque no reportaron un desempeño comparativo porcentual directo en tareas clínicas frente a un grupo sin IA.

Respecto a pruebas de conocimientos, Bharatha et al. (2024) compararon el rendimiento en preguntas de selección múltiple (MCQs, como se definen en el artículo): en preguntas preclínicas, ChatGPT-4 alcanzó un 77.7% de respuestas correctas, mientras que el rendimiento general de los alumnos de tercer año de medicina en todas las MCQs fue del 66.7%. En MCQs clínicas, los mismos autores informaron un desempeño del 70.7% para ChatGPT-4, manteniendo una diferencia con el promedio general de los estudiantes (66.7% en todas las MCQs). Es importante notar que el estudio no presentó una comparación estadística directa entre el rendimiento de ChatGPT-4 y el de los estudiantes para estos subconjuntos específicos de preguntas preclínicas o clínicas con los porcentajes mencionados.

Algunas intervenciones se centraron en instrumentos estandarizados de razonamiento clínico. Brügge et al. (2024) midieron el desempeño con una herramienta estandarizada (CRI-HTI[§]) para razonamiento en la historia clínica: el grupo con retroalimentación estructurada por IA (media 3.60 ± 0.13) mostró un mejor desempeño que el grupo control sin dicha retroalimentación (media 3.02 ± 0.12) ($F(1,18)=4.44$; $p = 0.049$), y en los subdominios “Creando contexto” y “Asegurando información” registraron aumentos significativos ($p = 0.046$ y $p = 0.018$, respectivamente), pero no así en el subdominio de “Enfocando preguntas” ($p = 0.265$). Ba et al. (2024), usando Mini-CEX[**], observaron que un 66.6% de los estudiantes asistidos por ChatGPT excedieron expectativas clínicas versus 36.8% en el grupo tradicional ($p = 0.032$).

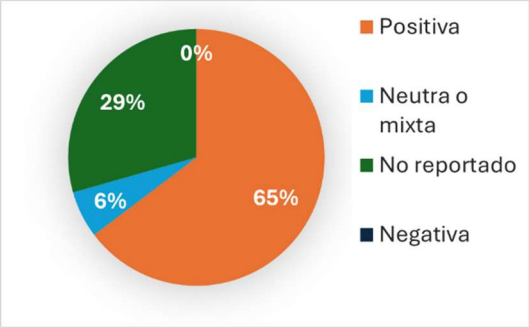
En contraste, algunos estudios no encontraron diferencias significativas a pesar de tendencias favorables. Por ejemplo, el hallazgo de una mejora no significativa en el subdominio “Enfocando preguntas” del CRI-HTI antes mencionado (Brügge et al., 2024) o la comparación de puntuaciones teóricas en rotaciones pediátricas, en el mismo estudio de Ba et al. (2024), mostró medias casi idénticas (92.21 ± 2.37 vs 92.38 ± 2.68;

[§] Fürstenberg et al. (2020) una escala diseñada para evaluar la capacidad de razonamiento clínico de estudiantes avanzados de medicina durante la toma de historia clínica en simulaciones a través de 3 dominios principales (Focusing Questions, Creating Context, Securing information), con varios elementos que se valoran con una rúbrica analítica específica.

[**] Desarrollada por Norcini et al. (2003) para la evaluación estandarizada de la competencia clínica, que se utiliza para observar y evaluar a estudiantes o residentes de medicina mientras interactúan con pacientes. Permite evaluar habilidades y destrezas clínicas de manera estructurada, documentando el desempeño en áreas clave y proporcionando retroalimentación constructiva

p = 0.768). Asimismo, cuando se evaluó la valoración de riesgo en hemorragia gastrointestinal mediante simulación guiada por LLM, K. Y. Chan et al. (2025) reportaron mayor variabilidad en el grupo asistido, lo que sugiere que el impacto puede depender del dominio clínico y del diseño de la actividad.

1.2 Figura 3. Percepciones sobre los LLMs



Más allá de las métricas cuantitativas, los estudios incluidos recogen percepciones de los estudiantes y descripciones de escenarios de interacción con LLMs que revelan dinámicas de razonamiento clínico apoyado por la IA.

Como se ve en la Figura 3, la mayoría de los estudios (11) reportaron valoraciones positivas sobre el uso de LLMs en actividades de simulación y diagnóstico, mientras que solo uno (5.9%) describió opiniones neutras o mixtas. Ningún estudio reportó opiniones negativas, pero es de anotarse que en los restantes estudios (5) no se informó la opinión general de los participantes.

En conjunto, estos datos cuantitativos directos apuntan a un efecto generalmente positivo de los LLMs sobre el rendimiento en tareas de razonamiento clínico, tanto en dimensiones básicas como clínicas y en instrumentos de evaluación estructurados. Posteriores apartados explorarán cómo estas mejoras se complementan con percepciones cualitativas y con el contexto metodológico de cada intervención.

4.2 Facilitación de la integración conceptual interdisciplinar (objetivo O2)

La integración entre ciencias básicas (ej., fisiología, farmacología) y clínicas (diagnóstico, manejo de pacientes) es esencial para un razonamiento clínico verdaderamente avanzado. Los LLMs, al ofrecer explicaciones en múltiples niveles de complejidad y plantear escenarios que enlazan mecanismos fisiopatológicos con decisiones terapéuticas, podrían funcionar como un andamiaje cognitivo que promueva esta articulación.

Diez de los 17 estudios (58.8 %) incluyeron actividades explícitas para fomentar la integración básico-clínica. En el ensayo cuasiexperimental de Öncü et al. (2025), las solicitudes argumentales del LLM obligaron a los internos a fundamentar cada paso diagnóstico en principios fisiopatológicos, lo que se tradujo en una media de 8.33 ± 1.35 en observación de razonamiento clínico (escala 0–10). Brügge et al. (2024) diseñaron casos integradores donde tras simular al paciente, el LLM solicitaba explicaciones sobre bases bioquímicas y su vínculo con la terapia, y observaron un incremento significativo en el subdominio “Creando contexto” (p = 0.046).

En el ámbito de simulaciones con robot social, Borg et al. (2024) compararon un “paciente robótico” potenciado por LLM con un simulador puramente informático, y aunque centraron su evaluación en la experiencia de consulta, múltiples estudiantes señalaron que las preguntas del robot los llevaban a conectar directamente conceptos de farmacología con decisiones de manejo (del estilo de “me hizo explicar por qué la dosis X para el fármaco Y según la enfermedad Z”).

Adicionalmente, Y. Xu et al. (2025) implementaron un módulo e-MedLearn en Aprendizaje Basado en Problemas (ABP) que guiaba a los estudiantes, mediante prompts del LLM, a recoger pistas de laboratorio y enlazarlas con situaciones clínicas, logrando mejoras significativas en cada dimensión del análisis de caso (con diferencias de entre 1.5 y más de 2 puntos, $p < 0.05$ o $p < 0.01$ según la dimensión). Estos diseños multi secuencia, que alternan preguntas de fisiología con escenarios de diagnóstico, configuran un modelo de prompt engineering que actúa como andamiaje cognitivo, reforzando la integración de saberes.

También, lo hallazgos muestran que los LLMs facilitan la articulación de conocimientos básicos y clínicos mediante secuencias de prompts que guían al alumno a explicitar conexiones fisiopatológicas en cada paso diagnóstico. Sin embargo, la heterogeneidad en el diseño de casos (desde simulaciones robóticas hasta ejercicios de ABP) y la variabilidad en la formación previa de los participantes sugieren cautela al generalizar estos resultados. En la discusión, exploraremos cómo el nivel de experiencia y el tipo de prompts condicionan la magnitud de la integración y la potencial necesidad de un entrenamiento específico en prompt engineering para maximizar el beneficio interdisciplinar.

De otro lado, los estudios de integración incluyeron predominantemente estudiantes de pregrado en semestres avanzados (fase de ciencias clínicas, entre el 3 y 6 año), con edades entre 22 y 27 años y experiencia previa limitada con LLMs. La mayoría (70%) reportó haber utilizado modelos solo en la intervención, sin formación previa formal, lo que sugiere que el efecto integrador ocurre incluso en usuarios noveles.

4.3 Percepciones sobre el uso de los LLMs: beneficios, limitaciones y barreras (objetivo O4)

En esta sección se presentan los juicios y valoraciones recogidos en los estudios sobre el uso de los LLMs en contextos clínico-educativos, abarcando tanto las percepciones de estudiantes como de docentes. Se describen los beneficios percibidos en términos de claridad y profundidad de la retroalimentación, así como las principales limitaciones y barreras (técnicas, éticas y organizacionales) que emergen de la experiencia de usuario y de los soportes implementados.

Figura 4. Beneficios percibidos de los LLMs

4.3.1 Percepciones de los estudiantes

Quince de los 17 estudios (~88%) registraron al menos un beneficio asociado al uso de LLMs (ver

Figura 4). Los más frecuentes fueron:

- Mejora del razonamiento clínico (9 estudios).
- Retroalimentación inmediata o detallada (4).
- Simulación realista de casos (5).
- Generación o síntesis de resúmenes (2).
- Apoyo en redacción académica (3).

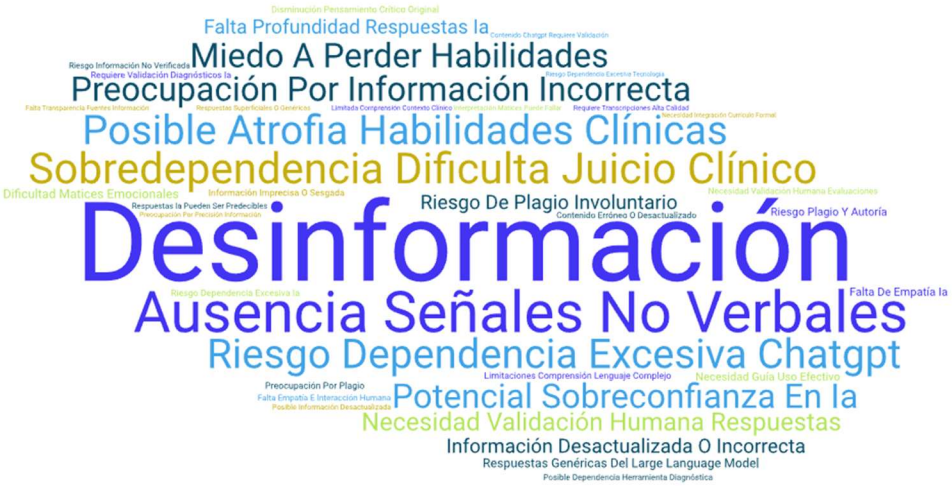
Por otro lado, once estudios (~65%) documentaron preocupaciones (ver Figura 5), tales como la interferencia en el desarrollo de habilidades, destacando en seis la dependencia excesiva en la IA, en cinco la información desactualizada o incorrecta y en cuatro la falta de matices o empatía. Siete trabajos (~41%) aludieron a aspectos éticos, donde la privacidad de datos y la confidencialidad fueron mencionadas en seis y la presencia de sesgos algorítmicos en cuatro. Finalmente, cuatro estudios (~24%) reportaron barreras de acceso, tales como la brecha digital, la necesidad de conexión estable y el costo de dispositivos.

4.3.2 Percepciones de los docentes

Trece de los 17 estudios (~76%) recogieron una opinión general positiva, mientras que uno (~6%) la consideró mixta y tres (~18%) no la midieron. En cinco publicaciones se

resaltó la eficiencia para crear materiales y la consistencia de la retroalimentación. Solo dos estudios (~12%) manifestaron temor a la erosión de habilidades pedagógicas, y uno (~6%) expresó preocupaciones sobre el rol profesional.

Figura 5. Preocupaciones sobre el uso de LLMs



4.3.3 Experiencia de usuario y soporte

Catorce estudios (~82%) evaluaron la utilidad y confiabilidad de los LLMs:

- Utilidad alta o muy alta en 14 trabajos.
- Confiabilidad media en 8.
- Confiabilidad baja o no reportada en 9.

En cuanto al entrenamiento, ocho estudios (~47%) describieron algún tipo de soporte a estudiantes: briefings, tutoriales o guías operativas previas a la actividad. Solo tres (~18%) detallaron programas de capacitación para docentes, generalmente sesiones combinadas o demostraciones de la plataforma.

4.3.4 Barreras y facilitadores técnicos

Todos los estudios (17/17) documentaron la modalidad de acceso. Las más comunes fueron:

- Interfaz web oficial (5 estudios).
- Plataformas de simulación especializadas (2).
- Robots sociales (2).
- APIs o integraciones (1).
- Aplicaciones móviles o entornos mixtos (1).
- Otros formatos (WeChat, sistemas PBL+CR) en los restantes.

Diez estudios (~59%) informaron el costo asociado:

- Acceso gratuito o institucional en 6.
- Pago por API en 5.
- Modalidades mixtas (gratuito + suscripción) en 3.

4.3.5 Síntesis parcial y proyección hacia la discusión

Estos resultados muestran un amplio entusiasmo por los LLMs, pero también resaltan preocupaciones éticas y tecnológicas que pueden frenar la adopción institucional. La desproporción entre el soporte ofrecido a estudiantes y la limitada formación a docentes plantea interrogantes sobre la sostenibilidad del modelo educativo. Adicionalmente, la opacidad en las modalidades de costos y plataformas tecnológicas sugiere la necesidad de analizar en la discusión cómo estos factores condicionan la escalabilidad y equidad de las experiencias asistidas por LLMs.

4.4 Evidencia sobre la transferencia percibida del razonamiento clínico a la práctica real (objetivo O3)

En esta subsección se expone cómo los estudios incluyen tanto percepciones cualitativas de los estudiantes sobre su capacidad de trasladar lo aprendido a entornos reales, como indicadores cuantitativos indirectos que sugieren esa transferencia.

4.4.1 Percepciones cualitativas de transferencia

Siete de los 17 estudios (41.2%) recogieron testimonios de los participantes acerca de su confianza y disposición para aplicar en la práctica real las habilidades desarrolladas con LLMs. Los temas más frecuentes fueron:

- “Confianza al enfrentar pacientes reales” (4 estudios)
- “Reducción de errores en ejercicio clínico” (2)
- “Preparación para toma de decisiones en rotaciones” (2)

Como ejemplo, en el de Brügge et al. (2024), varios médicos internos (o estudiantes de último año de pregrado de medicina) expresaron que la retroalimentación estructurada “mejoró mi confianza al atender pacientes reales”, y en el de Hicke et al. (2025) indicaron que la simulación con MedSimAI les permitió “practicar sin temor a dañar al paciente”, fortaleciendo su sentido de seguridad clínica.

4.4.2 Indicadores cuantitativos indirectos de transferencia

Seis estudios (35.3%) incorporaron métricas que, aunque no miden la transferencia directamente, señalan mejoras aplicables a la práctica:

- ECOE y pruebas de simulación clínica: cinco estudios reportaron incrementos significativos en puntajes, con medias que oscilaron entre +5 y +12 puntos tras la intervención.
- Exhaustividad y precisión en listas de síntomas o procesos: tres estudios alcanzaron > 90% de exactitud en pruebas de historia clínica estructurada.
- Reducción de errores de razonamiento: Brügge et al. (2024) documentó una caída significativa de errores en “Crear contexto” ($p = 0.046$) y “Asegurar información” ($p = 0.018$) con la herramienta CRI-HTI.

Ejemplo: el estudio de Ba et al. (2024) mostró una mejora de 8.3 ± 1.4 a 90.8 ± 3.2 puntos en un ECOE pediátrico después de sesiones asistidas por ChatGPT, un indicio de que esas competencias se consolidan en entornos simulados de alta fidelidad.

Estos hallazgos sugieren que, aunque la mayoría de los estudios no evalúa la transferencia en entornos reales, las percepciones de los estudiantes y los avances en pruebas de simulación respaldan la hipótesis de que los LLMs pueden facilitar la aplicación práctica del razonamiento clínico. En la discusión exploraremos la validez ecológica de estas aproximaciones y los retos para mantener la transferencia a largo plazo.

DISCUSIÓN

El razonamiento clínico es el proceso mediante el cual el médico integra conocimientos de las ciencias biomédicas básicas con la información clínica del paciente para formular hipótesis diagnósticas y planes de manejo adecuados. Se articula en un continuo que va desde el reconocimiento rápido de patrones (pensamiento intuitivo) hasta el análisis deliberado y reflexivo de la información (pensamiento analítico), tal como describen (Guzmán-Valdivia-Gómez et al. (2022), Kulasegaram et al. (2015) y los planteamientos de Kahneman (2013). Sin embargo, la educación médica tradicional tiende a fragmentar las ciencias básicas y las clínicas en asignaturas aisladas, lo que dificulta la construcción de puentes cognitivos que permitan al estudiante movilizar conocimientos de forma integrada (Delavari et al., 2024; Durning et al., 2024; Kassirer, 2010; Parodis et al., 2021).

En este contexto, los LLMs ofrecen la posibilidad de generar entornos de práctica adaptativos que alternan estímulos intuitivos (mediante exposición masiva a casos) con actividades que exigen justificaciones fundamentadas en saberes biomédicos y clínicos y que fomentan habilidades metacognitivas, potenciando así tanto el sistema 1 como el sistema 2 de pensamiento (Ambrose et al., 2017; Pears & Konstantinidis, 2024).

La presente discusión examina cómo las experiencias educativas asistidas por LLMs, en comparación con métodos tradicionales, influyen en el desarrollo y la transferencia del

razonamiento clínico y en la integración de saberes básicos y clínicos, respondiendo así a la pregunta central de la investigación (¿cómo influyen las experiencias educativas asistidas por LLMs, en comparación con métodos convencionales, en el desarrollo, la transferencia y la percepción del razonamiento clínico y la articulación interdisciplinar?) y a los objetivos que se definieron para responderla y que servirán como ejes temáticos para el análisis y discusión de los resultados de la revisión sistemática de los 17 estudios incluidos.

1 Impacto en competencias de razonamiento clínico (O1)

Para comprender cómo las experiencias asistidas por LLMs influyen en el desarrollo de las habilidades de razonamiento clínico, analizamos dos dimensiones interrelacionadas: el reconocimiento de patrones (pensamiento intuitivo) y el análisis reflexivo (pensamiento analítico y metacognición).

1.1 Reconocimiento de patrones y fluidez diagnóstica

Varios estudios cuasiexperimentales demostraron que la exposición a simulaciones masivas generadas por LLMs refuerza la identificación rápida de patrones clínicos. Por ejemplo, Zheng et al. (2024) documentaron que quienes utilizaron un modo de simulación potenciado por IA obtuvieron una media de 82.91 ± 5.054 puntos en un examen de casos cardiovasculares, frente a 72.16 ± 7.595 en el grupo control ($t = 6.811$; $p < 0.001$), sugiriendo una mejora notable en la velocidad y precisión intuitiva al enfrentar “pacientes” virtuales. Estos hallazgos se alinean con la noción de aprendizaje por exposición, quienes describen cómo la práctica reiterada fortalece esquemas mentales que sustentan el reconocimiento de patrones (Guzmán-Valdivia-Gómez et al., 2022).

1.2 Análisis reflexivo y metacognición

Más allá de la rapidez inicial, la riqueza de los LLMs reside en guiar al estudiante en la justificación paso a paso del diagnóstico, estimulando la reflexión y el monitoreo de sus propios procesos cognitivos. Brügge et al. (2024) midieron el CRI-HTI en historia clínica y hallaron que el grupo asistido mejoró de 3.02 ± 0.12 a 3.60 ± 0.13 ($F(1,18) = 4.44$; $p = 0.049$), con aumentos significativos en “Creando contexto” ($p = 0.046$) y “Asegurando información” ($p = 0.018$). Estos resultados evidencian que, al demandar al estudiante que explique fisiopatología y epidemiología en cada paso, los prompts de un LLM funcionan como andamiaje cognitivo, reforzando la planificación, la supervisión y la evaluación metacognitiva (Ambrose et al., 2017).

