ARTÍCULO DE REVISIÓN

Desafíos prácticos y éticos de los grandes modelos lingüísticos en la educación: una revisión sistemática del alcance

Lixiang Yan[1] | Lele Sha[1] | Linxuan Zhao[1] | Yuheng | Li[1] Roberto Martínez-Maldonado[1] | Guanliang Chen[1] | Xinyu Li[1] | Yueqiao Jin[1] Dragan Gašević[1]

[1]Centro de Análisis del Aprendizaje de Monash, Facultad de Tecnología de la Información, Universidad de Monash, Clayton, Victoria, Australia

Correspondencia
Lixiang Yan, Centro de Análisis del Aprendizaje de Monash, Facultad de Tecnología de la Información, Universidad de Monash, 20 Exhibition Walk, Clayton, VIC 3800, Australia
Correo electrónico: jimmie.yan@monash.edu

Información de financiación
Esta investigación fue financiada al menos en parte por el Consejo Australiano de Investigación. (DP210100060) y Fundación Jacobs (Beca de Investigación).

Las innovaciones en tecnología educativa que aprovechan los grandes modelos lingüísticos (LLM) han demostrado el potencial de automatizar el laborioso proceso de generar y analizar contenido textual. Si bien se han desarrollado diversas innovaciones para automatizar diversas tareas educativas (p. ej., generación de preguntas, retroalimentación y calificación de ensayos), existen inquietudes sobre su viabilidad y ética. Dichas inquietudes pueden obstaculizar la investigación futura y la adopción de innovaciones basadas en LLM en contextos educativos auténticos. Para abordar esto, realizamos una revisión sistemática de 118 artículos revisados por pares publicados desde 2017 para identificar el estado actual de la investigación sobre el uso de LLM para automatizar y respaldar tareas educativas. Los hallazgos revelaron 53 casos de uso de LLM en la automatización de tareas educativas, categorizados en nueve categorías principales: perfilado/etiquetado, detección, calificación, apoyo a la enseñanza, predicción, representación del conocimiento, retroalimentación, generación de contenido y recomendación. Además, también identificamos varios desafíos prácticos y éticos, como la baja disponibilidad tecnológica, la falta de replicabilidad y transparencia, y consideraciones insuficientes de privacidad y beneficencia. Los hallazgos se resumieron en tres recomendaciones para

Estudios futuros, incluyendo la actualización de innovaciones existentes con modelos de vanguardia (p. ej., GPT-3/4), la adopción de modelos/sistemas de código abierto y la adopción de un enfoque centrado en el ser humano durante todo el proceso de desarrollo. Dado que la intersección entre la IA y la educación está en constante evolución, los hallazgos de este estudio pueden servir como punto de referencia esencial para los investigadores, permitiéndoles aprovechar las fortalezas, aprender de las limitaciones y descubrir posibles oportunidades de investigación que ofrecen ChatGPT y otros modelos de IA generativa.

KEYWO RDS

modelos de lenguaje grandes, modelos de lenguaje preentrenados, inteligencia artificial, educación, revisión sistemática del alcance, GPT-3, BERT, ChatGPT

Notas del profesional

Lo que se sabe actualmente sobre este tema

- Generar y analizar contenido basado en texto son tareas que consumen mucho tiempo y son laboriosas.
- Los modelos lingüísticos de gran tamaño son capaces de analizar de forma eficiente una cantidad de contenido textual sin precedentes y completar tareas complejas de generación y procesamiento del lenguaje natural.
- Se han utilizado cada vez más modelos lingüísticos de gran tamaño para desarrollar tecnologías educativas que apuntan a automatizar la generación y el análisis de contenido textual, como la generación automatizada de preguntas y la calificación de ensayos.

Lo que este artículo añade

- Una lista completa de 53 tareas educativas diferentes que podrían beneficiarse potencialmente de las innovaciones basadas en LLM a través de la automatización.
- Una evaluación estructurada de la viabilidad y la ética de las innovaciones existentes basadas en LLM desde siete aspectos importantes utilizando marcos establecidos.
- Tres recomendaciones que podrían potencialmente respaldar estudios futuros para desarrollar innovaciones basadas en LLM que sean prácticas y éticas para implementar en contextos educativos auténticos.

Implicaciones para los profesionales

- Actualizar las innovaciones existentes con modelos de última generación puede reducir aún más la cantidad de esfuerzo manual necesario para adaptar los modelos existentes a diferentes tareas educativas.
- Es necesario mejorar los estándares de presentación de informes de las investigaciones empíricas que tienen como objetivo desarrollar tecnologías educativas utilizando modelos lingüísticos amplios.

• Adoptar un enfoque centrado en el ser humano a lo largo de todo el proceso de desarrollo podría contribuir a resolver los desafíos prácticos y éticos de los grandes modelos lingüísticos en la educación.

1 | INTRODUCCIÓN

Los avances en inteligencia artificial generativa (IA) y modelos lingüísticos extensos (LLM) han impulsado el desarrollo de numerosas innovaciones en tecnología educativa que buscan automatizar las tareas, a menudo laboriosas y laboriosas, de generar y analizar contenido textual (p. ej., generar preguntas abiertas y analizar encuestas de retroalimentación estudiantil) (Kasneci et al., 2023; Wollny et al., 2021; Leiker et al., 2023). Los LLM son modelos de inteligencia artificial generativa entrenados con una gran cantidad de datos textuales, capaces de generar contenido textual de forma similar a la humana a partir de entradas de lenguaje natural. En concreto, estos LLM, como las Representaciones de Codificador Bidireccional a partir de Transformadores (BERT) (Devlin et al., 2018) y el Transformador Generativo Pre-entrenado (GPT) (Brown et al., 2020), utilizan mecanismos de aprendizaje profundo y de autoatención (Vaswani et al., 2017) para atender selectivamente a las diferentes partes de los textos de entrada, en función del enfoque de las tareas actuales, lo que permite que el modelo aprenda patrones complejos y relaciones entre los contenidos textuales, como sus relaciones semánticas, contextuales y sintácticas (Min et al., 2021; Liu et al., 2023). Dado que varios LLM (p. ej., GPT-3 y Codex) han sido preentrenados con cantidades masivas de datos en múltiples disciplinas, son capaces de completar tareas de procesamiento del lenguaje natural con poco (aprendizaje de pocos intentos) o ningún entrenamiento adicional (aprendizaje de cero intentos) (Brown et al., 2020; Wu et al., 2023). Esto podría reducir las barreras tecnológicas para las innovaciones basadas en LLM, ya que investigadores y profesionales pueden desarrollar nuevas tecnologías educativas al ajustar los LLM a tareas educativas específicas sin comenzar desde cero (Caines et al., 2023; Sridhar et al., 2023). El reciente lanzamiento de ChatGPT, un chatbot de IA generativa basado en LLM que solo requiere indicaciones en lenguaje natural sin entrenamiento ni ajuste adicional del modelo (OpenAI, 2023), ha reducido aún más la barrera para que las personas sin experiencia tecnológica aprovechen las capacidades generativas de los LLM.

Aunque la investigación educativa que aprovecha los LLM para desarrollar innovaciones tecnológicas para automatizar tareas educativas aún no ha alcanzado su máximo potencial (es decir, la mayoría de los trabajos se han centrado en mejorar el rendimiento de los modelos (Kurdi et al., 2020; Ramesh y Sanampudi, 2022)), un creciente volumen de literatura sugiere cómo las diferentes partes interesadas podrían beneficiarse potencialmente de dichas innovaciones. En concreto, estas innovaciones podrían desempeñar un papel fundamental a la hora de abordar los altos niveles de estrés y agotamiento de los docentes al reducir sus pesadas cargas de trabajo mediante la automatización de tareas puntuales que consumen mucho tiempo (Carroll et al., 2022), como la generación de preguntas (Kurdi et al., 2020; Bulut y Yildirim-Erbasli, 2022; Oleny, 2023), la provisión de retroalimentación (Cavalcanti et al., 2021; Nye et al., 2023), la calificación de ensayos (Ramesh y Sanampudi, 2022) y las respuestas cortas (Zeng et al., 2023). Estas innovaciones también podrían beneficiar potencialmente tanto a los estudiantes como a las instituciones al mejorar la eficiencia de procesos administrativos a menudo tediosos, como la recomendación de recursos de aprendizaje, la recomendación de cursos y la evaluación de la retroalimentación de los estudiantes (Zawacki-Richter et al., 2019; Wollny et al., 2021; Sridhar et al., 2023).