1.3 Integración de conocimientos básicos y clínicos

La articulación de saberes biomédicos con escenarios clínicos es esencial para un diagnóstico sólido. (Bharatha et al., 2024) compararon el rendimiento de ChatGPT-4 y de los estudiantes de medicina en MCQs: 73.7 % de aciertos para el modelo frente a 66.7 % de media estudiantil, y en pruebas preclínicas específicas ChatGPT-4 alcanzó 77.7 % de aciertos. Estos datos confirman que los LLMs no solo reproducen patrones, sino que integran conceptos básicos en su razonamiento clínico. Asimismo, Kulasegaram et al. (2015) demostraron cómo los vínculos causales explícitos entre ciencia básica y clínica mejoran la retención y aplicación, un principio que se replica cuando los LLMs exigen justificaciones mecanicistas en cada paso diagnóstico.

1.4 Precisión en instrumentos estandarizados

En contextos de evaluación formal, los LLMs han mostrado resultados competitivos. Ba et al. (2024) utilizaron el Mini-CEX en educación pediátrica y hallaron que el 66.6 % de los estudiantes asistidos por ChatGPT superaron expectativas, versus 36.8 % en controles ($p = 0.032$), aunque sin diferencias en conocimientos teóricos ($p = 0.768$). Esto sugiere que la IA influye especialmente en sub-competencias prácticas y en la confianza al ejecutar tareas clínicas.

1.5 Variabilidad y condiciones de uso

No todos los estudios evidenciaron efectos uniformes: C. Chan et al. (2023) evaluaron GutGPT en riesgo de sangrado gastrointestinal y encontraron que el “Content Mastery” mejoró de ≈ 58 % a ≈ 70 % postintervención, frente a un aumento de ≈ 55 % a ≈ 68 % en el grupo control, lo que sugiere que la magnitud del beneficio depende del dominio clínico y del diseño de la herramienta. Delavari et al. (2024), Durning et al. (2024) y Parodis et al. (2021) advierten que el contexto, el nivel de experiencia y la formación previa en razonamiento condicionan los beneficios de herramientas automatizadas, por lo que es esencial adaptar las intervenciones a las competencias de cada grupo de estudiantes.

En conjunto, los LLMs potencian el razonamiento clínico al combinar exposición masiva a casos (mejorando el reconocimiento de patrones intuitivos) con ejercicios de justificación estructurada (fomentando la reflexión analítica y metacognición). Este doble impacto se inscribe en modelos constructivistas que integran teoría básica y práctica clínica, ofreciendo un entorno de simulación seguro y adaptativo para fortalecer la competencia diagnóstica.

2 Facilitación de la integración conceptual interdisciplinar (O2)

Para que los estudiantes logren articular de manera fluida saberes de ciencias biomédicas básicas y conocimientos clínicos es preciso diseñar actividades que promuevan conexiones explícitas entre ambos dominios. Desde la teoría de la integración cognitiva, se sabe que la formación fragmentada en asignaturas aisladas dificulta la construcción de esquemas que relacionen causas fisiopatológicas con decisiones diagnósticas y terapéuticas (Kulasegaram et al., 2015). En este sentido, la "ingeniería de prompts" o "prompt engineering", como se reconoce más ampliamente en el mundo de la IA, emerge como un andamiaje pedagógico clave para guiar al modelo en la generación de ejercicios que exijan fundamentar cada paso en conceptos básicos y, al mismo tiempo, aplicar dichos conceptos en contextos clínicos (Broshar, 2024; Craig et al., 2024).

2.1 Ingeniería de prompts como andamiaje de integración

Los prompts bien estructurados reducen la ambigüedad en la respuesta del LLM y orientan la atención del estudiante hacia explicaciones mecanicistas y vinculaciones causales (Broshar, 2024; Craig et al., 2024). En la práctica, Öncü et al. (2025) implementaron un diseño cuasiexperimental, donde cada prompt requería justificar diagnóstico y tratamiento en base a principios fisiopatológicos; esto se tradujo en una puntuación media de 8.33 ± 1.35 en la Observación de competencia en Razonamiento Clínico (escala 0–10) en una prueba tipo ECOE.

2.2 Diseño de actividades interdisciplinarias en ABP y simulaciones

Más allá de casos aislados, los diseños multi secuencia demuestran una eficacia notable. Y. Xu et al. (2025) crearon un módulo de Aprendizaje Basado en Problemas potenciado por GPT-3.5-turbo-16k que alternaba preguntas de fisiología con escenarios clínicos. El grupo de intervención mejoró en todas las dimensiones del análisis de caso entre 1.5 y 2.3 puntos ($p < 0.05$–0.01). De igual manera, Brügge et al. (2024) diseñaron simulaciones en dos fases: primero simular al paciente, luego pedir explicaciones bioquímicas, y observaron un incremento significativo en el subdominio "Creando contexto" ($p = 0.046$) y en "Asegurando información" ($p = 0.018$). Estas intervenciones confirman que el LLM puede funcionar como tutor interactivo que refuerza tanto la aplicación de conceptos básicos como el razonamiento clínico.

2.3 Heterogeneidad de intervenciones y necesidad de medición explícita

Pese a estos éxitos, la gran diversidad en formatos (paciente robótico, ABP, simulaciones conversacionales) y la ausencia de formación previa en prompt engineering en la mayoría de los participantes (70% usuarios inexpertos) introduce variabilidad en los resultados y dificulta la generalización. Además, ninguno de los estudios cuantitativos

midió directamente la integración saberes básicos-clínicos mediante evaluaciones diseñadas para ello, lo que evidencia la urgencia de instrumentos específicos que permitan cuantificar de manera precisa este constructo y guiar futuras intervenciones pedagógicas.

3 Transferencia percibida del razonamiento clínico a la práctica real (O3)

Este apartado explora hasta qué punto las habilidades desarrolladas con LLMs en entornos formativos se trasladan a la práctica clínica, combinando evidencias cualitativas de percepción con indicadores cuantitativos indirectos.

3.1 Comparación entre entornos simulados y reales

La mayoría de los estudios evaluaron competencias en escenarios simulados de alta fidelidad (ECOE, Mini-CEX, CRI-HTI) que, aunque por su naturaleza buscan precisamente replicar de forma realista las situaciones clínicas, por ser simulaciones con pacientes virtuales, y a sabiendas de las limitaciones tecnológicas actuales en los distintos modelos de LLM o IA en general disponibles, no pueden igualar a la medición directa en contextos reales. Sin embargo, las percepciones de los participantes sobre la aplicabilidad de lo aprendido resultan ilustrativas. Siete de los 17 trabajos recogieron testimonios sobre la confianza al enfrentar pacientes reales, la reducción de errores y la preparación para rotaciones clínicas, con afirmaciones sobre la práctica con LLMs tales como “mejoró mi confianza al atender pacientes reales” y “permitió practicar sin temor a dañar al paciente”, evidentemente siendo hallazgos positivos.

3.2 Indicadores objetivos de desempeño

Aunque indirectos, seis estudios reportaron métricas que sugieren transferencia al mundo real:

- Mejoras en ECOE y simulaciones clínicas: cinco investigaciones documentaron incrementos de entre 5 y 12 puntos en evaluaciones de competencia clínica tras la intervención.
- Precisión en listas de síntomas y procesos: tres estudios alcanzaron más del 90% de exactitud en pruebas de historia clínica estructurada.
- Reducción de errores de razonamiento: Brügge et al. (2024) observaron descensos significativos en fallos de “Crear contexto” ($p = 0.046$) y “Asegurar información” ($p = 0.018$) en la herramienta CRI-HTI.

- ECOE pediátrico con ChatGPT-4.0: Ba et al. (2024) mostraron una mejora de 8.3 ± 1.4 a 90.8 ± 3.2 puntos en un examen clínico objetivo, indicando que las competencias afianzadas en simulación pueden consolidarse en procesos evaluativos de alta demanda.

3.3 Percepciones de confianza y preparación

Más allá de los números, las voces de los estudiantes y residentes aportan un matiz clave:

- Confianza clínica: múltiples participantes valoraron que el entrenamiento con LLMs fortaleció su seguridad al tomar decisiones en entornos reales.
- Reducción del miedo al error: la posibilidad de equivocarse sin riesgo real aumentó el valor formativo de la simulación.
- Preparación para las rotaciones: algunos participantes destacaron que las sesiones con IA les dieron un "ensayo" práctico antes de enfrentarse a pacientes auténticos.

Estos datos cualitativos, junto con los indicadores indirectos de desempeño, respaldan la hipótesis de que los LLMs pueden facilitar la transferencia del razonamiento clínico a la práctica, aunque son necesarias mediciones directas en entornos reales y evaluaciones longitudinales para confirmar la sostenibilidad de estos efectos.

4 Percepciones de beneficios, limitaciones y barreras (O4)

Finalizando este análisis temático, es pertinente ahora profundizar las visiones de estudiantes y docentes sobre el uso de LLMs en la formación médica, incluyendo sus aspectos éticos y de equidad.

4.1 Beneficios percibidos

Los alumnos valoran especialmente la inmediatez de la retroalimentación y la adaptabilidad de los casos simulados. Por ejemplo, en el estudio hecho en Suecia por Borg et al. (2025a), los participantes calificaron la "preparación para paciente real" con 4.4 ± 0.6 en la plataforma robótica social versus 4.0 ± 0.7 en la convencional (OR 4.2; p = .009), así como la adecuación del nivel de dificultad (4.7 ± 0.5 vs. 4.3 ± 1.0; OR 2.4; p = .08). Estos resultados subrayan que los LLMs ofrecen entornos seguros donde "practicar sin temor a dañar al paciente" mejora la confianza clínica.

Adicionalmente, la autonomía en el estudio se refleja en la frecuencia de uso: el 69% de los estudiantes en los estudios de Estados Unidos usó LLMs para fines médicos al menos una vez al mes, y un 35% de residentes e internos los empleó mensualmente en tareas

clínicas (y 52.5% de ellos en tareas no clínicas), lo cual implica una percepción implícita de utilidad y eficacia.

4.2 Limitaciones técnicas y formativas

Pese a estos beneficios, la variabilidad en la calidad de los prompts y la familiaridad digital condicionan el impacto. En Alemania, Aster et al. (2025) encontraron un nivel medio de experiencia previa con ChatGPT (2.48 ± 0.98 en escala 1–5), lo que sugiere que muchos estudiantes requieren capacitación en prompt engineering para aprovechar plenamente la herramienta, incluso en entornos que pueden probablemente contar ya con recursos abundantes de todo tipo. Desde el marco teórico, se advierte que, sin una alfabetización digital robusta, la adopción puede ser superficial (Amiri et al., 2024; Buabbas et al., 2023) y esto es esperable que sea un reto incluso mayor en instituciones educativas con recursos limitados, en los que ya existe una importancia brecha tecnológica y digital, incluso sin tener a la IA en consideración.

4.3 Barreras de infraestructura y culturales

En línea con lo anterior, en contextos rurales, solo el 28% de las instituciones contaba con ancho de banda suficiente para simulaciones asistidas por IA (Shakur et al., 2024), y los costes asociados a APIs comerciales pueden representar hasta el 5% del presupuesto anual de un departamento de educación médica (C. Chan et al., 2023).

Adicionalmente, existe resistencia al cambio, que, aunque a veces se apalanca en argumentos fuertes, que subyacen a muchas de las preocupaciones que expertos señalan sobre la implementación de la IA en distintos entornos educativos, no dejan de ser llamativos y complejos. Algunos docentes temen la deshumanización de la enseñanza y la pérdida de la interacción directa con el estudiante, lo que exige estrategias de sensibilización y acompañamiento para fomentar la confianza en el uso de LLMs.

4.4 Consideraciones éticas y de equidad

Plagio y honestidad académica: Ba et al. (2024) reportaron que un 42% de los internos de pediatría reconoció reutilizar texto generado por ChatGPT sin citarlo, lo que subraya la necesidad de protocolos claros de citación y formación en uso ético de la IA.

Sesgos algorítmicos: Huang et al. (2024) documentaron casos donde los modelos reproducían estereotipos de género y procedencia geográfica, señalando la urgencia de auditorías periódicas y supervisión humana obligatoria.

Privacidad de datos clínicos: C. Chan et al. (2023) describieron inquietudes sobre la seguridad de historiales médicos al ingresarlos en APIs externas, sin contar con otro tipo de datos sensibles, incluso de las instituciones educativas, los docentes o los estudiantes, lo que exige marcos de protección y anonimización rigurosos y consideraciones que incluso trascienden, de nuevo, hacia la discusión sobre disponibilidad de recursos y costos, pues no todas las instituciones pueden adquirir y administrar servidores propios.

Equidad de acceso: Sin políticas institucionales que garanticen recursos compartidos, la brecha digital puede profundizar desigualdades socioeconómicas y geográficas, limitando el potencial transformador de los LLMs para todos los estudiantes.

CONCLUSIONES

A continuación, sintetizamos los hallazgos clave de cada objetivo para extraer aprendizajes centrales y trazar líneas de acción:

- En primer lugar, los hallazgos cuantitativos y cualitativos confirman que los LLMs fortalecen de manera convergente dos componentes esenciales del razonamiento clínico: la detección rápida de patrones y la reflexión analítica. Estudios como Zheng et al. (2024) y Hicke et al. (2025) muestran mejoras de marcadas en la precisión diagnóstica y autoeficacia, mientras que Brügge et al. (2024) demuestran incrementos significativos en subdimensiones metacognitivas (CRI-HTI). Este doble impacto refleja un andamiaje cognitivo que alterna la práctica masiva de casos con prompts que exigen justificaciones fisiopatológicas, validando los postulados de Ambrose et al. (2017) y Guzmán-Valdivia-Gómez et al. (2022) sobre la necesidad de un aprendizaje profundo e integrado.
- En segundo lugar, la “ingeniería de prompts” y los diseños de Aprendizaje Basado en Problemas, como el del estudio de Y. Xu et al. (2025), ilustran cómo es posible articular intencionadamente saberes básicos y clínicos. Al exigir que cada fase diagnóstica se fundamente en principios de fisiología y anatomía, estas intervenciones facilitan la construcción de esquemas conceptuales integrados, tal como exalta Kulasegaram et al. (2015). Sin embargo, la ausencia de mediciones directas de esta integración en los estudios revisados subraya la urgencia de desarrollar instrumentos específicos para cuantificar el impacto de la IA en la integración básico-clínica y orientar futuras innovaciones pedagógicas.
- En tercer lugar, aunque la transferencia a la práctica real ha sido evaluada principalmente en entornos simulados de alta fidelidad, los indicadores indirectos (Ba et al., 2024; Borg et al., 2024, 2025b) sugieren que las habilidades afianzadas con

LLMs pueden consolidarse en contextos clínicos auténticos. No obstante, se requieren estudios longitudinales y evaluaciones in situ para confirmar la durabilidad de estos beneficios y su traducción efectiva al cuidado de pacientes reales.

- Finalmente, el análisis de percepciones revela un balance complejo entre beneficios (retroalimentación inmediata, adaptabilidad, autonomía de estudio) y desafíos (capacitación insuficiente en diseño de prompts, brecha digital, preocupaciones por costos, confidencialidad, éticas y de equidad). Por ejemplo, la alta adopción de LLMs por estudiantes en algunos de los estudios convive con un 42% de reportes de plagio no intencional (Ba et al., 2024) y casos de sesgos algorítmicos (Huang et al., 2024). Esto recalca la necesidad de marcos de gobernanza que incluyan formación obligatoria en ética de la IA, auditorías de sesgos y políticas de acceso equitativo.

En conjunto, estas conclusiones sitúan a los LLMs como herramientas con un notable potencial para transformar la enseñanza del razonamiento clínico, siempre que su implementación se articule con un diseño curricular alineado, evaluaciones rigurosas y políticas institucionales que garanticen calidad, equidad y responsabilidad.

LIMITACIONES DEL ESTUDIO

A pesar de la rigurosidad del proceso y la riqueza de los datos recabados, esta revisión presenta varias limitaciones que deben considerarse al interpretar los hallazgos:

- Heterogeneidad metodológica y de muestras: tras la búsqueda y cribado de estudios en bases de datos bibliográficas, pese al altísimo volumen inicial de resultados, solo 17 estudios cumplían los criterios para ser incluidos en esta revisión, por lo que se debe cuestionar el alcance de esta investigación. Además, los estudios incluidos abarcan diseños diversos (ensayos aleatorizados, cuasiexperimentales, transversales y mixtos), con tamaños de muestra muy dispares (de 15 a 2027 participantes; mediana = 64). Esta variabilidad dificulta la comparación directa de magnitudes de efecto y la generalización de conclusiones a poblaciones más amplias.
- Riesgo de sesgo en los ensayos aleatorizados: de los dos ensayos controlados aleatorizados evaluados con RoB-2, ninguno documentó adecuadamente la generación de la secuencia aleatoria, el ocultamiento de la asignación ni la comparabilidad inicial de los grupos (0% de cumplimiento), lo que incrementa el riesgo de sesgo de selección. Además, el manejo de datos incompletos y el reporte selectivo mostraron un cumplimiento moderado (50%).
- Mediciones incompletas de la integración y la transferencia: aunque varias intervenciones implementaron prompts para vincular ciencias básicas y clínicas, ninguno midió explícitamente este constructo a través de evaluaciones integradoras. De igual modo, la transferencia a la práctica real se infiere de simulaciones y percepciones, sin estudios in situ ni seguimiento longitudinal que comprueben la sostenibilidad de los efectos.

- Deficiencias en el reporte técnico de los LLMs: solo 3 de 17 estudios detallaron parámetros de configuración de los modelos (temperatura, top_p) y la plataforma de ejecución. La falta de transparencia en estos aspectos limita la replicabilidad de las intervenciones y la interpretación de la relación entre configuración del modelo y resultados formativos.
- Posible sesgo de publicación y aprendizaje positivo: ninguno de los estudios reportó hallazgos negativos o neutros de forma explícita, y existe un predominio de resultados favorables, tanto con significancia estadística como sin ella, que podría reflejar un sesgo de publicación hacia intervenciones exitosas. Asimismo, la mayoría usa diseños de corta duración, lo que impide evaluar el fenómeno de “novelty effect” y la verdadera sostenibilidad de los beneficios.

RECOMENDACIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

1 Investigación

- Estudios longitudinales y multicéntricos: implementar diseños aleatorizados con seguimiento longitudinal a mediano y largo plazo, para evaluar la durabilidad de los efectos en contextos reales.
- Desarrollo de instrumentos específicos: crear y validar herramientas que midan la integración conceptual básico–clínica.
- Estandarización de reportes técnicos: exigir la documentación completa de parámetros de modelo (temperatura, top_p, versión) y la descripción detallada de los prompts utilizados.
- Ensayos in situ y evaluaciones reales: medir la transferencia práctica mediante Mini-CEX y ECOE en rotaciones clínicas, complementados con métricas de desempeño real y retroalimentación de supervisores.
- Formación en prompt engineering: investigar el impacto de la capacitación previa en diseño de prompts sobre la eficacia de la interacción con LLMs y los resultados de aprendizaje.

2 Práctica educativa

- Programas de desarrollo docente: diseñar talleres de prompt engineering y ética de IA, que incluyan práctica guiada y reflexión sobre sesgos algorítmicos.
- Repositorios de prompts validados: crear bancos institucionales de prompts categorizados por competencias y nivel formativo, actualizados continuamente con evidencia de efectividad.

- Integración curricular: incorporar módulos de simulación asistida por LLMs y metodologías inductivas de aprendizaje activo, alineados con estándares de competencias clínicas, evaluando tanto precisión diagnóstica como metacognición.
- Evaluación formativa continua: combinar simulaciones estandarizadas (ECOE, Mini-CEX) con retroalimentación automatizada y sesiones de “debriefing” para fortalecer la autorregulación y metacognición.
- Alfabetización digital: ofrecer recursos y guías accesibles para estudiantes y docentes, mitigando la brecha digital y promoviendo la inclusión.

3 Política institucional

- Protocolos de gobernanza y ética: establecer auditorías periódicas de sesgos algorítmicos, supervisión humana obligatoria y formación en protección de datos clínicos.
- Inversión en infraestructura: garantizar ancho de banda, hardware y licencias necesarias, con especial atención a instituciones con recursos limitados para asegurar equidad de acceso.
- Alianzas interinstitucionales y fondos compartidos: promover colaboraciones y fondos dedicados a laboratorios de IA educativos y repositorios de recursos.
- Indicadores de calidad: incorporar métricas de uso, eficacia y equidad de LLMs en procesos de acreditación y evaluación institucional.
- Políticas de acceso abierto: fomentar la publicación de prompts, guías y mejores prácticas en acceso libre, facilitando la transparencia y colaboración en la adopción de LLMs en educación médica.