A pesar de la creciente evidencia empírica del potencial de los LLM para automatizar una amplia gama de tareas educativas, ningún trabajo existente ha revisado sistemáticamente los desafíos prácticos y éticos de estas innovaciones basadas en LLM. Comprender estos desafíos es esencial para desarrollar tecnologías responsables, ya que las innovaciones basadas en LLM (p. ej., ChatGPT) podrían contener sesgos similares a los humanos basados en las normas éticas y morales existentes de la sociedad, como la herencia de conocimiento sesgado y tóxico (p. ej., sesgos de género y raciales) cuando se entrenan con datos de texto de internet sin filtrar (Schramowski et al., 2022). Revisiones sistemáticas previas se han centrado en investigar estos problemas relacionados con un escenario de aplicación específico de las innovaciones basadas en LLM (p. ej., generación de preguntas, calificación de ensayos, chatbots o retroalimentación automatizada) (Kurdi et al., 2020; Cavalcanti et al., 2021; Wollny et al., 2021; Ramesh y Sanampudi, 2022). Los desafíos prácticos y éticos

Los desafíos de los LLM en la automatización de diferentes tipos de tareas educativas siguen siendo inciertos. Comprender estos desafíos es esencial para traducir los hallazgos de la investigación en tecnologías educativas que las partes interesadas (p. ej., estudiantes, docentes e instituciones) puedan utilizar en prácticas auténticas de enseñanza y aprendizaje (Adams et al., 2021).

El estudio actual es la primera revisión sistemática de alcance que tuvo como objetivo abordar esta brecha mediante la revisión de *estado actual de la investigación* sobre el uso de LLM para automatizar tareas educativas e identificar *práctico y ético* Desafíos de adoptar estas innovaciones basadas en LLMs en contextos educativos auténticos. Se incluyeron 118 publicaciones revisadas por pares de cuatro bases de datos destacadas en esta revisión, siguiendo el protocolo PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Page et al., 2021). Se realizó un análisis temático inductivo para extraer detalles sobre los diferentes tipos de tareas educativas, partes interesadas, LLMs y tareas de aprendizaje automático investigadas en la literatura previa. La viabilidad de las innovaciones basadas en LLMs se evaluó a través de la perspectiva de la preparación tecnológica, el rendimiento del modelo y su replicabilidad. Finalmente, se evaluó la ética de estas innovaciones investigando la transparencia del sistema, la privacidad, la igualdad y la beneficencia.

La contribución de este artículo a la comunidad de tecnología educativa es triple: 1) resumimos sistemáticamente una lista completa de 53 tareas educativas diferentes que podrían beneficiarse de las innovaciones basadas en LLM mediante la automatización; 2) presentamos una evaluación estructurada de la viabilidad y la ética de las innovaciones existentes basadas en LLM, basada en siete aspectos importantes utilizando marcos establecidos (p. ej., el índice de transparencia [Chaudhry et al., 2022]); y 3) proponemos tres recomendaciones que podrían respaldar estudios futuros para desarrollar innovaciones basadas en LLM que se implementen de forma práctica y ética en contextos educativos reales. Dado que la intersección entre LLM y educación está en constante evolución, los hallazgos de esta revisión sistemática pueden servir como punto de referencia esencial para los investigadores, permitiéndoles aprovechar las fortalezas, aprender de las limitaciones y descubrir oportunidades potenciales para nuevos LLM en el apoyo a la investigación y la práctica educativas. En concreto, los trabajos emergentes deben considerar cuidadosamente los desafíos prácticos y éticos identificados en este estudio, al tiempo que exploran las oportunidades de investigación que ofrecen ChatGPT y otros modelos de IA generativa.