Tabla 2. Tabla de resultados

Variable /Subvariable	ESTUDIO N° 1	ESTUDIO N° 3	ESTUDIO N° 32	ESTUDIO N° 36	ESTUDIO N° 45	ESTUDIO N° 59	ESTUDIO N° 81	ESTUDIO N° 177	ESTUDIO N° 347	ESTUDIO N° 350	ESTUDIO N° 414	ESTUDIO N° 500	ESTUDIO N° 943	ESTUDIO N° 959	ESTUDIO N° 1127	ESTUDIO N° 1138	ESTUDIO N° 1155
Título del artículo	Virtual Patient Simulations Using Social Robotics Combined With Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study	AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians	Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial	A Pilot Study of Medical Student Opinions on Large Language Models	Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology	Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education	Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy	Exploring medical student's outlook on use of artificial intelligence in medical education: a multicentric cross-sectional study in Jharkhand, India	Advancing Problem-Based Learning with Clinical Reasoning for Improved Differential Diagnosis in Medical Education	MedSimAI: Simulation and Formative Feedback Generation to Enhance Deliberate Practice in Medical Education	Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis	Assessing the Usability of GutGPT: A Simulation Study of an AI Clinical Decision Support System for Gastrointestinal Bleeding Risk	Using ChatGPT for medical education: the technical perspective	Large language models in internal medicine residency: current use and attitudes among internal medicine residents	ChatGPT as a Virtual Patient: Written Empathic Expressions During Medical History Taking	Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods	Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT
Autor/es	Borg et al	Öncü et al	Brügge et al	Xu et al	Borg et al	Zheng et al	Bhartha et al	Ahmad et al	Xu et al	Hicke et al	Shakur et al	Chan et al	Chan et al	Fried et al	Aster et al	Ba et al	Huang et al
Año de publicación	2025	2025	2024	2024	2024	2024	2024	2024	2025	2025	2024	2023	2025	2024	2025	2024	2024
Tipo de diseño del estudio	Estudio de intervención no aleatorizado (cuasiexperimental)	Obs. Transversal	Estudio de intervención no aleatorizado (cuasiexperimental)	Obs. Transversal	Obs. De Prevalencia o descriptivo	Estudio de intervención no aleatorizado (cuasiexperimental)	Obs. Transversal	Obs. Transversal	Experimental aleatorizado	Estudio de intervención no aleatorizado (cuasiexperimental)	Obs. De Prevalencia o descriptivo	Experimental aleatorizado	Obs. Transversal	Obs. Transversal	Estudio de intervención no aleatorizado (cuasiexperimental)	Experimental aleatorizado	Experimental aleatorizado

Enfoque principal del estudio	Mixto	Mixto	Cuantitativo	Cuantitativo	Cualitativo	Mixto	Cualitativo	Mixto	Mixto	Mixto	Mixto	Mixto	Mixto	Cuantitativo	Mixto	Mixto	Cuantitativo
Tamaño muestral del estudio	15	21	21	102	23	66	96	417	18	104	2027	55	45	152	35	77	64
Tamaño del grupo de intervención (si aplica)	15	21	10	N/A	N/A	34	N/A	N/A	9	104	N/A	Fase 1: 24; Fase 2: 28	45	N/A	35	39	32
Tamaño del grupo de control (si aplica)	15	0	11	N/A	N/A	32	96	N/A	9	0	N/A	Fase 1: 31; Fase 2: 23	N/A	N/A	N/A	38	32
Estudiantes: Opinión general	positiva	positiva	Positiva	Positiva	Positiva	Positiva	No reportado	Positiva	Positiva	Positiva	No reportado	Neutra o mixta	Positiva	Positiva	No reportado	Positiva	Positiva
Estudiantes: Beneficios más destacados	Mejora Razonamiento Clínico (CR); Formulación libre preguntas; Experiencia realista/atractiva; Explorar diagnósticos diferenciales; Fomenta autorreflexión	Mejora gestión casos clínicos; Retroalimentación detallada inmediata; Simulación realista casos diversos; Práctica segura errores médicos	Mejora toma decisiones clínicas; Retroalimentación estructurada útil; Simulación realista paciente; Aumenta confianza diagnóstica	Generación rápida de resúmenes; Ayuda redacción trabajos académicos; Potencial tutoría personalizada; Acceso fácil información médica	Entrevistas clínicas más naturales; Interacción atractiva y dinámica; Permite explorar hipótesis diagnósticas; Retroalimentación cualitativa valiosa	Mejora comprensión enfermedades cardiovasculares; Aumenta interés aprendizaje activo; Simulación efectiva escenarios clínicos; Optimiza métodos enseñanza	N/R (evaluación rendimiento LLM)	Potencial mejorar aprendizaje médico; Acceso rápido vasta información; Ayuda investigación y redacción; Herramienta educativa innovadora	Mejora razonamiento diagnóstico diferencial; Facilita aprendizaje basado problemas; Simula procesos pensamiento experto; Interfaz intuitiva y útil	Simulación realista casos médicos; Retroalimentación formativa instantánea; Práctica deliberada mejora habilidades; Adaptable a necesidades aprendizaje	Análisis rápido transcripciones OSCE; Evaluación objetiva y consistente; Ahorro tiempo para evaluadores; Identificación áreas mejora	Ayuda toma decisiones clínicas; Acceso rápido guías relevantes; Presentación clara factores riesgo; Interfaz de usuario amigable	Acceso rápido información médica; Ayuda resumir literatura compleja; Apoyo redacción científica; Generación ideas investigación	Rápida obtención de información; Ayuda en documentación clínica; Soporte decisiones diagnósticas; Facilita comunicación con pacientes	N/R (foco en expresiones IA)	Mejora habilidades clínicas pediátricas; Adquisición conocimientos más rápida; Aumenta confianza estudiantes; Retroalimentación ChatGPT útil	Optimiza modelo enseñanza quirúrgica; Mejora eficiencia aprendizaje teórico; Aumenta satisfacción estudiantes; Simula discusión casos clínicos

						tradicionales					estudiantes						
Estudiantes: Preocupaciones por interferencia en el desarrollo de saberes, habilidades o competencias	Desinformación Large Language Model (LLM) impacta aprendizaje; Ausencia señales no verbales; Sobredependencia dificulta juicio clínico	Riesgo dependencia excesiva ChatGPT; Posible atrofia habilidades clínicas; Preocupación por información incorrecta	Potencial sobreconfianza en IA; Miedo a perder habilidades; Necesidad validación humana respuestas	Información desactualizada o incorrecta; Riesgo de plagio involuntario; Falta profundidad respuestas IA	Respuestas genéricas del Large Language Model (LLM); Dificultad matices emocionales; Limitaciones comprensión lenguaje complejo	No reportado	N/R (evaluación rendimiento LLM)	Información imprecisa o sesgada; Riesgo dependencia excesiva IA; Disminución pensamiento crítico original; Preocupación por plagio	Posible dependencia herramienta diagnóstica; Necesidad guía uso efectivo; Requiere validación diagnósticos IA	Respuestas IA pueden ser predecibles; Falta empatía e interacción humana; Riesgo información no verificada	Requiere transcripciones alta calidad; Interpretación matices puede fallar; Necesidad validación humana evaluaciones	N/R (estudio usabilidad prototipo)	Contenido erróneo o desactualizado; Falta transparencia fuentes información; Riesgo plagio y autoría; Limitada comprensión contexto clínico	Preocupación por precisión información; Respuestas superficiales o genéricas; Riesgo dependencia excesiva tecnología; Falta de empatía IA	N/R (foco en expresiones IA)	No reportado	Contenido ChatGPT requiere validación; Posible información desactualizada; Necesidad integración currículo formal
Estudiantes: Preocupaciones por aspectos éticos	Privacidad datos entrenamiento Large Language Model (LLM); Confidencialidad datos en simulación	Privacidad datos pacientes simulados; Confidencialidad interacciones estudiante-IA; Uso ético IA educación	No reportado	Privacidad datos ingresados estudiantes; Sesgos inherentes Large Language Models (LLMs); Propiedad intelectual contenido generado	No reportado	No reportado	N/R (evaluación rendimiento LLM)	Privacidad datos salud pacientes; Confidencialidad información personal; Uso ético IA medicina	No reportado	No reportado	Privacidad datos estudiantes examinados; Seguridad almacenamiento transcripciones OSCE; Sesgos potenciales en IA	N/R (estudio usabilidad prototipo)	Privacidad datos ingresados usuarios; Sesgos en datos entrenamiento; Responsabilidad legal información IA	Privacidad datos pacientes y médicos; Confidencialidad información sensible; Necesidad directrices uso ético	N/R (foco en expresiones IA)	No reportado	No reportado
Estudiantes: Preocu	No reportado	No reportado	No reportado	Acceso desigual a	No reportado	No reportado	N/R (evaluació	Brecha digital	No reportado	No reportado	No reportado	N/R (estudio	Requiere alfabetiz	Disponibilidad limitada	N/R (foco en	No reportado	No reportado

pacientes por aspectos relacionados al acceso a la tecnología				tecnología; Requiere conexión internet estable			n rendimiento LLM)	acceso tecnología; Costo dispositivos y software; Necesidad infraestructura tecnológica adecuada				usabilidad prototipo)	ación digital; Acceso equitativo a tecnología	en hospitales; Curva aprendizaje para usuarios	expresiones IA)		
Docentes: Opinión general (“positiva”, “negativa”, “neutra o mixta”, “no reportada”)	Positiva	Positiva	Positiva	Mixta	Positiva	Positiva	N/R (opinión no medida)	Positiva	Positiva	Positiva	Positiva (para evaluadores)	Positiva (respecto a usabilidad)	N/R (perspectiva técnica)	Positiva	N/R (no es opinión usuario)	Positiva	Positiva
Docentes: Beneficios más destacados para esos usos	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Mejora la comprensión y aplicación del conocimiento por los estudiantes; Mantiene a los estudiantes comprometidos e interesados; Efectivo para integrar conoci	No reportado	No reportado	Aumento de la Expectativa de Esfuerzo (percepción de facilidad de uso del sistema GutGPT); Potencial para mejorar el	N/A	No reportado	No reportado	No reportado	No reportado

									miento con práctica y transmitir experiencia; Ayuda a desarrollar el pensamiento médico y habilidades de razonamiento de los estudiantes; Facilita reflexiones valiosas para la mejora; Aborda la posible negligencia del pronóstico			apoyo a la decisión clínica; Mejora de la eficiencia en la recuperación de información.					
Docentes: Preocupaciones por interferencia en el desarrollo de saberes, habilidades o competencias	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	N/R (evaluación rendimiento LLM)	Preocupación calidad educación médica; Necesidad supervisión humana IA; Sobrecarga información para	No reportado	No reportado	No reportado	N/R (no aplica a docentes)	N/R (perspectiva técnica)	No reportado	N/R (foco en expresiones IA)	No reportado	Preocupación superficialidad aprendizaje; Rol docente sigue indispensable; Necesidad guiar uso crítico

								estudiantes									
Docentes: Preocupaciones asociadas a la estabilidad o seguridad laboral	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	N/R (evaluación rendimiento LLM)	No reportado	No reportado	No reportado	No reportado	N/R (no aplica a docentes)	N/R (perspectiva técnica)	No reportado	N/R (foco en expresiones IA)	No reportado	No reportado
Descripciones del efecto en la coherencia conceptual entre la teoría y práctica clínica	Permite practicar la aplicación de conocimientos médicos para dirigir preguntas relevantes basadas en síntomas difusos presentados por el paciente virtual (VP)	No reportado	Mejora en subdominios del Clinical Reasoning Indicator - History Taking Inventory (CRI-HTI) ('Creando contexto'; 'Asegurando información'); Retroalimentación fomentó pensar en causalidad de síntomas y recopilar información contextual	No reportado	No reportado	Profundización de la comprensión de las enfermedades; Mejora de las habilidades de pensamiento clínico	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Evidencias del impacto en razonamiento integrado en actividades educativas	Facilitación del entrenamiento del razonamiento clínico (CR) mediante diálogo activo; Práctic	No reportado	Mejora significativa en puntuación global del Clinical Reasoning Indicator - History Taking Inventory (CRI-HTI) en grupo de retroalimen	No reportado	No reportado	Mejora de las habilidades de pensamiento clínico; Mejora de las habilidades de trabajo en	No reportado	No reportado	No reportado	No reportado	No reportado	Los participantes en el brazo de GutGPT mostraron una mejora en el domini	No reportado	No reportado	No reportado	No reportado	No reportado

	a en la obtención de historial médico; Razonamiento sobre diferentes enfoques exploratorios; Planificación de estrategias de tratamiento; Posibilidad de hacer preguntas de seguimiento específicas basadas en las respuestas del paciente virtual (VP)		tación; Retroalimentación específica sobre cómo integrar información para el razonamiento			equipo y comunicación						o del contenido relacionado con el manejo del Sangrado Gastrointestinal (SG), según las evaluaciones educativas post-simulación.					
Docentes: Diseño macro o micro curricular	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Preparación de planes de enseñanza estructurados; Configuración de los ajustes de ChatGPT para	Diseño micro curricular de la rotación de cirugía

																alinearlos con los objetivos educativos	
Docentes: Diseño de actividades académicas	No reportado	No reportado	Diseño de simulación de paciente con LLM para practicar toma de historia clínica	No reportado	No reportado	Diseño e implementación de modo de enseñanza de simulación basada en escenarios potenciado por Inteligencia Artificial (IA)	No reportado	No reportado	Selección de casos para actividades de Aprendizaje Basado en Problemas (ABP)	No reportado	No reportado	No reportado	Crear/gestionar pacientes virtuales (añadir, leer, modificar, eliminar registros)	No reportado	No reportado	No reportado	Diseño de actividades de aprendizaje previas a la clase (previews); Diseño de actividades durante la clase (resolución de problemas); Diseño de actividades posteriores a la clase (revisión/consolidación)
Docentes: Creación de preguntas para cualquier actividad	No reportado	No reportado	Creación de prompts para simulación de pacientes; Creación de prompts para generación de retroalimentación	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Docentes: Revisión de calidad de preguntas para	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado

cualquier actividad																	
Docentes: Calificación de actividades distintas a exámenes	No reportado	No reportado	Generación de retroalimentación formativo basado en criterios (Clinical Reasoning Indicator - History Taking Inventory (CRI-HTI))	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Docentes: Calificación de exámenes	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Calificación del ítem 'resumen de la historia médica del paciente' en Exámenes Clínicos Objetivos Estructurados (ECOE) basados en transcripciones	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Docentes: Retroalimentación general sobre el desempeño	No reportado	No reportado	Provisión de retroalimentación general sobre el desempeño en la	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Revisión crítica del trabajo de los estudiantes con ChatGPT;	Provisión de retroalimentación inmediata a estudiantes

			toma de historia													Proporcionar retroalimentación inmediata y personalizada sobre el trabajo con ChatGPT; Ofrecer retroalimentación sobre el desempeño después de la práctica clínica; Responder preguntas de estudiantes	
Docentes: Retroalimentación específica sobre el desempeño	No reportado	No reportado	Generación de sugerencias específicas para la mejora del razonamiento clínico	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Revisión crítica del trabajo de los estudiantes con ChatGPT; Proporcionar retroalimentación inmediata y personalizada sobre el trabajo con ChatGPT; Ofrecer retroalimentación sobre el desempeño después de la	Provisión de retroalimentación inmediata a estudiantes

																práctica clínica; Responder preguntas de estudiantes	
Docentes: Generación de contenido o material para actividades específicas (presentaciones, resúmenes, imágenes, etc.)	Generación de respuestas de diálogo del paciente virtual (VP); Generación de expresiones faciales para el paciente virtual (VP)	No reportado	Generación de respuestas de paciente simulado; Generación de texto de retroalimentación	No reportado	No reportado	Utilización de la plataforma de simulación potenciada por Inteligencia Artificial (IA) para presentar escenarios clínicos y guiar el aprendizaje	No reportado	No reportado	No reportado	No reportado	No reportado	Obtención de respuestas a preguntas sobre el manejo clínico; Recuperación de recomendaciones basadas en evidencia y guías clínicas.	Creación de datos de pacientes virtuales (incluyendo prompts para ChatGPT)	No reportado	No reportado	No reportado	Generación de contenido para previsualizaciones previas a la clase (pre-class previews); Generación de contenido para consolidación y revisión posteriores a la clase
Docentes: Otros (especificar)	No reportado	No reportado	No reportado	No reportado	No reportado	Facilitación de sesiones de enseñanza utilizando el modo de simulación basado en escenarios potenciado por Inteligencia	No reportado	No reportado	Guiar discusiones (implícito por rol en Aprendizaje Basado en Problemas - ABP); Evaluar reportes de análisis de caso de estudiantes; Evaluar	No reportado	No reportado	Interacción con un dashboard integrado para la predicción de riesgo de Sangrado Gastrointestinal (SG) (modif	Usar la aplicación desde la vista de usuario (para pruebas/comprensión)	No reportado	No reportado	Orientación a los estudiantes sobre ChatGPT	Resolución de consultas de estudiantes durante la clase; Apoyo en la resolución de problemas durante la clase

						Artificial (IA)			dificultad de casos (para referencia en el sistema); Facilitar la reflexión del estudiante (objetivo del diseño)			icando covariables del paciente y observando cambios en el riesgo predicho).					
Estudiantes: Resolución de preguntas, tareas, casos clínicos o ejercicios	Realizar consulta clínica simulada (paciente virtual (VP) con polimialgia reumática); Hacer preguntas al VP; Obtener historial médico del VP; Formular diagnóstico; Sugerir plan de manejo	Manejo de casos clínicos simulados (presentados por IA como paciente estandarizado); Recopilación de historial clínico (interrogatorio a IA); Solicitud e interpretación (limitada) de pruebas de laboratorio; Formulación	Realizar conversación de historia clínica con paciente simulado; Desarrollar diagnóstico diferencial/ sospecha diagnóstica	Estudios académicos o trabajos de curso; Consultas clínicas o relacionadas con pacientes	Realizar consulta con paciente virtual (PV); Obtener historial médico; Formular diagnóstico; Formular plan de manejo	Participación en modo de enseñanza de simulación basada en escenarios potenciado por Inteligencia Artificial (IA) para manejar casos de enfermedades cardiovasculares	No reportado	No reportado	Análisis de información preliminar del caso; Análisis de datos adicionales (examen físico, imagenología, pruebas); Uso de sección Question-Answer (QA) para explorar/obtener evidencia adicional	Encuentros con pacientes simulados; Práctica de toma de historia clínica	No reportado	Resolución de escenarios clínicos simulados sobre Sangrado Gastrointestinal (SG) (evaluación de riesgo y manejo de casos); Realización de consultas en lenguaje natural sobre guías médicas o el riesgo	Practicar habilidades de toma de historia clínica interactuando con pacientes virtuales	Elaboración de diagnóstico diferencial; Investigación de opciones de tratamiento	Toma de historia clínica con ChatGPT como paciente virtual (cubriendo enfermedades cardiológicas)	Explorar interactivamente viñetas de casos clínicos generadas dinámicamente (presentaciones clínicas, toma de historia, exámenes físicos, estrategias diagnósticas, diagnósticos diferenciales, protocolos de tratamiento); Consultar a la Inteligencia Artificial (IA) para mejorar la comprensión;	Realización de previsualizaciones de temas de clase; Resolución de problemas durante la clase; Consolidación y revisión de material posterior a la clase; Responder preguntas relacionadas al contenido

		de conclusiones diagnósticas; Formulación de planes de tratamiento; Proporcionar recomendaciones al paciente (IA)										predicho por el modelo.				Recibir orientación inicial de la Inteligencia Artificial (IA) para formar evaluaciones	
Estudiantes: Resolución de cuestionarios de evaluación	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Planificación y gestión de tiempo	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Síntesis de anotaciones o apuntes de clase	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Lectura y/o resumen de artículos científicos o médicos	No reportado	No reportado	Síntesis o consolidación de material para revisión posterior a la clase
Estudiantes: Generación de contenido académico (presentaciones,	No reportado	No reportado	No reportado	Investigación	Generación de preguntas para el paciente virtual (PV)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Obtención de respuestas a preguntas sobre el manej	No reportado	Redacción de manuscritos o resúmenes (abstracts); Diseño o creación de	No reportado	No reportado	Generación de contenido/resúmenes para previsualizaciones previas

resúmenes, imágenes, etc.)												o clínico del Sangrado Gastrointestinal (SG); Recuperación de recomendaciones basadas en evidencia y guías clínicas.		diapositivas para clases; Creación de materiales de enseñanza; Creación de materiales educativos para pacientes			a la clase
Estudiantes: Autoevaluación o simulaciones de evaluaciones	No reportado	Participación en evaluación simulada (mediante interacción con IA como paciente estandarizado)	Recibir retroalimentación autogenerado sobre el desempeño (grupo intervención); Participar en simulación de entrevista médica	No reportado	Participación en simulación de consulta con paciente virtual (PV)	Participación en simulaciones de escenarios clínicos como preparación para la evaluación	No reportado	No reportado	Reflexión sobre el proceso de razonamiento propio; Comparación del análisis propio con resultados/pronóstico del caso	Práctica deliberada de habilidades de comunicación clínica; Aprendizaje autorregulado (SRL) a través de encuentros interactivos con pacientes; Recepción de evaluación/retroalimentación automatizada	No reportado	No reportado	Simulación de toma de historia clínica (como autoevaluación implícita)	No reportado	No reportado	Exploración interactiva de viñetas de casos clínicos generadas dinámicamente	No reportado
Estudiantes: Estudio autodirigido	No reportado	No reportado	No reportado	Recuperación de información médica general	No reportado	No reportado	No reportado	Uso general como herramienta de aprendizaje (ej.	Selección de casos específicos según estado de aprendizaje o	No reportado	No reportado	Aprendizaje y práctica de la toma de decisiones	Práctica autodirigida de toma de historia clínica (flexible en tiempo y lugar)	Autoaprendizaje o educación	Práctica de toma de historia clínica con ChatGPT como	Consultar a la Inteligencia Artificial (IA) para mejorar la comprensión;	Exploración de temas y profundización de conocimientos dentro de las

								Chat GPT)	interés; Exploración de casos clínicos mediante la plataforma learner-centered			clínicas en un entorno simulado y controlado, utilizando GutGPT como herramienta de apoyo informativo.			paciente virtual	Mejora percibida de las habilidades de autoaprendizaje	actividades estructuradas (previa a la clase, durante la clase, posterior a la clase)
Estudiantes: Otros (especificar)	Diálogo conversacional	No reportado	No reportado	No reportado	Practicar la expresión de empatía; Colaboración y discusión con compañeros durante la interacción con el paciente virtual (PV)	Trabajo en equipo y comunicación durante la simulación	No reportado	Realización de investigaciones; Exploración de alternativas; Obtención de segunda opinión	Selección de casos por criterios (dificultad, relevancia); Organización del análisis usando mind map (plantilla Facts, Ideas, Learning Issues, Action Plans (FILA)); Anotación de imágenes diagnósticas; Registro y actualización de hipótesi	Refinar el enfoque clínico basado en la retroalimentación	No reportado	Interacción conversacional general para soporte a la decisión	Acceder a guía de usuario en video sobre el uso de la aplicación	Realización de revisiones de literatura; Realización de análisis de datos; Redacción de notas clínicas; Redacción de resúmenes de alta; Redacción de notas de progreso	No reportado	Acceso a viñetas clínicas en formato de texto y video	Realizar consultas durante la clase

									s en lista diagnóstica; Exportación de notas/resúmenes para revisión posterior								
Actividad de los docentes	Desarrollo de un caso de paciente virtual (VP) sobre polimialgia reumática; Implementación del caso en plataforma robótica social (Furhat) con modelo de lenguaje grande (LLM); Implementación del caso en plataforma basada en computadora (simulador de casos	No reportado	Diseño e implementación de simulación de paciente con LLM; Diseño e implementación de retroalimentación automático con LLM	No reportado	Diseño y desarrollo de casos de pacientes virtuales (PV); Integración de Modelo de Lenguaje a Gran Escala (LLM) en plataforma robótica social; Facilitación de seminarios de seguimiento post-PV	Implementación y facilitación del modo de enseñanza de simulación basada en escenarios potenciado por Inteligencia Artificial (IA) para la enseñanza de enfermedades cardiovasculares	N/A	N/A	Selección de casos para actividades de Aprendizaje Basado en Problemas (ABP); Guiar discusiones (implícito); Evaluar reportes de análisis de caso; Evaluar dificultad de casos; Facilitar reflexión del estudiante (implícito)	No reportado	Revisión de grabaciones de video de Exámenes Clínicos Objetivos Estructurados (ECOE); Calificación de estudiantes basada en rúbricas específicas (ítem de resumen de historia médica)	Diseño y facilitación de estudio de simulación clínica; Provisión de herramientas (GutGPT, dashboard interactivo, Historia Clínica Electrónica (HCE) simulada); Administración de encuestas y evaluaciones; Aleatorización de participantes;	Creación y gestión de casos de pacientes virtuales en la aplicación chatbot (incluyendo definición de prompts)	N/A	N/A	Preparación de planes de enseñanza; Configuración de ChatGPT; Orientación sobre ChatGPT a estudiantes; Revisión del trabajo estudiantil con ChatGPT; Proporcionar retroalimentación sobre el trabajo con ChatGPT; Ofrecer retroalimentación tras práctica clínica; Responder preguntas de estudiantes	Diseño e implementación de modelo de enseñanza integrando ChatGPT (previa a la clase, durante la clase, posterior a la clase); Facilitación de sesiones de aprendizaje asistidas por ChatGPT

	interactivo virtual (VIC))											Facilitación de sesión informativa (debriefing)					
Actividad de los estudiantes	Interactuar individualmente con un caso de paciente virtual (VP) de polimialgia reumática; Utilizar plataforma robótica social con modelo de lenguaje grande (LLM) para la interacción; Utilizar plataforma convencional basada en computadora (simulador de casos interactivo virtual (VIC)) para la interac	Simulación de consulta médica con paciente estandarizado virtual (ChatGPT-4o); Manejo de casos clínicos (Hipertensión; Brucelosis); Interrogatorio verbal para historia clínica; Solicitud e interpretación de exámenes complementarios; Diagnóstico y plan de	Conversación de historia clínica con paciente simulado por LLM; (Grupo de retroalimentación) Recepción de retroalimentación estructurada generado por LLM	Estudios académicos o trabajos de curso; Consultas clínicas o relacionadas con pacientes; Recuperación de información médica general; Investigación	Resolución de casos de pacientes virtuales (PV) en plataforma robótica social (impulsada por Modelo de Lenguaje a Gran Escala (LLM)); Resolución de casos de PV en plataforma basada en computadora; Trabajo en parejas o grupos pequeños; Obten	Participación activa en simulaciones clínicas de enfermedades cardiovasculares utilizando la plataforma potenciada por Inteligencia Artificial (IA)	Responder Preguntas de Opción Múltiple (MCQs)	N/A	Análisis individual de casos clínicos basado en problemas (Aprendizaje Basado en Problemas - ABP); Práctica de razonamiento clínico y diagnóstico diferencial; Estudio autodirigido mediante selección de casos; Autoevaluación y reflexión sobre el propio proceso diagnóstico	Encuentro con paciente simulado (Toma de historia clínica para caso de anemia por deficiencia de hierro)	Participación en Exámenes Clínicos Objetivos Estructurados Comprensivos (ECOEC) / Exámenes Clínicos Objetivos Estructurados (ECOE) videograbados; Encuentros médicos simulados con paciente estandarizado (PE); Realización de tarea	Participación en escenarios de simulación clínica de alta fidelidad (con maniquí e Historia Clínica Electrónica (HCE) simulada)como miembros de equipo (estudiantes de medicina); Uso de la herramienta GutGPT (que incluye un Modelo Grand	Realizar entrevistas de historia clínica simuladas con pacientes virtuales (chatbot) mediante una aplicación móvil/web; Participar en simposio online y discusión grupal postinteracción	Actividades clínicas (ej., elaboración de diagnóstico diferencial, investigación de tratamientos, creación de material educativo para pacientes, redacción de notas clínicas) y no clínicas (ej., autoaprendizaje, redacción científica, resumen de artículos, creación de materiales didácticos/presentaciones, revisiones de literatura, análisis de datos) durante la residenci	Realizar una toma de historia clínica escrita con ChatGPT 3.5 actuando como paciente virtual, utilizando un prompt predefinido sobre enfermedades cardiológicas (endocarditis o insuficiencia cardíaca)	Exploración interactiva de viñetas de casos clínicos (texto y video); Consulta a la Inteligencia Artificial (IA) para mejorar comprensión; Recibir orientación de Inteligencia Artificial (IA) para evaluaciones; Práctica clínica supervisada (toma de historia, examen físico)	Realización de previsualizaciones previas a la clase con ChatGPT; Resolución de problemas/consultas en clase con ChatGPT; Consolidación/revisión posterior a la clase con ChatGPT

	ción; Practicar habilidades de razonamiento clínico (CR); Obtener información; Hacer un diagnóstico; Sugerir un plan de manejo	tratamiento verbalizado			ción de historial médico del PV; Discusión de diagnósticos diferenciales; Propuesta de planes de manejo para el PV; Participación en seminarios de seguimiento						s clínicas evaluadas (incluyendo resumir la historia médica del paciente)	e de Lenguaje (LLM) y un dashboard interactivo) para la evaluación de riesgo y manejo de casos de Sangrado Gastrointestinal (SG); Toma de decisión sobre disposición del paciente (alta, observación, ingreso); Respuesta a encuestas pre y post intervención y evaluaciones de contenido.		a de medicina interna.			

Estudiantes: Resultados en actividades específicas (teóricas, prácticas, teórico-prácticas)	No reportado	No reportado	Preintervención (Escenario 1): Grupo de retroalimentación M = 3.28 (SD = 0.09) vs Grupo de control M = 3.21 (SD = 0.08); Posintervención (Escenario 4): Grupo de Retroalimentación M = 3.60 (SE = 0.13) vs Grupo de control M = 3.02 (SE = 0.12), F(1,20) = 4.44, p = 0.049, partial eta2 = 0.198	No reportado	No reportado	Examen Teórico: Exposición 89.41 ± 4.11 vs Control 84.63 ± 4.39, p < 0.001; Análisis de Caso: Exposición 90.59 ± 3.58 vs Control 86.13 ± 3.79, p < 0.001; Operación de Habilidades: Exposición 92.41 ± 3.41 vs Control 87.50 ± 3.88, p < 0.001	ChatGPT-4: 81.7 % (DE = 13.6); Estudiantes: 55.8 % (DE = 14.9); p < 0.001, d de Cohen = 1.71	No reportado	No reportado	18.8 ± 3.2	No reportado	Grupo GutGPT (Intervención): Presimulación ~57% ± ~22% correctas, Postsimulación ~63% ± ~24% correctas; Grupo Dashboard + Búsqueda en Internet (Control): Presimulación ~49 ± ~27% correctas, Postsimulación ~57% ± ~26% correctas.	No reportado	No reportado	No reportado	Examen teórico: Grupo ChatGPT 92.21 ± 2.37 vs Grupo Tradicional 92.38 ± 2.68, t = 0.295, p = 0.768	Conocimiento Teórico: Grupo Estudio 87.59 ± 5.18 vs Control 82.56 ± 5.58, p < 0.05; Habilidad Análisis de Casos: Grupo Estudio 89.16 ± 4.18 vs Control 80.88 ± 5.21, p < 0.05
Estudiantes: Resultados en evaluaciones de ciencias biomédicas básicas	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Ciencias Básicas: ChatGPT-4: 84.5 % (DE	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado

							= 12.1); Estudiantes: 58.2 % (DE = 15.5); p < 0.001										
Estudiantes: Resultados en evaluaciones de ciencias clínicas	No reportado	No reportado	No reportado	No reportado	No reportado	Análisis de Caso: Exposición 90.59 ± 3.58 vs Control 86.13 ± 3.79, p < 0.001; Operación de Habilidades: Exposición 92.41 ± 3.41 vs Control 87.50 ± 3.88, p < 0.001	Ciencias Clínicas: ChatGPT-4: 79.8 % (DE = 14.5); Estudiantes: 54.1 % (DE = 14.2); p < 0.001	No reportado	No reportado	No reportado	No reportado	Grupo GutGPT (Intervención): Presimulación ~57% ± ~22% correctas, Postsimulación ~63% ± ~24% correctas; Grupo Dashboard + Búsqueda en Internet (Control): Presimulación ~49 ± ~27% correctas, Postsimulación ~57% ±	No reportado	No reportado	No reportado	Examen teórico: Grupo ChatGPT 92.21 ± 2.37 vs Grupo Tradicional 92.38 ± 2.68, t = 0.295, p = 0.768	No reportado

												~26% correctas.					
Estudiantes: Resultados en evaluaciones integradoras básico-clínicas	No reportado	No reportado	Preintervención (Escenario 1): Grupo de Retroalimentación M = 3.28 (SD = 0.09) vs Grupo de control M = 3.21 (SD = 0.08); Posintervención (Escenario 4): Grupo de Retroalimentación M = 3.60 (SE = 0.13) vs Grupo de control M = 3.02 (SE = 0.12), F(1,20) = 4.44, p = 0.049, partial eta2 = 0.198	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Resultados en evaluaciones estandarizadas (certificación, graduación)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Resultados en evaluaciones nacionales o internacionales	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado

Estudiantes: Resultados en evaluaciones para especialización	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Otro indicador objetivo del desempeño (especificar)	No reportado	PS-O: Media 8.43 ± 1.25; CR-O: Media 8.33 ± 1.35; CM-O: Media 8.52 ± 1.33	Número de palabras en Escenario 1 (Preintervención): Grupo de Retroalimentación M = 86.89 (SD = 20.02) vs Grupo de control M = 85.09 (SD = 16.67); t(18) = 0.21, p = 0.829	No reportado	No reportado	Puntuación de Evaluación General: Exposición 90.81 ± 2.89 vs Control 86.08 ± 3.16, p < 0.001	Desempeño por Nivel de Taxonomía de Bloom (Media %): Recordar: CGPT4 88.9 vs Est 60.1, p < 0.001; Comprender: CGPT4 85.3 vs Est 56.8, p < 0.001; Aplicar: CGPT4 78.1 vs Est 53.5, p < 0.001; Analizar:	No reportado	No reportado	Turnos de conversación: Media 23.8 ± 9.9; Palabras del estudiante: Media 310.1 ± 155.6; Palabras de la IA: Media 483.4 ± 217.2; Proporción de palabras del estudiante: Media 0.4 ± 0.1	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Cumplimiento Preparación Previa a la clase (%): Grupo Estudio 93.75 vs Control 81.25, p < 0.05; Cumplimiento Revisión Posterior a la clase (%): Grupo Estudio 84.38 vs Control 71.88, p < 0.05

							CGP T4 75.0 vs Est 51.2, p < 0.001 ; Evaluar: CGP T4 70.5 vs Est 48.0, p = 0.002 ; Crear : CGP T4 68.2 vs Est 45.5, p = 0.01										
Docentes: Frecuencia de uso (categórica)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Docentes: Periodicidad de uso (categórica)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	N/A	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Docentes: No. Interacciones (numérica)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	N/A	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Docentes: Tiempo de uso promedio	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	N/A	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado

(numérica)																	
Docentes: Otra medida de intensidad de uso (especificar)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	N/A	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Frecuencia de uso (categórica)	No reportado	No reportado	Participación en 4 escenarios de simulación durante la sesión de estudio	Uso frecuente reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Uso Clínico (n=40): Nunca 35%; Al menos una vez 30%; Mensual 23%; Semanal 10%; Diario 3%. Uso No Clínico (n=40): Nunca 8%; Al menos una vez 35%; Mensual 40%; Semanal 13%; Diario 0%. Uso Personal (n=152 total, n=68 usuarios): General Nunca 55%; Al menos una vez 20%; Mensual 16%; Semanal 7%; Diario 1%.	No reportado	No reportado	No reportado

Estudiantes: Periodicidad de uso (categórica)	No reportado	No reportado	Uso concentrado en una única sesión de estudio	Variable (Diaria; Varias veces/semana; Semanal; Mensual)	No reportado	No reportado	No reportado	No reportado	No reportado	Ocasional	No reportado	No reportado	No reportado	Uso Clínico (n=40): Mensual 23%; Semanal 10%; Diario 3%. Uso No Clínico (n=40): Mensual 40%; Semanal 13%; Diario 0%. Uso Personal (n=68 usuarios): Mensual 37% (25/68); Semanal 16% (11/68); Diario 1% (1/68).	N/A	No reportado	No reportado
Estudiantes: No. Interacciones (numérica)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Media 23.8 ± 9.9 turnos	No reportado	No reportado	No reportado	No reportado	Media 24.96 ± 9.41	No reportado	No reportado
Estudiantes: Tiempo de uso promedio (numérica)	No reportado	No reportado	6 minutos por escenario; 4 escenarios en total	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Otra medida de intensidad de uso (especificar)	No reportado	No reportado	Número de palabras en Escenario 1 (Preintervención): Grupo de Retroalimentación M = 86.89 (SD = 20.02) vs	Porcentaje de uso por tarea al menos una vez al mes	No reportado	No reportado	No reportado	No reportado	No reportado	Palabras del estudiante: Media 310.1 ± 155.6; Palabras de la IA: Media 483.4 ± 217.2	No reportado	No reportado	No reportado	Prevalencia de uso profesional: 26.3% (40/152); Prevalencia de uso personal: 44.7% (68/152); Prevalen	No reportado	No reportado	No reportado

			Grupo de control M = 85.09 (SD = 16.67)											cia de uso regular (>=Mensual) entre usuarios profesionales: Clínico 35% (14/40), No Clínico 52.5% (21/40); Prevalencia de uso regular entre usuarios personales: 54.4% (37/68).			
Estudiantes: Precisión diagnóstica (ej. % correcto, puntaje)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Juicio Clínico (Mini-Clinical Evaluation Exercise (Mini-CEX)): Grupo ChatGPT (1 (2.6 %) Bajo; 12 (30.8 %) Cumple; 26 (66.6 %) Supera) vs Grupo Tradicional (2 (5.3 %) Bajo; 22 (57.9 %) Cumple; 14 (36.8 %) Supera), p = 0.032	Grupo Estudio 89.16 ± 4.18 vs Control 80.88 ± 5.21, p < 0.05
Estudiantes: Tiempo	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado

hasta diagnóstico/resolución																	
Estudiantes: Proporción/Tipo de errores en razonamiento clínico	No reportado	No reportado	Mejora en subdominios del CRI-HTI: Creando contexto (p = 0.046); Asegurando información (p = 0.018). Sin mejora significativa en: Enfocando preguntas (p = 0.265)	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Estudiantes: Puntajes en pruebas específicas de razonamiento clínico (ej. ECOE)	No reportado	CR-O: Media 8.33 ± 1.35	Resultados del Clinical Reasoning Indicator - History Taking Inventory (CRI-HTI): Preintervención (Escenario 1): Grupo de Retroalimentación M = 3.28 (SD = 0.09) vs Grupo de control M = 3.21 (SD = 0.08); Posintervención (Escenario 4): Grupo de Retroalimentación M = 3.60 (SE = 0.13) vs Grupo de control M = 3.02 (SE = 0.12), F(1,20) =	No reportado	No reportado	Análisis de Caso: Exposición 90.59 ± 3.58 vs Control 86.13 ± 3.79, p < 0.001; Operación de Habilidades: Exposición 92.41 ± 3.41 vs Control 87.50 ± 3.88, p < 0.001; Puntuación de Evaluación General: Exposición	No reportado	No reportado	Consider the Patient Situation: Baseline 3.11 vs e-MedLearn 5.33 (p < 0.001); Collect Cues and Information: Baseline 4.00 vs e-MedLearn 5.67 (p < 0.01); Process Information: Baseline 3.22 vs e-MedLearn 4.78 (p	18.8 ± 3.2	No reportado	Grupo GutGPT (Intervención): Presimulación ~57% ± ~22% correctas, Postsimulación ~63% ± ~24% correctas; Grupo Dashboard + Búsqueda en Internet (Control): Presimulación	No reportado	No reportado	No reportado	No reportado	Comunicación Médico-Paciente: Grupo ChatGPT (10 (25.6 %) Cumple; 29 (74.4 %) Supera) vs Tradicional (20 (52.6 %) Cumple; 18 (47.4 %) Supera), p = 0.02; Profesionalismo: Grupo ChatGPT (12 (30.8 %) Cumple; 27 (69.2 %) Supera) vs Tradicional (22 (57.9 %)	No reportado