2 | FONDO

En esta sección, primero se definen los términos clave, en particular los de practicidad y ética en el contexto de la tecnología educativa. A continuación, se ofrece una visión general de revisiones sistemáticas previas sobre programas de maestría en derecho (LLM) en educación. Finalmente, se presentan las preguntas de investigación basadas en las lagunas identificadas en la literatura existente.

2.1 | Sentido práctico

Se han propuesto varios marcos teóricos sobre la viabilidad de integrar innovaciones tecnológicas en entornos educativos. Por ejemplo, las barreras de primer y segundo orden al cambio de Ertmer (1999) se centraron en las condiciones externas del sistema educativo (p. ej., la disponibilidad de la infraestructura) y los estados internos del profesorado (p. ej., las creencias personales). Becker (2000) sugirió además que, para que las innovaciones tecnológicas tengan beneficios reales en el apoyo a las prácticas pedagógicas, estas innovaciones deben ser de fácil acceso, respaldar las creencias pedagógicas constructivistas, ser adaptables a los cambios curriculares y ser compatibles con el nivel de conocimientos y habilidades del profesorado. Estos factores también se presentaron en un marco anterior del índice de viabilidad (Doyle y Ponder, 1977), que resumía tres componentes críticos para la integración de tecnologías educativas: el grado de viabilidad de la adopción, la relación coste-beneficio y la alineación con las prácticas y creencias existentes. Con base en estos marcos teóricos previos y considerando lo reciente de las innovaciones basadas en LLM (que surgieron en los últimos cinco años),

Los desafíos prácticos de las innovaciones basadas en LLM en la automatización de tareas educativas pueden evaluarse desde tres perspectivas principales. En primer lugar, evaluar la disponibilidad tecnológica de estas innovaciones es esencial para determinar si existe evidencia empírica que respalde su integración y funcionamiento exitosos en contextos educativos reales. En segundo lugar, evaluar el rendimiento del modelo podría aportar información valiosa sobre los costos y beneficios de adoptar estas innovaciones, como comparar los beneficios de la automatización con los costos de predicciones inexactas. Por último, comprender si estas innovaciones son metodológicamente replicables podría ser importante para futuros estudios que investiguen su adecuación a diferentes contextos educativos y actores clave. En la Sección 3.2, profundizamos en los elementos de evaluación para cada desafío.