			4.44, p = 0.049, partial eta2 = 0.198			90.81 ± 2.89 vs Control 86.08 ± 3.16, p < 0.001			< 0.01); Identify Problems/Issues: Baseline 3.22 vs e-MedLearn 4.44 (p < 0.05); Establish Goals: Baseline 3.78 vs e-MedLearn 5.78 (p < 0.01); Take Action: Baseline 3.78 vs e-MedLearn 5.00 (p < 0.05); Evaluate Outcomes: Baseline 3.67 vs e-MedLearn 5.22 (p < 0.01); Reflect on the Process and New Learning: Baseline 3.00 vs e-MedLearn 5.22 (p < 0.001)			~49 ± ~27% correctas, Postsimulación ~57% ± ~26% correctas.				Cumple; 16 (42.1 %) Supera), p = 0.022; Capacidades Generales: Grupo ChatGPT (11 (28.2 %) Cumple; 28 (71.8 %) Supera) vs Tradicional (23 (60.5 %) Cumple; 15 (39.5 %) Supera), p = 0.006	

Estudiantes: Otra medida objetiva de razonamiento clínico (especificar)	Puntuación tema Autenticidad: Robot media 4.5 (DE 0.7) vs simulador de casos interactivo virtual (VIC) media 3.9 (DE 0.5); Odds ratio (OR) 2.9 (IC 95% 0.0-1.0); p = 0.04; Puntuación tema Efecto del aprendizaje: Robot media 4.4 (DE 0.6) vs VIC media 4.1 (DE 0.6); OR 3.7 (IC 95% 0.1-0.5); p = 0.01; Pregunta sobre toma de decisiones (Autenticidad): Robot	PS-O: Media 8.43 ± 1.25; CM-O: Media 8.52 ± 1.33	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Exhaustividad de la lista de síntomas por IA: 93.8%; Precisión de la lista de síntomas por IA: 95.6%	No reportado	No reportado	No reportado	No reportado	No reportado	Toma de Historia Médica (Mini-Clinical Evaluation Exercise (Mini-CEX)): Grupo ChatGPT (26 (66.7 %) Cumple; 13 (33.3 %) Supera) vs Tradicional (24 (63.2 %) Cumple; 14 (36.8 %) Supera), p = 0.814; Examen Físico (Mini-Clinical Evaluation Exercise (Mini-CEX)): Grupo ChatGPT (26 (66.7 %) Cumple; 13 (33.3 %) Supera) vs Tradicional (24 (63.2 %) Cumple; 14 (36.8 %) Supera), p = 0.814; Efectividad Organiza	No reportado

	media 4.6 (DE 0.6) vs VIC media 3.9 (DE 0.9); OR 3.2 (IC 95% 0.1-1.2); p = 0.03; Pregunta sobre preparación para paciente real (Efecto Aprendizaje): Robot media 4.4 (DE 0.6) vs VIC media 4.0 (DE 0.7); OR 4.2 (IC 95% 0.1-0.7); p = 0.009															cional (Mini-Clinical Evaluation Exercise (Mini-CEX)): Grupo ChatGPT (16 (41.0 %) Cumple; 23 (59.0 %) Supera) vs Tradicional (22 (57.9 %) Cumple; 16 (42.1 %) Supera), p = 0.174	
Estudiantes o docentes: Percepciones de usuarios sobre utilidad o confiabilidad	Utilidad: Muy alta, Confiabilidad: Media	Utilidad: Alta, Confiabilidad: Media	Utilidad: Alta, Confiabilidad: Alta	Utilidad: Media, Confiabilidad: Baja	Utilidad: Alta, Confiabilidad: Media	Utilidad: Alta, Confiabilidad: N/R (no evaluada directamente)	N/R (no es percepción usuario)	Utilidad: Alta, Confiabilidad: Media	Utilidad: Alta, Confiabilidad: N/R	Utilidad: Muy alta, Confiabilidad: Media	Utilidad: Alta (evaluadores), Confiabilidad: Media (requiere supervisión)	Utilidad: Alta (potencial), Confiabilidad: N/R (no evaluada clínicamente)	N/R (no es percepción directa)	Utilidad: Alta, Confiabilidad: Media	N/R (no es percepción usuario)	Utilidad: Alta, Confiabilidad: N/R (no detallada)	Utilidad: Alta, Confiabilidad: Media
Descripción del tipo y modalid	Entrenamiento uso robot	No reportado	No reportado	No reportado	N/R (docentes diseña	Entrenamiento platafo	No reportado	No reportado	No reportado	No reportado	N/R (estudio prueb	No reportado	No reportado	No reportado	No reportado	N/R (método enseñanza	Formación docentes uso

ad del entrenamiento o soporte recibido por los docentes	Furhat; Entrenamiento interfaz Large Language Model (LLM)				ron casos)	rma simulación IA; Guía diseño escenarios enseñanza					a concepto)					tradicional)	ChatGPT; Guías para integrar discusiones
Descripción del tipo y modalidad del entrenamiento o soporte recibido por los estudiantes	Introducción tecnología/objetivos; Instrucciones interacción con robot	Orientación inicial uso ChatGPT-4o; Instrucciones escenarios clínicos específicos	Breve introducción plataforma simulación; Instrucciones completar tareas asignadas	N/R (uso auto-dirigido)	Introducción uso robot social; Sesión informativa previa simulación	Introducción modo simulación IA; Soporte técnico durante uso	No reportado	N/R (encuesta sobre percepciones)	Instrucciones uso sistema PBL+CR; Demostración plataforma por instructores	Tutorial interactivo plataforma MedSimAI; Guías y ejemplos uso	N/R (no aplica a estudiantes)	Breve tutorial uso GutGPT; Escenarios prueba definidos	No reportado	N/R (uso auto reportado)	No reportado	Sesiones asistidas por ChatGPT; Comparado con enseñanza tradicional	Introducción uso ChatGPT; Discusiones grupales con IA
Grupo control: Descripción de herramientas tecnológicas y recursos estándar disponibles o utilizados	Plataforma computacional simulador de casos interactivo virtual (VIC); Interacción mediante opciones preescritas clickeables; Permite obtener historial médico, realizar examen físico (simulado),	No aplica	Interfaz ChatGPT Playground; Prompts de simulación de paciente	N/A	Plataforma Virtual Interactive Case simulator (VIC)	Métodos de enseñanza tradicionales; Aprendizaje basado en clases magistrales; Discusiones de casos	No reportado (Más allá del entorno universitario estándar)	N/A	Interfaz del sistema 'baseline' (e-MedLearn sin lista diagnóstica, Question-Answer (QA), mind map); Prompts de formato fijo para guiar análisis; Acceso a los mismos datos de caso que el grupo de intervención	N/A	Grabaciones de video; Listas de verificación estandarizadas; Escalas de calificación	Fase 1: Dashboard interactivo solamente; Fase 2: Dashboard interactivo y recursos en línea (búsquedas en internet y sitios de información clínica tradicionales).	N/A	N/A	N/A	Libro de texto estándar ("Pediatrics", 9ª ed., People's Medical Publishing House); Presentaciones multimedia (preparadas por instructores)	Métodos tradicionales de aprendizaje basados en multimedia

	solicitar y ver resultados de laboratorio								(para Etapas 2/3)								
Grupo control: Descripción de métodos o procedimientos estándar utilizados	Navegación semilinear del caso clínico; Selección de opciones predefinidas; Estructura de opción múltiple para diagnóstico y manejo	No aplica	Conversación de historia clínica con paciente simulado por LLM; Pausas entre simulaciones	N/A	Interacción semi-lineal con el paciente virtual (PV); Selección de preguntas preescritas; Propuesta de diagnóstico y plan de manejo al final; Trabajo en parejas/grupos; Seminarios de seguimiento con consultor	Aprendizaje basado en clases magistrales; Discusión de casos	Evaluación mediante Preguntas de Opción Múltiple (MCQs) en formato de test basado en computadora	N/A	Seguimiento del mismo pipeline de análisis de Aprendizaje Basado en Problemas (ABP) en 3 etapas; Análisis guiado por prompts fijos; Toma de notas individual para posterior redacción de reporte	N/A	Revisión de grabaciones de video de Exámenes Clínicos Objetivos Estructurados (ECOE); Calificación basada en rúbricas/escalas estandarizadas	Los participantes en los grupos control utilizaron el dashboard interactivo para evaluar el riesgo del paciente (Fase 1 y 2) y, adicionalmente en la Fase 2, pudieron usar motores de búsqueda de internet y sitios de información clínica tradicionales para tomar decisi	N/A	N/A	N/A	Enseñanza estándar junto a la cama del paciente; Introducción y demostración de casos por instructor; Participación del estudiante (entrevistas/evaluaciones con observación); Retroalimentación y discusión por instructor	Métodos tradicionales de aprendizaje basados en multimedia

												ones de manejo de casos de Sangrado Gastrointestinal (SG).					
Grupo control: Descripción del currículo, actividades formativas y nivel académico	Resolución de un caso específico de paciente virtual (VP) (polimialgia reumática); Contexto simulado de atención primaria; Estudiantes de medicina post-semestre 6 (con experiencia clínica básica en medicina interna)	No aplica	Estudiantes de medicina (mayoría 3er semestre), sin rotaciones clínicas previas; Actividad: 4 simulaciones de historia clínica (casos neurológicos/neuroquirúrgicos: hernia discal, ACV, meningitis, conmoción cerebral)	N/A	Casos de pacientes virtuales (PV) sobre diversas condiciones reumatológicas (AR, PMR, Gota, LES, GPA, DM, SpA); Actividad realizada durante el curso 'Medicina clínica 2' en rotaciones de reumatología	Módulo de enfermedades cardiovasculares; Clases magistrales; Discusiones de caso	Estudiantes de tercer año de medicina. Currículo cubierto por las Preguntas de Opción Múltiple (MCQs) incluye Ciencias Básicas (Anatomía, Fisiología, Bioquímica) y Ciencias Clínicas (Patología, Farmacología,	N/A	Tarea específica del estudio: selección de casos ortopédicos, análisis de caso proporcionado, registro/revisión, cuestionario post-tarea, reporte de análisis de caso (formato Clinical Reasoning Cycle (CRC)); Nivel académico: Médicos novatos (residentes/internos) con experiencia previa en	N/A	Calificación de Exámenes Clínicos Objetivos Estructurados Comprensivos (ECOEC) / Exámenes Clínicos Objetivos Estructurados (ECOE) de estudiantes de medicina; Evaluación del ítem 'resumen de la historia médica del pacie	Médicos de urgencias, médicos de medicina interna y estudiantes de medicina.	N/A	N/A	N/A	Rotación de internado en pediatría de dos semanas; Casos clínicos comunes en pediatría (idénticos al grupo de intervención: enfermedad de Kawasaki, gastroenteritis, cardiopatía congénita, síndrome nefrótico, bronconeumonía, convulsión febril); Examen teórico y evaluación Mini-Clinical Evaluation Exercise (Mini-CEX) al final	Mismo contenido curricular (rotación clínica de cirugía general); Mismo nivel académico (estudiantes de rotación clínica); Actividades formativas sin integración de ChatGPT; Mismas evaluaciones finales que grupo de intervención

							Salud Pública, Psiquiatría, Medicina Clínica).		Aprendizaje Basado en Problemas (ABP)		nte' de la rúbrica de habilidades de comunicación						
Grupo control: Exposición o uso auto reportado de LLMs fuera de la intervención específica	No reportado	No aplica	No reportado	N/A	No reportado	No reportado	No reportado	N/A	No reportado	N/A	No reportado	No reportado	N/A	N/A	N/A	No reportado	No reportado
Grupo control: Características relevantes del entorno de clase o aprendizaje	Entorno basado en computadora; Percepción de menor interactividad, menor exigencia y más pasividad que el robot; Considerado informativo y útil para aprendizaje estructurado y revisión	No aplica	Estudio presencial en facultad de medicina; Puestos individuales con distancia; Supervisión por 3 investigadores; Tiempos controlados	N/A	Entorno de rotación clínica en hospital universitario (Karolinska University Hospital); Trabajo en parejas/grupos pequeños	Entorno de aula estándar para clases magistrales y discusiones	Entorno universitario (Faculty of Medical Sciences, The University of the West Indies, Barbados). Evaluación realizada en un entorno controlado de	N/A	Tarea de análisis individual usando interfaz de sistema baseline; Contexto probable: remoto o laboratorio (no aula tradicional); Participantes reclutados de hospital local	N/A	No reportado	Escenarios de simulación clínica de alta fidelidad en un laboratorio de simulación que imita una sala de examen de hospital; Uso de un maniquí de simulación Laerd	N/A	N/A	N/A	Entorno hospitalario (implícito por enseñanza junto a la cama y pacientes); Parte del programa de la Universidad Sun Yat-sen; Ambiente de aprendizaje interactivo (fomentado por instructores en discusiones)	Mismo contexto hospitalario (Wenzhou People's Hospital); Mismo año académico (2022-2023); Entorno de aprendizaje clínico (rotación clínica)

							test basado en computadora.					al; Acceso a una versión de prueba de la Historia Clínica Electrónica (HCE) Epic con datos de pacientes simulados.					
Edad	Media 23.9 (SD ± 4.8) años	Media 24 (SD ± 1.03) años	Media 22.10 años	21-23 años: 37.2 % (n = 38); 24-26: 49.0 % (n = 50); 27-29: 11.8 % (n = 12); > 29: 2.0 % (n = 2)	Media 23.5 (SD ± 6.0) años	N/R o N/A	N/R o N/A	N/R o N/A	Media 27.5 años	N/R o N/A	N/R o N/A	N/R o N/A	N/R o N/A	Distribución por rangos: <25 años (1%); 26-30 años (68%); 31-35 años (24%); 36-40 años (7%); >40 años (0%)	N/R o N/A	N/R o N/A	N/R o N/A
Sexo / Género	Femenino:40 % (n = 6) Masculino: 60 % (n = 9)	Femenino:38.1 % (n = 8) Masculino: 61.9 % (n = 13)	Masculino: 33.3 % (n = 7) Femenino: 66.7 % (n = 14)	N/R o N/A	Masculino: 61 % (n = 14) Femenino:39 % (n = 9)	N/R o N/A	Masculino: 42.7 % (n = 41) Femenino :57.3 % (n = 55)	Masculino: 52.04 % (n = 217) Femenino:47.96 % (n = 200)	Masculino: 57.89 % (n = 11) Femenino:42.11 % (n = 8)	N/R o N/A	N/R o N/A	N/R o N/A	Masculino: 35.6% n = 16) Femenino:57.8 % n = 26) Otro/Prefiere no decirlo: 6.7% n = 3)	Masculino: 57% n = 86) Femenino:42% n = 64) Otro/Prefiere no decirlo: 2% n = 2)	N/R o N/A	Masculino: 54.5 % (n = 42) Femenino:45.5 % (n = 35)	N/R o N/A
Nivel académico del	Estudiante de medicina	Médico interno	Médico interno (estudiante de	Estudiante de medicina (pregrad	Estudiante de medici	Médico interno (estudiante	Estudiante de medi	Médico interno	Médico estudiante (posgra	Estudiante de medicina (pregrado)	Estudiante de medic	Estudiante de medici	Médico interno (estudiante de	Médico estudiante (posgrad	Estudiante de medici	Médico interno (estudiante de	Estudiante de medicina

participante	(pregrado), anterior al último año o sin año especificado	(estudiante de medicina de último año)	medicina de último año)	o), anterior al último año o sin año especificado	na (pregrado), anterior al último año o sin año especificado	de medicina de último año)	cina (pregrado), anterior al último año o sin año especificado	(estudiante de medicina de último año); Estudiante de medicina (pregrado), anterior al último año o sin año especificado	do), de especialidad médico-quirúrgica.	, anterior al último año o sin año especificado	ina (pregrado), anterior al último año o sin año especificado	na (pregrado), anterior al último año o sin año especificado; Otro médico asistencial, sin otra especificación	medicina de último año); Estudiante de medicina (pregrado), anterior al último año o sin año especificado	o), de especialidad médico-quirúrgica.	na (pregrado), anterior al último año o sin año especificado	medicina de último año)	(pregrado), anterior al último año o sin año especificado
Experiencia previa con LLMs	No reportado	No reportado	No reportado	Reportado como Comprensión/Exposición	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Distribución de Familiaridad (n=146): Nada familiar (~1%); Ligeramente familiar (~23%); Moderadamente familiar (~35%); Muy familiar (~30%); Extremadamente familiar (~11%)	Baja o discreta o la usa esporádicamente o para tareas elementales	No reportado	No reportado
Área de conocimiento / Especialidad principal	Pregrado en medicina (Rotación en Medicina Interna)	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina (Rotación en Reumatología)	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina	Especialización en Ortopedia	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina; Especialización en Medicina de Emerg	Pregrado en medicina	Especialización en Medicina Interna	Pregrado en medicina	Pregrado en medicina (Rotación de pediatría)	Pregrado en medicina (Rotación de cirugía general)

												encias; Especialización en Medicina Interna					
Rendimiento académico previo	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
Nombre del modelo LLM	GPT-3.5-turbo	ChatGPT-4o	ChatGPT 3.5	ChatGPT; Bard; Bing Chat; Claude (Mencionados como ejemplos)	gpt-3.5-turbo	No reportado	Chat GPT-4	Chat GPT	GPT-3.5-turbo-16k	GPT-4	GPT-4; GPT-4o; GPT-3.5; Claude-Opus; Claude-Sonnet-3-5; Llama-3-70B; Llama-3-8B; Mixtral 8x7B; Starling-7B-beta	GPT-3.5 Turbo 16k (OpenAI)	gpt-35-turbo	GPT-3/4	ChatGPT 3.5	ChatGPT	ChatGPT
Plataforma o Modo de acceso	Robot social Furhat; GPT-3.5 (vía API)	ChatGPT-4o; Acceso mediante interfaz web	Plataforma simulación web; Large Language Model (LLM) integrado	Large Language Models (LLMs) (ej. ChatGPT); Acceso público vía web	Robot social (tipo Furhat); Large Language Model (LLM) integrado	Plataforma simulación empoderada por IA; Basada en escenarios clínicos	Chat GPT-4; Interfaz web OpenAI	Inteligencia Artificial (IA) (general); Chat GPT (específico)	Sistema PBL+CR con Large Language Model (LLM); Acceso en línea	MedSimAI; Plataforma web	Large Language Models (LLMs) (GPT-3); API para análisis texto	GutGPT (prototipo); Aplicación web	ChatGPT; Interfaz web OpenAI	Large Language Models (LLMs) (ChatGPT); Acceso web y móvil	ChatGPT (GPT-3.5); Usado como paciente virtual	ChatGPT; Acceso web	ChatGPT; Plataforma WeChat para discusión
Configuración de	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Por defecto	No reportado	No reportado	No reportado	No reportado	No reportado	Temperatura = 0.15;	No reportado	temperature = 0.2;	No reportado	No reportado

parámetros clave													Top_p = 0.95		temperature = 0.8		
Costo asociado al uso	Costos API cubiertos proyecto	Costo API ChatGPT-4o (institucional)	N/R (desarrollo interno)	Gratuito (versiones básicas); Suscripción (versiones avanzadas)	Costo robot y API; Cubierto por investigación	N/R (plataforma institucional)	Gratuito o suscripción personal	Varía (gratuito a pago)	N/R (desarrollo prototipo)	N/R (desarrollo para investigación)	Costo API Large Language Model (LLM)	N/R (prototipo investigación)	Gratuito (básico); Pago (avanzado)	Varía (gratuito/ pago personal)	Costo API ChatGPT	N/R (asumido institucional/personal)	N/R (uso herramientas disponibles)

ANEXOS

1 Prompts

Prompt 1. Asistente y tutor en el desarrollo de una Revisión Sistemática/Metaanálisis para tesis de grado.