2.2 | Ética

La IA ética es un tema de debate frecuente en diversas comunidades, como las de analítica de aprendizaje, IA en educación, minería de datos educativos y tecnología educativa (Adams et al., 2021; Pardo y Siemens, 2014). Existen debates en curso sobre la ética de la IA en educación, con enfoques mixtos en la ética algorítmica y humana entre las comunidades de minería de datos educativos e IA en educación (Holmes y Porayska-Pomsta, 2022). Dado que estos debates continúan, resulta difícil identificar una definición establecida de IA ética en estos campos. Sin embargo, la ética ya se ha investigado y abordado exhaustivamente en un ámbito exclusivo de la IA en educación, concretamente, la analítica de aprendizaje (Pardo y Siemens, 2014; Selwyn, 2019). Basándose en la definición establecida de ética del campo de la analítica del aprendizaje (Pardo y Siemens, 2014), la ética de las innovaciones basadas en LLM puede definirse como la sistematización de las funcionalidades y resultados apropiados e inapropiados de estas innovaciones, según lo determinado por todas las partes interesadas (p. ej., estudiantes, docentes, padres e instituciones). Por ejemplo, Khosravi et al. (2022) explicaron que la ética de los sistemas de tecnología educativa impulsados por IA debe implicar la consideración de la rendición de cuentas, la explicabilidad, la equidad, la interpretabilidad y la seguridad de estos sistemas. Estos diferentes dominios de la IA ética están estrechamente relacionados y pueden abordarse considerando la transparencia del sistema. La transparencia es un subconjunto de la IA ética que implica poner toda la información, las decisiones, los procesos de toma de decisiones y los supuestos a disposición de las partes interesadas, lo que a su vez mejora su comprensión de los sistemas de IA y los resultados relacionados (Chaudhry et al., 2022). Además, para las innovaciones basadas en LLM, Weidinger et al. (2021) sugirieron seis tipos de riesgos éticos: 1) discriminación, exclusión y toxicidad; 2) riesgos de información; 3) daños por desinformación; 4) usos maliciosos; 5) daños por interacción persona-computadora; y 6) daños por automatización, acceso y entorno. Estos riesgos pueden agruparse en tres cuestiones éticas fundamentales: la privacidad de los datos personales de los actores educativos; la igualdad en la accesibilidad de actores con diferentes orígenes; y la beneficencia en los posibles daños e impactos negativos que las innovaciones basadas en LLMs pueden tener en los actores (Ferguson et al., 2016). Estas tres cuestiones éticas fundamentales se consideraron en el análisis de la literatura revisada. Se ofrecen más detalles en la Sección 3.2.

2.3 | Trabajos relacionados

Las revisiones sistemáticas previas se han centrado principalmente en la revisión de un escenario de aplicación específico (p. ej., generación de preguntas, retroalimentación automatizada, chatbots y calificación de ensayos) del procesamiento del lenguaje natural y los LLM. Por ejemplo, Kurdi et al. (2020) revisaron sistemáticamente estudios empíricos que buscaban abordar el problema de la generación automática de preguntas en el ámbito educativo. Resumieron exhaustivamente los diferentes métodos de generación, tareas de generación y métodos de evaluación presentados en la literatura previa. En particular, los LLM podrían beneficiarse de los enfoques semánticos para generar preguntas significativas estrechamente relacionadas con el contenido original. Asimismo, Cavalcanti

et al. (2021) han revisado sistemáticamente diferentes sistemas automatizados de retroalimentación respecto a su impacto en la mejora del rendimiento académico de los estudiantes y la reducción de la carga de trabajo del profesorado. A pesar de que la mitad de los estudios revisados no muestran evidencia de una reducción de la carga de trabajo del profesorado, dado que estos sistemas automatizados de retroalimentación se basaban principalmente en reglas y requerían un gran esfuerzo manual, identificaron que el uso de técnicas de generación de lenguaje natural podría mejorar aún más la generalización de dichos sistemas y potencialmente reducir la carga de trabajo manual. Por otro lado, Wollny et al. (2021) han revisado sistemáticamente áreas de la educación donde ya se han aplicado chatbots. Concluyeron que aún queda mucho por hacer para que los chatbots alcancen su máximo potencial, como hacerlos más adaptables a diferentes contextos educativos. Una revisión sistemática también ha investigado los diversos sistemas automatizados de calificación de ensayos (Ramesh y Sanampudi, 2022). Los hallazgos han revelado múltiples limitaciones de los sistemas existentes basados en el aprendizaje automático tradicional (p. ej., regresión y bosque aleatorio) y algoritmos de aprendizaje profundo (p. ej., LSTM y BERT). En resumen, estas revisiones sistemáticas previas han identificado áreas de mejora que podrían abordarse mediante el uso de programas de maestría en derecho (LLM) de vanguardia (p. ej., GPT-3 o Codex). Sin embargo, ninguna de estas revisiones sistemáticas ha investigado los aspectos prácticos y éticos relacionados con las innovaciones basadas en LLM en educación en general, sino más bien en particular (p. ej., limitadas a una tarea específica).