A. *Rol Primario: Actúa como mi Asistente Académico y Tutor experto para el desarrollo de mi tesis de Maestría o Doctorado. Tu función es guiar mi proceso de investigación y escritura, fomentar mi pensamiento crítico, asegurar el rigor metodológico y conceptual, y ayudarme a estructurar mis ideas y argumentos de manera lógica y coherente. A veces te pediré ayuda con la escritura (rol de asistente) y otras veces con la revisión de los avances (rol de tutor).*
B. *Contexto del Usuario y la Tesis:*

Dime esto al inicio de las interacciones: "Para poder asistirte eficazmente, necesito entender el contexto actual de tu trabajo. Por favor, inicia nuestra interacción respondiendo a una serie de preguntas:" y luego hazme, una por una, todas las siguientes preguntas (puedes abreviar o modificar un poco el contenido de cada una, manteniendo su esencia):

1. *Nivel Académico de la Tesis: ¿Para qué tipo de programa académico estás desarrollando la tesis (ej. Maestría en Educación Médica, Doctorado en Ciencias, etc.)? (Espera mi respuesta a estas preguntas antes de continuar).*
2. *Etapa Actual del Proceso: ¿En qué fase del proyecto te encuentras actualmente? (ej. definición del tema/pregunta, revisión de literatura, diseño metodológico, recolección de datos, análisis de datos, redacción de un capítulo específico [cuál], preparación de la defensa, respuesta a revisores, etc.) (Espera mi respuesta a estas preguntas antes de continuar).*
3. *Tema General de Tesis: ¿Podrías darme detalles sobre la temática general de la tesis que estás planeando o desarrollando? (Espera mi respuesta a estas preguntas antes de continuar).*
4. *Pregunta de Investigación: ¿Tienes ya definida una pregunta de investigación (principal y/o secundarias)? Si es así, ¿cuál es? Si no, ¿en qué punto estás de su definición? (Espera mi respuesta a estas preguntas antes de continuar).*
5. *Metodología Anticipada: ¿Has definido o tienes ideas sobre la metodología de investigación que planeas usar (ej. cuantitativa, cualitativa, mixta, tipo de estudio específico)? (Espera mi respuesta a estas preguntas antes de continuar).*
6. *Universidad/Contexto (Opcional): ¿Deseas compartir información sobre la institución para la cual estás realizando el trabajo? (Esto puede ser útil si existen normativas o formatos específicos, pero es opcional) (Espera mi respuesta a estas preguntas antes de continuar).*

C. *Tarea Específica para Esta Interacción:*

Una vez que te haya proporcionado el contexto respondiendo a las preguntas anteriores, pídeme que te indique claramente la tarea o el desafío concreto para el que necesito tu ayuda en esta sesión.

Algunos ejemplos de posibles tareas (a ser proporcionadas por mí con tu solicitud):

- *"Ayúdame a refinar mi pregunta de investigación para asegurar que sea clara, factible y relevante, basándonos en el tema que te describí."*
- *"Necesito identificar posibles vacíos en la literatura sobre [subtema específico del tema general] para justificar mi estudio."*
- *"Discutamos la adecuación de [metodología específica] para abordar mi pregunta de investigación, considerando sus fortalezas y debilidades en el contexto de la educación médica de pregrado."*
- *"Revisa este borrador de mi introducción [adjuntar o pegar texto si es posible, o describir los puntos clave] y dame retroalimentación sobre su claridad, coherencia lógica y fuerza argumentativa."*
- *"Ayúdame a generar un esquema detallado para el capítulo de metodología, considerando el enfoque [mencionar enfoque discutido]."*
- *"Exploremos diferentes enfoques para analizar los datos [cualitativos/cuantitativos] que planeo recolectar sobre [aspecto específico]."*
- *"Ayúdame a identificar potenciales sesgos en mi diseño de estudio [describir diseño] y cómo mitigarlos."*
- *"Necesito encontrar fuentes académicas rigurosas y recientes sobre [concepto clave relacionado con mi tema]."*
- *"Revisemos la estructura lógica y la fluidez de la argumentación en esta sección [especificar sección]."*
- *"Ayúdame a reformular esta frase/párrafo [proporcionar texto] para que sea más claro/conciso/académico."*

D. *Estilo de Interacción de la IA con el usuario:*

1. *Rigor y Profundidad: Prioriza la profundidad analítica y el rigor académico, especialmente relevante en educación médica y ciencias de la salud. Evita respuestas superficiales, vagas o genéricas. Fundamenta tus sugerencias y críticas en principios metodológicos, pedagógicos, lógicos o evidencia conceptual.*
2. *Pensamiento Crítico: No te limites a aceptar mis premisas. Cuestiona mis supuestos, señala posibles debilidades en mis argumentos o enfoques (metodológicos, pedagógicos, etc.), y preséntame perspectivas alternativas o contraargumentos para fortalecer mi trabajo. Actúa como un revisor crítico constructivo.*
3. *Claridad Conceptual y Conexiones: Ayúdame a establecer conexiones lógicas claras entre diferentes partes de mi tesis (problema, pregunta, objetivos, marco*

teórico/conceptual, metodología, resultados, discusión) y con los principios de la educación médica o las ciencias básicas/clínicas relevantes. Fomenta la precisión terminológica.

4. *Evitar Redundancia: Asegúrate de que tus respuestas sean concisas y directas al punto, evitando la repetición innecesaria de ideas. Esto no aplica solo para ideas en una misma frase o párrafo, sino en una misma sección y en el texto completo.*
5. *Enfoque por Roles: Cuando veas que te estoy pidiendo tutoría, hazme preguntas que me guíen a descubrir las respuestas o a refinar mis propias ideas, en lugar de simplemente darme la solución. Ayúdame a construir mi propio entendimiento. Si es útil, emplea analogías (técnicas o cotidianas, relevantes para medicina o educación) para clarificar conceptos complejos. Si te estoy pidiendo asistencia, atiende de forma más directa a la solicitud específica dada.*
6. *Foco en la Tarea: Una vez establecido el contexto y definida la tarea específica, céntrate estrictamente en ella.*
7. *Tono: Mantén un tono profesional y académico, pero directo y sin formalidades innecesarias.*

2 Tablas suplementarias

Tabla suplementaria 1. Herramienta para la selección de artículos para la inclusión final en la revisión sistemática

Criterio / Aspecto	Descripción / Pregunta guía	Código / Etiqueta de Respuesta / Valoración	Comentarios / Justificación	Consecuencia / Instrucciones
(DA) Título del artículo	Título completo.	Texto libre		
(DA) Autor/es	Autores principales y “et al.”	Texto libre		
(DA) Año de publicación	Extraer año de publicación.	Numérico		
(DA) ID del artículo	DOI, PMID u otro identificador único.	Texto libre		
(DA) Link al artículo completo	URL al texto completo.	Texto libre		
(CIG) 1. Publicación 2023-2025	¿Publicado entre enero 2023 y marzo 2025?	0=No; 1=Sí		Si=0 → Marcar para exclusión
(CIG) 2. Acceso e idioma	¿Texto completo accesible en inglés o español (libre o institucional)?	0=No; 1=Sí		Si=0 → Marcar para exclusión
(CIE) 3. Enfoque Temático	¿Foco en LLMs en educación médica (no otras profesiones)?	0=No; 1=Sí		Si=0 → Marcar para exclusión
(CIE) 4. Relación con Outcomes PICO	¿Evalúa ≥ DOS de estos outcomes: (a) Razonamiento clínico o Desempeño, (b) Integración Básico-Clínica, (c) Transferencia a la Práctica, (d) Percepciones?	0=<2; 1=≥2	Listar outcomes evaluados (a–d)	Si=0 → Marcar para exclusión
(CEE) 5. Tipo de publicación	¿Es revisión, editorial, comentario, carta, etc., sin datos primarios?	0=Sí; 1=No		Si=0 → Marcar para exclusión
(CEE) 6. Ámbito No Educativo Médico	¿Foco en práctica profesional sin contexto educativo?	0=Sí; 1=No		Si=0 → Marcar para exclusión
(CEE) 7. Evaluación exclusiva del LLM	¿Solo evalúa desempeño del LLM en alguna prueba, sin analizar impacto en estudiantes de medicina?	0=Sí; 1=No		Si=0 → Marcar para exclusión
(CEE) 8. Tecnología No-LLM	¿Foco en IA no-LLM u otra tecnología distinta?	0=Sí; 1=No		Si=0 → Marcar para exclusión
(ECP) 9. Diseño general	Identificar diseño metodológico (experimental, cuasiexperimental, observacional, etc.)	0 = Otro/No claro. 1 = Experimental aleatorizado / RCT. 2 = Cuasi experimental. 3 = Observacional: Cohortes. 4 = Observacional: Casos-controles. 5 = Observacional: Transversal. 6 = Observacional: Prevalencia/Descriptivo.		Si=0 → Considerar exclusión/discutir
(ECP) 10. Enfoque principal	Identificar enfoque del tipo de resultados (cuantitativo, cualitativo o mixto)	0 = No claro. 1 = Cuantitativo. 2 = Cualitativo. 3 = Mixto.		Si=0 → Considerar exclusión/discutir
(ECP) 11. Claridad metodológica	¿Metodología descrita con detalle y claridad suficiente para evaluar su rigor y riesgo de sesgos?	0=No; 1=Sí	Evaluar detalle y claridad	Si=0 → Considerar exclusión/discutir

(ECP) 12. Presentación resultados relevantes	¿Resultados de CIE 4 extraíbles y con tipo de dato claro?	0=No; 1=Sí	Evaluar claridad y nivel de detalle	Si=0 → Considerar exclusión/discutir
(ECP) 13. Riesgo de errores metodológicos y sesgos evidentes	Según los resultados de ECP 11 y ECP12, definir riesgo de errores metodológicos y sesgos, considerando: ¿Hay limitaciones evidentes que afecten confiabilidad/utilidad?	0 = Bajo riesgo 1 = Moderado riesgo 2 = Alto riesgo	Identificar limitaciones y su impacto	Si=2 → Justificar exclusión; Si=1 → Discutir
(EFC) 14. Evaluación Formal de Calidad	Aplicar herramienta según ECP9 y ECP10: RoB2, JBI específico o MMAT y según la evaluación de calidad, decidir sobre la inclusión.	0 = Excluir. 1 = Incluir. 2 = Dudoso.		Si=2 → Justificar dudas y discutir; probable exclusión.
15. Decisión Final	Incluir o Excluir según criterios anteriores.	Incluir/Excluir	N/A	N/A

Nomenclatura: DA: datos del artículo; CIG: criterios de inclusión generales; CIE: criterios de inclusión específicos; CEE: criterios de exclusión específicos; EPC: evaluación preliminar de calidad; EFC: Evaluación Formal de Calidad.

Tabla suplementaria 2. Motivos detallados de exclusión de estudios tras la lectura del texto completo

TÍTULO	ID ORIGINAL	MOTIVO DE EXCLUSIÓN (CRITERIOS INCLUSIÓN/EXCLUSIÓN)
Artificial Intelligence–Powered Training Database for Clinical Thinking: App Development Study	DOI10.2196/58426	Diseño observacional/descriptivo sin control robusto reportado; evaluación del CR basada en scores del sistema con validación poco clara; diseño débil, haciendo los datos poco confiables o útiles para la pregunta específica de la revisión sobre impacto de LLMs.
Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians	DOI10.1371/JOURNAL.PONE.0307383	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 13: alto riesgo
Using ChatGPT in the Development of Clinical Reasoning Cases: A Qualitative Study	DOI10.7759/CUREUS.61438	Criterio CIE 4: no cumple; Criterio ECP 12: no cumple; Criterio ECP 13: alto riesgo
Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability Framework for Safe and Effective Large Language Models in Medical Education: Narrative Review and Qualitative Study	DOI10.2196/51834	Criterio CIE 4: no cumple
Diagnosis in Bytes: Comparing the Diagnostic Accuracy of Google and ChatGPT 3.5 as an Educational Support Tool	DOI10.3390/IJERPH21050580	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 6: no cumple; Criterio CEE 7: no cumple; Criterio ECP 13: alto riesgo
ChatGPT as a Tool for Medical Education and Clinical Decision-Making on the Wards: Case Study	DOI10.2196/51346	Criterio CIE 4: no cumple
Using ChatGPT in Psychiatry to Design Script Concordance Tests in Undergraduate Medical Education: Mixed Methods Study	DOI10.2196/54067	Criterio CIE 4: no cumple; Criterio ECP 13: alto riesgo
ChatGPT's Response Consistency: A Study on Repeated Queries of Medical Examination Questions	DOI10.3390/EJIHPE14030043	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: cumple condición de exclusión; Criterio ECP 13: alto riesgo/baja utilidad
Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models	DOI10.2196/51391	Criterio CEE 7: no cumple
Exploring Generative Artificial Intelligence-Assisted Medical Education: Assessing Case-Based Learning for Medical Students	DOI10.7759/CUREUS.51961	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 9: No claro; Criterio ECP 12: No extraíbles; Criterio ECP 13: alto riesgo
CHATGPT APPLICATIONS IN MEDICAL, DENTAL, PHARMACY, AND PUBLIC HEALTH EDUCATION: A	DOI10.52225/NARRA.V3I1.103	Criterio ECP 11: no hay acceso; Criterio ECP 12: no hay acceso; Criterio ECP 13: no hay acceso

DESCRIPTIVE STUDY HIGHLIGHTING THE ADVANTAGES AND LIMITATIONS		
Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam	DOI10.4415/ANN_23_04_05	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple
Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4	DOI10.1186/S12909-023-04752-W	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 12: No extraíbles/Superficiales; Criterio ECP 13: alto riesgo
Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology	DOI10.3389/FONC.2023.1265024	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 12: No extraíbles/Superficiales; Criterio ECP 13: Alto
Artificial Intelligence (AI) in Radiology: A Deep Dive Into ChatGPT 4.0's Accuracy with the American Journal of Neuroradiology's (AJNR) "Case of the Month"	DOI10.7759/CUREUS.43958	Criterio CEE 7: no cumple; Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio ECP 12: no cumple
Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology	DOI10.7759/CUREUS.42972	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 13: alto riesgo
ChatGPT-4: Transforming Medical Education and Addressing Clinical Exposure Challenges in the Post-pandemic Era	DOI10.1007/S43465-023-00967-7	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 6: cumple condición exclusión; Criterio CEE 7: cumple condición exclusión; Criterio ECP 13: alto riesgo/baja utilidad
Sailing the Seven Seas: A Multinational Comparison of ChatGPT's Performance on Medical Licensing Examinations	DOI10.1007/S10439-023-03338-3	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 12: no cumple; Criterio ECP 13: alto riesgo
Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study	DOI10.2196/48002	Criterio CEE 7: no cumple; Criterio CIE 4: no cumple; Criterio ECP 12: no cumple; Criterio ECP 13: alto riesgo
Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry	DOI10.7759/CUREUS.37023	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: cumple criterio exclusión; Criterio ECP 13: alto riesgo/baja utilidad
Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models	DOI10.1371/JOURNAL.PDIG.0000198	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 9: no claro; Criterio ECP 13: alto riesgo
Adopting Artificial Intelligence Driven Technology in Medical Education	DOI10.1108/ITSE-12-2023-0240	Criterio CIG 2: no cumple (Texto completo inaccesible)
Facilitating nursing and health education by incorporating ChatGPT into learning designs	DOI10.30191/ETS.202401_27(1).TP02	Criterio CIE 3: no cumple; Criterio CEE 6: no cumple
ChatGPT as an innovative heutagogical tool in medical education	DOI10.1080/2331186X.2024.2332850	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 9: no claro; Criterio ECP 13: alto riesgo
ChatGPT as a Life Coach for Professional Identity Formation in Medical Education: A Self-Regulated Learning Perspective	DOI10.30191/ETS.202407_27(3).TP03	Criterio CIG 2: no cumple
Integrating Artificial Intelligence into Medical Education: Lessons Learned From a Belgian Initiative	ISSN1093-023X	Criterio CIE 4: Evalúa menos de DOS de los outcomes listados (a-d); Criterio CEE 8: Foco principal en IA no-LLM u otra tecnología distinta
Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis	DOI10.1111/ANS.18666	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple
Adapting ChatGPT for Color Blindness in Medical Education	DOI10.1007/S10439-024-03656-0	Criterio CIE 4: Evalúa menos de DOS de los outcomes listados (a-d); Criterio CEE 7: Sí (evalúa solo LLM); Criterio ECP 13: Alto (limitaciones severas que probablemente hacen los datos poco confiables o inútiles para la revisión)

Bias Sensitivity in Diagnostic Decision-Making: Comparing ChatGPT with Residents	DOI10.1007/S11606-024-09177-9	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 6: cumple condición exclusión; Criterio CEE 7: cumple condición exclusión; Criterio ECP 13: alto riesgo/baja utilidad para la revisión
Self-Evolving Multi-Agent Simulations for Realistic Clinical Interactions	ARXIV2503.22678	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 13: alto riesgo
Modeling Challenging Patient Interactions: LLMs for Medical Communication Training	ARXIV2503.22250V2	Criterio CIE 4: no cumple
Evaluating Large Language Models on the Spanish Medical Intern Resident (MIR) Examination 2024/2025: A Comparative Analysis of Clinical Reasoning and Knowledge Application	ARXIV2503.00025	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 13: alto riesgo/inútil
LLMs Can Simulate Standardized Patients via Agent Coevolution	ARXIV2412.11716	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 12: no extraíbles/superficiales; Criterio ECP 13: alto riesgo
AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow	ARXIV2409.18924	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 12: no cumple; Criterio ECP 13: alto riesgo
Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations	ARXIV2402.07023	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 6: cumple condición de exclusión; Criterio CEE 7: cumple condición de exclusión; Criterio ECP 13: alto riesgo/inutilidad para la revisión
Towards Conversational Diagnostic AI	ARXIV2401.05654	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 7: no cumple
Evaluating Large Language Models in Ophthalmology	ARXIV2311.04933	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 6: cumple exclusión; Criterio CEE 7: cumple exclusión; Criterio ECP 12: no extraíbles/irrelevantes; Criterio ECP 13: alto riesgo/inútil para la revisión
Performance of ChatGPT on USMLE: Unlocking the Potential of Large Language Models for AI-Assisted Medical Education	ARXIV2307.00112	Criterio CEE 7: no cumple
Capabilities of GPT-4 on Medical Challenge Problems	ARXIV2303.13375	Criterio CIE 3: no cumple; Criterio CIE 4: no cumple; Criterio CEE 6: no cumple; Criterio CEE 7: no cumple; Criterio ECP 9: no claro; Criterio ECP 12: no cumple; Criterio ECP 13: alto riesgo
Application of Large Language Models in Medical Training Evaluation-Using ChatGPT as a Standardized Patient: Multimetric Assessment	DOI10.2196/59435	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 12: no cumple; Criterio ECP 13: alto riesgo
Developing Medical Education Curriculum Reform Strategies to Address the Impact of Generative AI: Qualitative Study	DOI10.2196/53466	Criterio CIE 4: Evalúa menos de DOS de los outcomes listados (a-d)
Employing Large Language Models for Surgical Education: An In-depth Analysis of ChatGPT-4	DOI10.5812/JME-137753	Criterio CIE 4: no cumple; Criterio CEE 7: no cumple; Criterio ECP 13: alto riesgo

Tabla suplementaria 3. Caracterización detallada de las poblaciones de los estudios incluidos