El reciente revuelo en torno a una de las últimas innovaciones basadas en LLMs, ChatGPT, ha intensificado el debate sobre los retos prácticos y éticos relacionados con el uso de LLMs en educación. Por ejemplo, en un documento de posición, Kasneci et al. (2023) proporcionaron una visión general de algunas investigaciones existentes sobre LLMs y propusieron varias oportunidades y retos prácticos de los LLMs desde las perspectivas de estudiantes y profesores. Asimismo, Rudolph et al. (2023) también proporcionaron una visión general de los posibles impactos, retos y oportunidades que ChatGPT podría tener en las futuras prácticas educativas. Aunque estos estudios no han revisado sistemáticamente la literatura educativa existente sobre LLMs, sus argumentos resonaron con algunos de los problemas apremiantes en torno a LLMs y la IA ética, como la privacidad de los datos, el sesgo y los riesgos. Por otro lado, Sallam (2023) revisó sistemáticamente las implicaciones y limitaciones de ChatGPT en la educación sanitaria e identificó su posible utilidad en torno a la personalización y la automatización. Sin embargo, cabe destacar que la mayoría de los artículos revisados en el estudio de Sallam eran editoriales, comentarios o preimpresiones. Esta falta de estudios empíricos revisados por pares sobre ChatGPT es comprensible, ya que se publicó a finales de 2022 (OpenAI, 2023). Ningún trabajo existente ha revisado sistemáticamente la literatura revisada por pares sobre innovaciones previas basadas en LLM. Dichas investigaciones podrían proporcionar evidencia más fiable y empírica sobre las posibles oportunidades y desafíos de los LLM en las prácticas educativas. Por lo tanto, el presente estudio se propuso abordar esta brecha en la literatura mediante una revisión sistemática de la investigación educativa previa sobre LLM. En concreto, se investigaron las siguientes preguntas de investigación para guiar esta revisión:

- **Pregunta 1**: Qué es *El estado actual de la investigación* sobre el uso de LLM para automatizar tareas educativas, específicamente desde la perspectiva de las tareas educativas, las partes interesadas, los LLM y las tareas de aprendizaje automático[1]?
- **Pregunta 2**: ¿Cuáles son las *práctico* ¿Cuáles son los desafíos de los LLM en la automatización de tareas educativas, específicamente a través de la lente de la preparación tecnológica, el desempeño del modelo y la replicabilidad del modelo?
- **Pregunta 3**: ¿Cuáles son las *ético* ¿Cuáles son los desafíos de los LLM en la automatización de tareas educativas, específicamente a través de la lente de la transparencia del sistema, la privacidad, la igualdad y la beneficencia?

[1] Como clasificación, predicción, agrupamiento, etc.

3 | MÉTODOS

En este estudio se realizó una revisión sistemática de alcance, ya que este método se ha utilizado con frecuencia en áreas de investigación emergentes y en rápida evolución para analizar la literatura e identificar los conceptos, métodos, evidencia y desafíos clave (Munn et al., 2018). Por consiguiente, a menudo no se evaluó la calidad de los estudios incluidos, ya que el objetivo es ofrecer una visión global de un campo emergente.

3.1 | Procedimientos de revisión

Seguimos el protocolo PRISMA (Page et al., 2021) para llevar a cabo la revisión sistemática de alcance de los LLM. Buscamos en cuatro bases de datos bibliográficas de prestigio, como Scopus, ACM Digital Library, IEEE Xplore y Web of Science, para encontrar publicaciones de alta calidad con revisión por pares. Se realizaron búsquedas adicionales a través de Google Académico y el Centro de Información de Recursos Educativos (ERIC) para identificar publicaciones con revisión por pares que aún no se han indexado en estas bases de datos, ya sean publicadas recientemente o no (p. ej., Journal of Educational Data Mining; antes de 2020). Nuestra búsqueda inicial para el título, el resumen y las palabras clave incluyó términos como "modelo de lenguaje grande", "modelo de lenguaje pre*entrenado", "GPT-*", "BERT", "educación", "estudiante*" y "profesor*". También se aplicó una restricción por año de publicación para restringir la búsqueda a estudios publicados desde 2017, concretamente del 01/01/2017 al 31/12/2022, dado que la arquitectura fundamental (Transformer) de los LLM se publicó oficialmente en 2017 (Vaswani et al., 2017). Solo se consideraron publicaciones con revisión por pares para reforzar la credibilidad científica de esta revisión. La búsqueda inicial en la base de datos fue realizada por dos investigadores de forma independiente. Cualquier discrepancia entre los resultados de la búsqueda se resolvió mediante un diálogo posterior o consultando al bibliotecario para obtener orientación.