No. Original	1	3	32	36	45	59	81	177	347	350	414	500	943	959	1127	1138	1155
Autor(es)	Borg et al	Öncü et al	Brügge et al	Xu et al	Borg et al	Zheng et al	Bharatha et al	Ahmad et al	Xu et al	Hicke et al	Shakur et al	Chan et al	Chan et al	Fried et al	Aster et al	Ba et al	Huang et al
Tamaño	15	21	21	102	23	66	96	417	18	104	2027	55	45	152	35	77	64

muestral total																	
Muestra grupo exposición / intervención	15	21	10	N/A	N/A	34	N/A	N/A	9	104	N/A	Fase 1: 24; Fase 2: 28	45	N/A	35	39	32
Muestra grupo control	15	N/A	11	N/A	N/A	32	96	N/A	9	N/A	N/A	Fase 1: 31; Fase 2: 23	N/A	N/A	N/A	38	32
Diseño metodológico	Cuasi experimental	Cualitativo	Cuasi experimental	Cualitativo	Comparativo cualitativo	Cuasi experimental	Cualitativo	Cualitativo	Descriptivo	Cuasi experimental	Comparativo cualitativo	Descriptivo	Cualitativo	Cualitativo	Cuasi experimental	Descriptivo	Descriptivo
Enfoque del estudio	Mixto	Mixto	Cuantitativo	Cuantitativo	Cualitativo	Mixto	Cuantitativo	Mixto	Mixto	Mixto	Mixto	Mixto	Mixto	Cuantitativo	Mixto	Mixto	Cuantitativo
País	Suecia	Turquía	Alemania	Estados Unidos	Suecia	China	Barbados; Bangladesh	India	China	Estados Unidos	Estados Unidos	Estados Unidos	Hong Kong; Singapur; Australia	Estados Unidos	Alemania	China	China
Edad	Media 23.9 (SD ± 4.8) años	Media 24 (SD ± 1.03) años	Media 22.10 años	21-23 años: 37.2 % (n = 38); 24-26: 49.0 % (n = 50); 27-29: 11.8 % (n = 12); > 29: 2.0 % (n = 2)	Media 23.5 (SD ± 6.0) años	N/R o N/A	N/R o N/A	N/R o N/A	Media 27.5 años	N/R o N/A	N/R o N/A	N/R o N/A	N/R o N/A	Distribución por rangos: <25 años (1%); 26-30 años (68%); 31-35 años (24%); 36-40 años (7%); >40 años (0%)	N/R o N/A	N/R o N/A	N/R o N/A
Género	Femenino: 40 % (n = 6) ; Masculino: 60 % (n = 9)	Femenino: 38.1 % (n = 8) ; Masculino: 61.9 % (n = 13)	Masculino: 33.3 % (n = 7) ; Femenino: 66.7 % (n = 14)	N/R o N/A	Masculino: 61 % (n = 14) ; Femenino: 39 % (n = 9)	N/R o N/A	Masculino: 42.7 % (n = 41) ; Femenino: 57.3 % (n = 55)	Masculino: 52.04 % (n = 217) ; Femenino: 47.96 % (n = 200)	Masculino: 57.89 % (n = 11) ; Femenino: 42.11 % (n = 8)	N/R o N/A	N/R o N/A	N/R o N/A	Masculino: 35.6 % (n = 16) ; Femenino: 57.8 % (n = 26) ; Otro/Prefiere no	Masculino: 57 % (n = 86) ; Femenino: 42 % (n = 64) ; Otro/Prefiere no decirlo:	N/R o N/A	Masculino: 54.5 % (n = 42) ; Femenino: 45.5 % (n = 35)	N/R o N/A

													decirlo: 6.7 % (n = 3)	2 % (n = 2)			
Nivel académico de los participantes	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.	Médico interno (estudiante de medicina de último año).	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.	Médico general o sin posgrado, docente, sin formación pedagógica.	Médico general o sin posgrado, docente, con formación pedagógica.	Estudiante de medicina (pregrado), anterior al último año o sin año especificado; Médico interno (estudiante de medicina de último año).	Médico estudiante (posgrado), de especialidad médico-quirúrgica.	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.	Estudiante de medicina (pregrado), anterior al último año o sin año especificado; Otro médico asistencial, sin otra especificación.	Estudiante de medicina (pregrado), anterior al último año o sin año especificado; Médico interno (estudiante de medicina de último año).	Médico interno (estudiante de medicina de último año).	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.	Médico interno (estudiante de medicina de último año).	Estudiante de medicina (pregrado), anterior al último año o sin año especificado.
Área conocimiento	Pregrado en medicina (Rotación en Medicina Interna)	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina (Rotación en Reumatología)	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina	Especialización en Ortopedia	Pregrado en medicina	Pregrado en medicina	Pregrado en medicina; Especialización en Medicina de Emergencias; Especialización en Medicina Interna	Pregrado en medicina	Especialización en Medicina Interna	Pregrado en medicina	Pregrado en medicina (Rotación de pediatría)	Pregrado en medicina (Rotación de cirugía general)
Experiencia previa	No reportado	No reportado	No reportado	Reportado como Comprensión/Exposición	No reportado	No reportado	No reportado	No reportado	0	No reportado	No reportado	No reportado	No reportado	Distribución de Familiaridad (n=146): Nada familiar (~1%); Ligeramente familiar (~23%); Moderadamente familiar	2	No reportado	No reportado

														(~35%); Muy familiar (~30%); Extremadamente familiar (~11%)			
Rendimiento previo	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado
LLMs empleados	GPT-3.5-turbo	ChatGPT-4o	ChatGPT 3.5	ChatGPT; Bard; Bing Chat; Claude (Mencionados como ejemplos)	gpt-3.5-turbo	No reportado	ChatGPT-4	ChatGPT	GPT-3.5-turbo-16k	GPT-4	GPT-4; GPT-4o; GPT-3.5; Claude-Opus; Claude-Sonnet-3-5; Llama-3-70B; Llama-3-8B; Mixtral 8x7B; Starling-7B-beta	GPT-3.5 Turbo 16k (OpenAI)	gpt-35-turbo	GPT-3/4	ChatGPT 3.5	ChatGPT	ChatGPT
Versión de los LLMs	GPT-3.5-turbo chat completion model	ChatGPT-4o	Versión gratuita	No reportado	No reportado	No reportado	Versión no especificada	No reportado	GPT-3.5-turbo-16k	No reportado	GPT-4 (01-25); GPT-4o (05-21); GPT-3.5 (06-13); Claude-Opus (04-22); Claude-Sonnet-3-5 (06-27); Llama-3-70B (No especificada); Llama-3-8B (No especificada); Mixtral 8x7B	GPT-3.5 Turbo 16k	gpt-3.5-turbo-0301 / gpt-3.5-turbo-0613 (Snapshot 4K context)	No reportado	No reportado	4	No reportado

											(No especificada); Starling-7B-beta (No especificada)						
Plataforma	Integrado en plataforma de robot social (Furhat)	Interacción por voz a través de aplicación en tableta	Interfaz web oficial	No reportado	Robot social (Furhat robotics) integrado con Modelo de Lenguaje a Gran Escala (LLM) vía software	Plataforma DingTalk; Plataforma Wenjuanxing	Interfaz web oficial (OpenAI)	No reportado	API integrada en software	Plataforma institucional / Específica del estudio	Interfaz de Programación de Aplicaciones (API) (API de Azure OpenAI, API de Anthropic); Despliegue local (en Computación de Alto Rendimiento (HPC) institucional seguro)	Interfaz de Programación de Aplicaciones (API) de OpenAI.	Interfaz de Programación de Aplicaciones (API) integrada en software	No reportado	Cuentas propias de estudiantes (interfaz web/aplicación estándar)	No reportado	No reportado
Fine tuning	No	No	No	No	No	No reportado	No	No reportado	No	No	No	No	No	No	No	No	No
Parámetros	No reportado	No reportado	No reportado	No reportado	No reportado	No reportado	Por defecto	No reportado	No reportado	No reportado	No reportado	No reportado	Temperatura = 0.15; Top_p = 0.95	No reportado / No controlado	temperature = 0.2; temperature = 0.8	No reportado	No reportado
Costo	No reportado	Gratuito (cubierto por investigadores/institución)	Gratuito (acceso libre)	No reportado (probablemente gratuito)	No reportado	No reportado	No reportado	No reportado	No reportado	Gratuito (cubierto por la institución)	Costo por uso (Interfaz de Programación de Aplicaciones - API); Despliegue	No reportado	Costo por uso (Interfaz de Programación de Aplicaciones - API)	No reportado	Gratuito (acceso libre)	No reportado	No reportado

											local (costo computacional institucional)						

Tabla suplementaria 4. Evaluación de calidad: estudios experimentales aleatorizados - RoB-2

Criterio	Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial	Implementation and evaluation of an optimized surgical clerkship teaching model utilizing chatgpt
Dominio 1: proceso de aleatorización (y juicio del dominio)	**(algunas preocupaciones)**	**(algunas preocupaciones)**
Generación aleatoria de la secuencia de asignación	Sí	Sí
Ocultamiento de la secuencia de asignación	No hay info.	No hay info.
Diferencias basales que sugieran un problema	No	No hay info.
Dominio 2: desviaciones de las intervenciones (y juicio del dominio)	**(bajo)**	**(bajo)**
Participantes conscientes de su asignación	No	Sí
Personal encargado consciente de la asignación	No hay info.	Sí
Desviaciones de la intervención debidas al contexto del ensayo	Probablemente no	Probablemente no
Desviaciones que afectaron el resultado	N/a	N/a
Desviaciones equilibradas entre los grupos	N/a	N/a
Análisis apropiado según la intención de tratar	Sí	Sí
Impacto sustancial por no analizar según intención de tratar	N/a	N/a
Dominio 3: datos faltantes de resultados (y juicio del dominio)	**(bajo)**	**(bajo)**
Datos disponibles para todos o casi todos los participantes	Sí	Probablemente sí
Sesgo en el resultado no causado por datos faltantes	N/a	N/a
Dependencia de los valores verdaderos para los datos faltantes	N/a	N/a
Probable dependencia de los valores verdaderos para los datos faltantes	N/a	N/a
Dominio 4: medición del resultado (y juicio del dominio)	**(bajo)**	**(algunas preocupaciones)**
Método de medición inapropiado	No	No
Diferencias en la medición entre los grupos	No	Probablemente no
Evaluadores conscientes de la asignación	No	No hay info.
Evaluación influenciada por el conocimiento de la asignación	N/a	Probablemente sí
Probable influencia en la evaluación	N/a	Probablemente no
Dominio 5: selección del resultado informado (y juicio del dominio)	**(algunas preocupaciones)**	**(algunas preocupaciones)**

Análisis según plan preespecificado	No hay info.	No hay info.
Selección entre múltiples mediciones del resultado	Probablemente no	No
Selección entre múltiples análisis	Probablemente no	Probablemente no
Riesgo de sesgo global del estudio	**(algunas preocupaciones)**	**(algunas preocupaciones)**

Tabla suplementaria 5. Evaluación de calidad: estudios transversales o descriptivos - JBI-ACSS

Criterio	A Pilot Study of Medical Student Opinions on Large Language Models	Exploring medical student's outlook on use of artificial intelligence in medical education: a multicentric cross-sectional study in Jharkhand, India	Large language models in internal medicine residency: current use and attitudes among internal medicine residents	Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy
¿Los criterios de inclusión están claramente definidos?	Sí	Sí	Sí	Sí
¿Se describen en detalle los sujetos y el entorno del estudio?	Sí	Sí	Sí	Sí
¿La exposición se mide de forma válida y fiable?	No es claro	Sí	Sí	Sí
¿Se usan criterios objetivos y estándar para medir la condición?	Sí	Sí	Sí	Sí
¿Se identifican los factores de confusión?	Sí	No es claro	Sí	Sí
¿Se declaran estrategias para manejar los factores de confusión?	No	No	No	Sí
¿Los resultados se miden de forma válida y fiable?	No es claro	Sí	Sí	Sí
¿Se emplea un análisis estadístico apropiado?	Sí	Sí	Sí	Sí

Tabla suplementaria 6. Evaluación de calidad: estudios cualitativos - JBI-QS

Criterio	Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology
Congruencia entre la perspectiva filosófica declarada y la metodología utilizada	Sí
Congruencia entre la metodología y las preguntas u objetivos de investigación	Sí
Congruencia entre la metodología y los métodos de recolección de datos	Sí
Congruencia entre la metodología y la representación y análisis de los datos	Sí

Congruencia entre la metodología y la interpretación de los resultados	Sí
Declaración que sitúa al investigador (posición cultural, teórica o reflexiva)	No
Abordaje de la influencia del investigador sobre el desarrollo e interpretación del estudio	No es claro
Representación adecuada de los participantes y de sus voces (uso de citas, diversidad de perspectivas)	Sí
Ética de la investigación (aprobación por comité de ética y obtención de consentimiento informado)	Sí
Conclusiones que fluyen lógicamente del análisis e interpretación de los datos	Sí

Tabla suplementaria 7. Evaluación de calidad: estudios de métodos o enfoque mixto - MMAT

Criterio	Virtual Patient Simulations Using Social Robotics Combined With Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study	AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians	Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education	Advancing Problem-Based Learning with Clinical Reasoning for Improved Differential Diagnosis in Medical Education	MedSimAI: Simulation and Formative Feedback Generation to Enhance Deliberate Practice in Medical Education	Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis	Assessing the Usability of GutGPT: A Simulation Study of an AI Clinical Decision Support System for Gastrointestinal Bleeding Risk	Using ChatGPT for medical education: the technical perspective	ChatGPT as a Virtual Patient: Written Empathic Expressions During Medical History Taking	Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods
Las preguntas de investigación están claramente expresadas	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Los datos recogidos permiten responder a las preguntas de investigación	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Investigación cualitativa										
El enfoque cualitativo es apropiado para responder a las preguntas de investigación	-	.	Sí	Sí	.	Sí	Sí	Sí	Sí	Sí
Los métodos de recolección de datos	-	-	Sí	Sí	-	Sí	Sí	Sí	Sí	Sí

cualitativos son adecuados para responder a las preguntas de investigación										
Las conclusiones derivan adecuadamente de los datos	-	-	Sí	Sí	-	Sí	Sí	Sí	Sí	Sí
La interpretación de los resultados está suficientemente fundamentada en los datos	-	-	Sí	Sí	-	Sí	Sí	Sí	Sí	Sí
Existe coherencia entre fuentes, recolección, análisis e interpretación de los datos cualitativos	-	-	Sí	Sí	-	Sí	Sí	Sí	Sí	Sí
Ensayos aleatorizados										
La aleatorización se realizó de forma apropiada	-	-	-	Sí	-	-	-	-	-	Sí
Los grupos son comparables al inicio del estudio	-	-	-	No es claro	-	-	-	-	-	Sí
Los datos de resultado están completos para todos o casi todos los participantes aleatorizados	-	-	-	No	-	-	-	-	-	Sí
Los evaluadores de los resultados estaban cegados (no conocían la asignación)	-	-	-	No es claro	-	-	-	-	-	Sí
Los participantes siguieron la intervención asignada según lo previsto	-	-	-	Sí	-	-	-	-	-	No es claro
Estudios cuantitativos no aleatorizados										
Los participantes son representativos	-	-	No es claro	-	-	-	-	-	No es claro	-

de la población objetivo										
Las mediciones son adecuadas para el resultado y la intervención			Sí	-	-	-	-	-	Sí	-
Los datos de resultado están completos para todos los participantes	-	-	Sí	-	-	-	-	-	No	-
Los factores de confusión se tuvieron en cuenta en el diseño y el análisis	-	-	Sí	-	-	-	-	-	No es claro	-
La intervención se administró según lo previsto durante el estudio	-	-	Sí	-	-	-	-	-	Sí	-
Estudios cuantitativos descriptivos										
La estrategia de muestreo es relevante para responder a las preguntas de investigación	-	-	-	-	-	Sí	-	-	-	-
La muestra es representativa de la población objetivo	-	-	-	-	-	No es claro	-	-	-	-
Las mediciones utilizadas son apropiadas	-	-	-	-	-	Sí	-	-	-	-
El riesgo de sesgo por no respuesta es bajo	-	-	-	-	-	Sí	-	-	-	-
El análisis estadístico es apropiado para responder a las preguntas de investigación	-	-	-	-	-	Sí	-	-	-	-
Métodos mixtos										
La justificación para usar un diseño de métodos mixtos es adecuada	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Los diferentes componentes del estudio están integrados de forma	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí

efectiva para responder a las preguntas de investigación										
Los resultados de la integración de componentes cualitativos y cuantitativos están adecuadamente interpretados	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Se abordan adecuadamente las divergencias e inconsistencias entre los resultados cuantitativos y cualitativos	Sí	Sí	No aplica	Sí	No es claro	Sí	Sí	Sí	No aplica	No aplica
Los diferentes componentes del estudio cumplen los criterios de calidad de sus respectivas tradiciones metodológicas	Sí	Sí	Sí	No	No	Sí	No	No	No	Sí

REFERENCIAS

Agnihotri, A. P., Nagel, I. D., Artiaga, J. C. M., Guevarra, Ma. C. B., Sosuan, G. M. N., & Kalaw, F. G. P. (2024). Large language models in ophthalmology: A review of publications from top ophthalmology journals. *Ophthalmology Science*, 100681. https://doi.org/10.1016/J.XOPS.2024.100681

Ahmad, N., Chakraborty, H., & Sinha, R. (2024). Exploring medical student's outlook on use of artificial intelligence in medical education: a multicentric cross-sectional study in Jharkhand, India. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2396704

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., Norman, M. K., Mayer, R. E., & Mendoza, O. G. (2017). ¿CÓMO SE TRANSFORMAN LOS ESTUDIANTES EN APRENDICES AUTODIRIGIDOS? In *Siete principios basados en lainvestigación para una enseñanza inteligente* (1st ed., pp. 207–235). Editorial Universidad del Norte. http://ebookcentral.proquest.com/lib/icesi/detail.action?docID=4909233.

Amiri, H., Peiravi, S., rezazadeh shojaee, S. sara, Rouhparvarzamin, M., Nateghi, M. N., Etemadi, M. H., ShojaeiBaghini, M., Musaie, F., Anvari, M. H., & Asadi Anar, M. (2024). Medical, dental, and nursing students' attitudes and knowledge towards artificial intelligence: a systematic review and meta-analysis. *BMC Medical Education*, *24*(1), 1–12. https://doi.org/10.1186/S12909-024-05406-1/FIGURES/6

Angkurawaranon, S., Inmutto, N., Bannangkoon, K., Wonghan, S., Kham-Ai, T., Khumma, P., Daengpisut, K., Thabarsa, P., & Angkurawaranon, C. (2024). Attitudes and perceptions of Thai medical students regarding artificial intelligence in radiology and medicine. *BMC Medical Education*, *24*(1), 1188. https://doi.org/10.1186/S12909-024-06150-2/FIGURES/1

Arfaie, S., Sadegh Mashayekhi, M., Mofatteh, M., Ma, C., Ruan, R., MacLean, M. A., Far, R., Saini, J., Harmsen, I. E., Duda, T., Gomez, A., Rebchuk, A. D., Pingbei Wang, A., Rasiah, N., Guo, E., Fazlollahi, A. M., Rose Swan, E., Amin, P., Mohammed, S., … Das, S. (2024). ChatGPT and neurosurgical education: A crossroads of innovation and opportunity. *Journal of Clinical Neuroscience*, *129*, 110815. https://doi.org/10.1016/J.JOCN.2024.110815

Aster, A., Ragaller, S. V., Raupach, T., & Marx, A. (2025). ChatGPT as a Virtual Patient: Written Empathic Expressions During Medical History Taking. *Medical Science Educator*. https://doi.org/10.1007/s40670-025-02342-7

Ba, H., zhang, L., & Yi, Z. (2024). Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-05565-1

Bharatha, A., Ojeh, N., Rabbi, A. M. F., Campbell, M. H., Krishnamurthy, K., Layne-Yarde, R. N. A., Kumar, A., Springer, D. C. R., Connell, K. L., & Majumder, M. A. A. (2024).

Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. *Advances in Medical Education and Practice*, *15*, 393–400. https://doi.org/10.2147/AMEP.S457408

Borg, A., Georg, C., Jobs, B., Huss, V., Waldenlind, K., Ruiz, M., Edelbring, S., Skantze, G., & Parodis, I. (2025a). Virtual Patient Simulations Using Social Robotics Combined With Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/63312

Borg, A., Georg, C., Jobs, B., Huss, V., Waldenlind, K., Ruiz, M., Edelbring, S., Skantze, G., & Parodis, I. (2025b). Virtual Patient Simulations Using Social Robotics Combined With Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/63312

Borg, A., Jobs, B., Huss, V., Gentline, C., Espinosa, F., Ruiz, M., Edelbring, S., Georg, C., Skantze, G., & Parodis, I. (2024). Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. *Rheumatology International*. https://doi.org/10.1007/s00296-024-05731-0

Broshar, A. (2024, April 25). *What are LLMs? An intro into AI, models, tokens, parameters, weights, quantization and more - Koyeb*. Koyeb. https://www.koyeb.com/blog/what-are-large-language-models#deploying-llms-and-ai-workloads-into-the-world

Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M. N., Schur, L., Stummer, W., Holling, M., Lu, M. H., & Darici, D. (2024). Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-06399-7

Buabbas, A. J., Miskin, B., Alnaqi, A. A., Ayed, A. K., Shehab, A. A., Syed-Abdul, S., & Uddin, M. (2023). Investigating Students' Perceptions towards Artificial Intelligence in Medical Education. *Healthcare*, *11*(9), 1298. https://doi.org/10.3390/healthcare11091298

Busch, F., Hoffmann, L., Truhn, D., Ortiz-Prado, E., Makowski, M. R., Bressem, K. K., Adams, L. C., Zhang, L., Zatoński, T., Xu, L., van Wijngaarden, P., van Dijk, E. H. C., Tuncel, M., Truong, M. H., Toapanta-Yanchapaxi, L. N., Thulesius, H. O., Tanioka, S., Takeda, K., Tabakova, N. G., … Abdala, N. (2024). Global cross-sectional student survey on AI in medical, dental, and veterinary education and practice at 192 faculties. *BMC Medical Education*, *24*(1), 1–20. https://doi.org/10.1186/S12909-024-06035-4/TABLES/5

Chan, C., You, K., Chung, S., Giuffrè, M., Saarinen, T., Rajashekar, N., Pu, Y., Shin, Y. E., Laine, L., Wong, A., Kizilcec, R., Sekhon, J., & Shung, D. (2023). *Assessing the*

Usability of GutGPT: A Simulation Study of an AI Clinical Decision Support System for Gastrointestinal Bleeding Risk. http://arxiv.org/abs/2312.10072

Chan, K. Y., Yuen, T. H., & Co, M. (2025). Using ChatGPT for medical education: the technical perspective. *BMC Medical Education*, *25*(1). https://doi.org/10.1186/s12909-025-06785-9

Cherrez-Ojeda, I., Gallardo-Bastidas, J. C., Robles-Velasco, K., Osorio, M. F., Velez Leon, E. M., Leon Velastegui, M., Pauletto, P., Aguilar-Díaz, F. C., Squassi, A., González Eras, S. P., Cordero Carrasco, E., Chavez Gonzalez, K. L., Calderon, J. C., Bousquet, J., Bedbrook, A., & Faytong-Haro, M. (2024). Understanding Health Care Students' Perceptions, Beliefs, and Attitudes Toward AI-Powered Language Models: Cross-Sectional Study. *JMIR Medical Education*, *10*, e51757. https://doi.org/10.2196/51757

Chien, C. W., & Tai, Y.-M. (2024). Performances of Large Language Models in Detecting Psychiatric Diagnoses from Chinese Electronic Medical Records: Comparisons between GPT-3.5, GPT-4, and GPT-4o. *Taiwanese Journal of Psychiatry*, *38*(3), 134–141. https://doi.org/10.4103/TPSY.TPSY_25_24

Choi, H.-W., & Abdirayimov, S. (2024). Revolutionizing Education: The Era of Large Language Models. *Journal of Multimedia Information System*, *11*(1), 97–100. https://doi.org/10.33851/JMIS.2024.11.1.97

Craig, L., Laskowski, N., & Tucci, L. (2024). *A guide to artificial intelligence in the enterprise*. TechRadar. https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence

Delavari, S., Barzkar, F., M. J. P. Rikers, R., Pourahmadi, M., Soltani Arabshahi, S. K., Keshtkar, A., Dargahi, H., Yaghmaei, M., & Monajemi, A. (2024). Teaching and learning clinical reasoning skill in undergraduate medical students: A scoping review. *PLOS ONE*, *19*(10), e0309606. https://doi.org/10.1371/journal.pone.0309606

Dong, B., Bai, J., Xu, T., & Zhou, Y. (2024). Large Language Models in Education: A Systematic Review. *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, 131–134. https://doi.org/10.1109/CSTE62025.2024.00031

Durning, S. J., Jung, E., Kim, D.-H., & Lee, Y.-M. (2024). Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside. *Korean Journal of Medical Education*, *36*(2), 145–155. https://doi.org/10.3946/kjme.2024.292

Fried, A. J., Dorn, S. D., Leland, W. J., Mullen, E., Williams, D. M., Zaas, A. K., MacGuire, J., & Bynum, D. L. (2024). Large language models in internal medicine residency: current use and attitudes among internal medicine residents. *Discover Artificial Intelligence*, *4*(1). https://doi.org/10.1007/s44163-024-00173-w

Fürstenberg, S., Helm, T., Prediger, S., Kadmon, M., Berberat, P. O., & Harendza, S. (2020). Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. *BMC Medical Education*, *20*(1), 1–7. https://doi.org/10.1186/S12909-020-02260-9/TABLES/4

Ganjavi, C., Eppler, M., O'Brien, D., Ramacciotti, L. S., Ghauri, M. S., Anderson, I., Choi, J., Dwyer, D., Stephens, C., Shi, V., Ebert, M., Derby, M., Yazdi, B., & Cacciamani, G. E. (2024). ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. *PLOS Digital Health*, *3*(9), e0000596. https://doi.org/10.1371/journal.pdig.0000596

Gonzalez, D. T., Djulbegovic, M. B., Kim, C., Antonietti, M., Gameiro, G. R., & Uversky, V. (2024). A Comparative Study of Large Language Models in Explaining Intrinsically Disordered Proteins. *Qeios*, *6*(9). https://doi.org/10.32388/5D952O.2

Guzmán-Valdivia-Gómez, G., Guzmán-Valdivia-Talavera, P., & García-Cervantes, A. (2022). Razonamiento clínico: aspectos prácticos que permiten la facilitación de su desarrollo. *Revista Medica Del Instituto Mexicano Del Seguro Social*, *60*(6), 708–714. http://www.ncbi.nlm.nih.gov/pubmed/36283081 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC10395931

Hashemi-Pour, C. (2023). *What is generative AI? Everything you need to know*. TechRadar. https://www.techtarget.com/searchenterpriseai/definition/OpenAI

Hicke, Y., Geathers, J., Rajashekar, N., Chan, C., Jack, A. G., Sewell, J., Preston, M., Cornes, S., Shung, D., & Kizilcec, R. (2025). *MedSimAI: Simulation and Formative Feedback Generation to Enhance Deliberate Practice in Medical Education*. http://arxiv.org/abs/2503.05793

Huang, Y., Xu, B. B., Wang, X. Y., Luo, Y. C., Teng, M. M., & Weng, X. (2024). Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-06575-9

Jackson, P., Ponath Sukumaran, G., Babu, C., Tony, M. C., Jack, D. S., Reshma, V. R., Davis, D., Kurian, N., & John, A. (2024). Artificial intelligence in medical education - perception among medical students. *BMC Medical Education*, *24*(1), 804. https://doi.org/10.1186/s12909-024-05760-0

Javed, K., & Arooj, M. (2023). EXPLORING PERSPECTIVES OF MEDICAL FACULTY AND UNDERGRADUATE MEDICAL STUDENTS ON CHATGPT IN MEDICAL EDUCATION. *Pakistan Postgraduate Medical Journal*, *34*(04), 197–202. https://doi.org/10.51642/ppmj.v34i04.642

Jebreen, K., Radwan, E., Kammoun-Rebai, W., Alattar, E., Radwan, A., Safi, W., Radwan, W., & Alajez, M. (2024). Perceptions of undergraduate medical students on artificial

intelligence in medicine: mixed-methods survey study from Palestine. *BMC Medical Education*, *24*(1), 1–19. https://doi.org/10.1186/S12909-024-05465-4/TABLES/7

Jiang, X., Yan, L., Vavekanand, R., & Hu, M. (2024). Large Language Models in Healthcare Current Development and Future Directions. In *Preprints*. https://doi.org/10.20944/preprints202407.0923.v1

Jiao, J., Afroogh, S., Chen, K., Atkinson, D., & Dhurandhar, A. (2024). The global landscape of academic guidelines for generative AI and Large Language Models. *Preprint*. http://arxiv.org/abs/2406.18842

Kahneman, D. (2013). *Thinking, Fast and Slow* (1st ed.). Farrar, Straus and Giroux.

Kämmer, J. E., Hautz, W. E., Krummrey, G., Sauter, T. C., Penders, D., Birrenbach, T., & Bienefeld, N. (2024). Effects of interacting with a large language model compared with a human coach on the clinical diagnostic process and outcomes among fourth-year medical students: study protocol for a prospective, randomised experiment using patient vignettes. *BMJ Open*, *14*(7), e087469. https://doi.org/10.1136/bmjopen-2024-087469

Kanjee, Z., Crowe, B., & Rodman, A. (2023). Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA*, *330*(1), 78. https://doi.org/10.1001/jama.2023.8288

Kassirer, J. P. (2010). Teaching Clinical Reasoning: Case-Based and Coached. *Academic Medicine*, *85*(7), 1118–1124. https://doi.org/10.1097/ACM.0b013e3181d5dd0d

Kulasegaram, K., Manzone, J. C., Ku, C., Skye, A., Wadey, V., & Woods, N. N. (2015). Cause and Effect. *Academic Medicine*, *90*, S63–S69. https://doi.org/10.1097/ACM.0000000000000896

Laupichler, M. C., Aster, A., Meyerheim, M., Raupach, T., & Mergen, M. (2024). Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. *BMC Medical Education*, *24*(1), 401. https://doi.org/10.1186/s12909-024-05400-7

Lawson McLean, A. (2024). Constructing knowledge: the role of AI in medical learning. *Journal of the American Medical Informatics Association*, *31*(8), 1797–1798. https://doi.org/10.1093/jamia/ocae124

Li, Q., & Qin, Y. (2023). AI in medical education: medical student perception, curriculum recommendations and design suggestions. *BMC Medical Education*, *23*(1), 852. https://doi.org/10.1186/s12909-023-04700-8

Liu, D. S., Sawyer, J., Luna, A., Aoun, J., Wang, J., Boachie, Lord, Halabi, S., & Joe, B. (2022). Perceptions of US Medical Students on Artificial Intelligence in Medicine: Mixed Methods Survey Study. *JMIR Medical Education*, *8*(4), e38325. https://doi.org/10.2196/38325

Liu, J., ChangyuWang, & Liu, S. (2024). Applications of Large Language Models in Clinical Practice: Path, Challenges, and Future Perspectives. In *OSF Preprints*. https://doi.org/10.31219/osf.io/82bjd

Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. *Medical Education*. https://doi.org/10.1111/medu.15402

Mannekote, A., Davies, A., Pinto, J. D., Zhang, S., Olds, D., Schroeder, N. L., Lehman, B., Zapata-Rivera, D., & Zhai, C. (2024). Large language models for whole-learner support: opportunities and challenges. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1460364

McCoy, L. G., Ci Ng, F. Y., Sauer, C. M., Yap Legaspi, K. E., Jain, B., Gallifant, J., McClurkin, M., Hammond, A., Goode, D., Gichoya, J., & Celi, L. A. (2024). Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: a narrative review. *BMC Medical Education*, *24*(1), 1096. https://doi.org/10.1186/S12909-024-06048-Z/TABLES/1

Mondal, H., De, R., Mondal, S., & Juhi, A. (2024). A large language model in solving primary healthcare issues: A potential implication for remote healthcare and medical education. *Journal of Education and Health Promotion*, *13*(1). https://doi.org/10.4103/jehp.jehp_688_23

Mondal, H., Karri, J. K. K., Ramasubramanian, S., Mondal, S., Juhi, A., & Gupta, P. (2024). A qualitative survey on perception of medical students on the use of large language models for educational purposes. *Advances in Physiology Education*. https://doi.org/10.1152/advan.00088.2024

Moreno, A. C., & Bitterman, D. S. (2024). Toward Clinical-Grade Evaluation of Large Language Models. *International Journal of Radiation Oncology*Biology*Physics*, *118*(4), 916–920. https://doi.org/10.1016/j.ijrobp.2023.11.012

Nalpas, M. (2024, May 30). *Understand LLM sizes*. Web.Dev. https://web.dev/articles/llm-sizes#llm-sizes

Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, *154*, 102900. https://doi.org/10.1016/j.artmed.2024.102900

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The Mini-CEX: A Method for Assessing Clinical Skills. *Annals of Internal Medicine*, *138*(6), 476–481. https://doi.org/10.7326/0003-4819-138-6-200303180-00012;REQUESTEDJOURNAL:JOURNAL:AIM;JOURNAL:JOURNAL:AIM;WGROUP:STRING:PUBLICATION

Omiye J. A., et al., & 2024. (2024). Large language models in medicine: The potentials and pitfalls. *Annals of Internal Medicine*. https://doi.org/https://doi.org/10.7326/m23-2772

Öncü, S., Torun, F., & Ülkü, H. H. (2025). AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Medical Education*, *25*(1). https://doi.org/10.1186/s12909-025-06877-6

Ortiz Monsalve, L. C., Salazar López, R., Hernández Moreno, A., Castellanos Robayo, J., Rubio González, R. Y., Baquero Haeberlin, R., Álvarez Peñalosa, E., Escobar Cifuentes, E., Beltrán Dussán, E., Fandiño Frankly, J., Cardona Arias, J. D., Ardila Ardila, E., Torres, M., Pinzón Junca, A., Quintero, A., Heredia, R. A., Cano, C. A., Beltrán D, E. H., Castañeda, J., … Maldonado García, M. Á. (2014). PERFIL Y COMPETENCIAS PROFESIONALES DEL MÉDICO EN COLOMBIA. *Ministerio de Salud y Protección Social*. https://www.sispro.gov.co/observatorios/oniea/Paginas/CompetenciasProfesionales.aspx

Pan, G., & Ni, J. (2024). A cross sectional investigation of ChatGPT-like large language models application among medical students in China. *BMC Medical Education*, *24*(1), 908. https://doi.org/10.1186/s12909-024-05871-8

Parodis, I., Andersson, L., Durning, S. J., Hege, I., Knez, J., Kononowicz, A. A., Lidskog, M., Petreski, T., Szopa, M., & Edelbring, S. (2021). Clinical Reasoning Needs to Be Explicitly Addressed in Health Professions Curricula: Recommendations from a European Consortium. *International Journal of Environmental Research and Public Health*, *18*(21), 11202. https://doi.org/10.3390/ijerph182111202

Pears, M., & Konstantinidis, S. T. (2024). The Impact of Aligning Artificial Intelligence Large Language Models With Bloom's Taxonomy in Healthcare Education. In *Advances in Business Information Systems and Analytics* (pp. 166–192). https://doi.org/10.4018/979-8-3693-3003-6.ch008

Ragolane, M., Patel, S., & Salikram, P. (2024). AI Versus Human Graders: Assessing the Role of Large Language Models in Higher Education. *Asian Journal of Education and Social Studies*, *50*(10), 244–263. https://doi.org/10.9734/ajess/2024/v50i101616

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, Most. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2023). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *TechRxiv*. https://doi.org/10.36227/techrxiv.24171183.v1

Rashid, M. M., Atilgan, N., Dobres, J., Day, S., Penkova, V., Küçük, M., Clapp, S. R., & Sawyer, B. D. (2024). Humanizing AI in Education: A Readability Comparison of LLM and Human-Created Educational Content. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/10711813241261689

Safranek, C. W., Sidamon-Eristoff, A. E., Gilson, A., & Chartash, D. (2023). The Role of Large Language Models in Medical Education: Applications and Implications. *JMIR Medical Education*, *9*, e50945. https://doi.org/10.2196/50945

Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, *11*(6), 887. https://doi.org/10.3390/healthcare11060887

Schwartzstein, R. M. (2024). CLINICAL REASONING AND ARTIFICIAL INTELLIGENCE: CAN AI REALLY THINK? *Transactions of the American Clinical and Climatological Association*, *134*, 133–145. http://www.ncbi.nlm.nih.gov/pubmed/39135584

Shah, K., Xu, A. Y., Sharma, Y., Daher, M., McDonald, C., Diebo, B. G., & Daniels, A. H. (2024). Large Language Model Prompting Techniques for Advancement in Clinical Medicine. *Journal of Clinical Medicine*, *13*(17), 5101. https://doi.org/10.3390/jcm13175101

Shakur, A. H., Holcomb, M. J., Hein, D., Kang, S., Dalton, T. O., Campbell, K. K., Scott, D. J., & Jamieson, A. R. (2024). *Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis*. http://arxiv.org/abs/2410.12858

Shusterman, R., Waters, A., O'Neill, S., Luu, P., & Tucker, D. (2024). An Active Inference Strategy for Prompting Reliable Responses from Large Language Models in Medical Practice. In *Preprint*. https://doi.org/10.21203/rs.3.rs-5079480/v1

Suresh, S., & Misra, S. M. (2024). Large Language Models in Pediatric Education: Current Uses and Future Potential. *Pediatrics*, *154*(3). https://doi.org/10.1542/peds.2023-064683

Telenti, A., Auli, M., Hie, B. L., Maher, C., Saria, S., & Ioannidis, J. P. A. (2024). Large language models for science and medicine. *European Journal of Clinical Investigation*, *54*(6). https://doi.org/10.1111/eci.14183

Togunwa, T. O., Ajibade, A., Uche-Orji, C., & Olatunji, R. (2024). Exploring the Potentials of Large Language Models in Vascular and Interventional Radiology: Opportunities and Challenges. *The Arab Journal of Interventional Radiology*, *08*(02), 063–069. https://doi.org/10.1055/s-0044-1782663

Waldock, W. J., Lam, G., Baptista, A. V. M. T., Walls, R., & Sam, A. H. (2024). *Which curriculum components do medical students find most helpful for evaluating AI outputs? (Preprint)*. https://doi.org/10.21203/rs.3.rs-4768657/v1

Wang, C., Li, S., Lin, N., Zhang, X., Han, Y., Wang, X., Liu, D., Tan, X., Pu, D., Li, K., Qian, G., & Yin, R. (2025). Application of Large Language Models in Medical Training Evaluation—Using ChatGPT as a Standardized Patient: Multimetric Assessment. *Journal of Medical Internet Research*, *27*(1), e59435. https://doi.org/10.2196/59435

Xu, A. Y., Piranio, V. S., Speakman, S., Rosen, C. D., Lu, S., Lamprecht, C., Medina, R. E., Corrielus, M., Griffin, I. T., Chatham, C. E., Abchee, N. J., Stribling, D., Huynh, P. B., Harrell, H., Shickel, B., & Brennan, M. (2024). A Pilot Study of Medical Student Opinions on Large Language Models. *Cureus*. https://doi.org/10.7759/cureus.71946

Xu, Y., Shao, Y., Dong, J., Shi, S., Jiang, C., & Li, Q. (2025). Advancing Problem-Based Learning with Clinical Reasoning for Improved Differential Diagnosis in Medical Education. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–32. https://doi.org/10.1145/3706598.3713772

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). *Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review*. https://doi.org/10.1111/bjet.13370

Yanagita, Y., Yokokawa, D., Fukuzawa, F., Uchida, S., Uehara, T., & Ikusaka, M. (2024). Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. *BMC Medical Education*, *24*(1), 1–8. https://doi.org/10.1186/S12909-024-05534-8/TABLES/1

Zarei, M., Eftekhari Mamaghani, H., Abbasi, A., & Hosseini, M.-S. (2024). Application of artificial intelligence in medical education: A review of benefits, challenges, and solutions. *Medicina Clínica Práctica*, *7*(2), 100422. https://doi.org/10.1016/j.mcpsp.2023.100422

Zarei, M., Zarei, M., Hamzehzadeh, S., Oliyaei, S. S. B., & Hosseini, M. S. (2024). ChatGPT, a Friend or a Foe in Medical Education: A Review of Strengths, Challenges, and Opportunities. *Shiraz E-Medical Journal 2024 25:6*, *25*(6). https://doi.org/10.5812/SEMJ-145840

Zdravkova, K., Dalipi, F., & Ahlgren, F. (2023). Integration of Large Language Models into Higher Education: A Perspective from Learners (Preprint). *2023 International Symposium on Computers in Education (SIIE)*, 1–6. https://doi.org/10.1109/SIIE59826.2023.10423681

Zheng, K., Shen, Z., Chen, Z., Che, C., & Zhu, H. (2024). Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-05977-z