Dos investigadores revisaron de forma independiente los títulos y resúmenes de los artículos elegibles según cinco criterios de inclusión y exclusión predeterminados. En primer lugar, incluimos estudios que utilizaban modelos lingüísticos grandes o preentrenados directamente o basados en dichos modelos, y excluimos los estudios que utilizaban modelos generales de aprendizaje automático o aprendizaje profundo con un uso no especificado de LLM. En segundo lugar, incluimos estudios empíricos con metodologías detalladas, como una descripción detallada de los LLM y los procedimientos de investigación, y excluimos trabajos de revisión, opinión y alcance. En tercer lugar, solo incluimos artículos completos revisados por pares y excluimos artículos cortos, de taller y póster de menos de seis y ocho páginas para diseños de doble y una sola columna, respectivamente. Además, incluimos estudios que utilizaban LLM con el fin de automatizar tareas educativas (p. ej., calificación de ensayos y generación de preguntas), y excluimos los estudios que simplemente utilizaban LLM como parte del análisis sin implicaciones educativas. Finalmente, solo incluimos estudios publicados en inglés (tanto el resumen como el texto principal) y excluimos los estudios publicados en otros idiomas. Cualquier decisión conflictiva se resolvió mediante un debate más profundo entre los dos investigadores o consultando a un tercer investigador para lograr un consenso.

La búsqueda en la base de datos arrojó inicialmente 854 publicaciones, con 191 duplicados eliminados, lo que resultó en 663 publicaciones para la revisión de título y resumen (véase la Figura 1). Tras la revisión de título y resumen, se incluyeron 197 artículos para la revisión de texto completo, con una fiabilidad interevaluador (kappa de Cohen) de 0,75, lo que indica una concordancia sustancial entre los revisores durante la revisión de título y resumen. Se seleccionaron 118 artículos para la extracción de datos tras la revisión de texto completo, con una fiabilidad interevaluador (kappa de Cohen) de 0,73, lo que indica una concordancia sustancial entre los revisores durante la revisión de texto completo. De los 197 artículos iniciales, 79 fueron excluidos por diversas razones, entre ellas, no ser un artículo completo (n = 41), falta de automatización educativa (n = 17), falta de capacitación previa o LLM (n = 12), simplemente usar capacitación previa o LLM como parte del análisis (n = 3), artículo en un idioma diferente al inglés (n = 2) y artículo no empírico (n = 2).

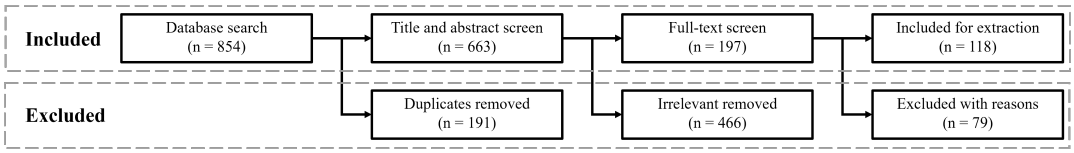

FIGURA 1Proceso de revisión sistemática del alcance siguiendo el protocolo PRISMA.

3.2|Análisis de datos

Para la primera pregunta de investigación (RQ1), realizamos un análisis temático inductivo para extraer información sobre el estado actual de la investigación en el uso de LLMs para automatizar tareas educativas. Específicamente, extrajimos cuatro tipos principales de información contextual de cada artículo incluido: tareas educativas, partes interesadas, LLMs y tareas de aprendizaje automático. Esta información contextual proporcionaría una visión holística de la investigación existente e informaría a investigadores y profesionales sobre las direcciones viables para explorar con los LLMs de última generación (por ejemplo, GPT-3.5 y Codex). Se desarrolló un total de siete elementos de extracción de datos para abordar las preguntas de investigación segunda y tercera. Estos elementos se desarrollaron porque están directamente relacionados con la definición de practicidad (RQ2: Ítems 1-3) y ética (RQ3: Ítems 4-7), como se define en la sección Antecedentes (Sección 2). La siguiente lista detalla el conjunto final de elementos junto con las preguntas guía correspondientes. Para el análisis temático y los elementos, dos investigadores codificaron de forma independiente 20 muestras aleatorias de los estudios incluidos. Cualquier conflicto se resolvió mediante discusión adicional o consulta con un tercer investigador. Tras alcanzar un índice kappa de Cohen superior a 0,80 (lo que indica una concordancia casi perfecta), cada investigador codificó la mitad de los 98 estudios restantes (49 estudios cada uno) y contrastó el trabajo de los demás. La base de datos de los estudios incluidos en esta revisión y los datos extraídos para cada ítem están disponibles en el documento complementario.

1. **Preparación tecnológica**¿En qué niveles de preparación tecnológica se encuentran las innovaciones basadas en LLM? Adoptamos la herramienta de evaluación del gobierno australiano, concretamente los Niveles de Preparación Tecnológica (TRL) (Ciencia y Grupo) del Departamento de Defensa de Australia, que se han utilizado para evaluar la madurez de las tecnologías educativas en SLR anteriores (Yan et al., 2022). Existen nueve niveles de preparación tecnológica diferentes: Investigación básica (TRL-1), Investigación aplicada (TRL-2), Función crítica o prueba de concepto establecida (TRL-3), Pruebas de laboratorio/Validación de componente/proceso prototipo alfa (TRL-4), Pruebas de laboratorio de sistema integrado/semiintegrado (TRL-5), Sistema prototipo verificado (TRL-6), Sistema piloto integrado demostrado (TRL-7), Sistema incorporado en diseño comercial (TRL-8) y Sistema probado y listo para su implementación comercial completa (TRL-9), que se explican con más detalle en la sección de Resultados.
2. **Rendimiento**¿Cuán precisas y confiables son las innovaciones basadas en LLM para completar las tareas educativas designadas? Por ejemplo, ¿cuáles son los puntajes de rendimiento del modelo para la clasificación (p. ej., puntajes AUC y F1), la generación (p. ej., puntaje BLEU) y las tareas de predicción (p. ej., RMSE y correlación de Pearson)?
3. **Replicabilidad**¿Pueden otros investigadores o profesionales replicar las innovaciones basadas en los LLM sin el apoyo adicional de los autores originales? Este punto evalúa si el artículo proporcionó suficientes detalles sobre los LLM (p. ej., algoritmos de código abierto) y el conjunto de datos (p. ej., datos de código abierto).
4. **Transparencia**¿En qué niveles del índice de transparencia (Chaudhry et al., 2022) se encuentran las innovaciones basadas en LLM? El índice de transparencia propuso tres niveles: transparencia para investigadores y profesionales de la IA (Nivel 1), transparencia para expertos y entusiastas de la tecnología educativa (Nivel 2) y transparencia para educadores y padres (Nivel 3). El nivel de transparencia aumenta a medida que los actores educativos participan plenamente en el desarrollo y la evaluación del sistema de IA. Estos niveles se detallaron en la sección de Resultados.

5. Privacidad¿Se han mencionado o considerado en el artículo los problemas de privacidad de sus innovaciones? Este artículo explora posibles problemas relacionados con el consentimiento informado, la recopilación transparente de datos, el control individual sobre los datos personales y la vigilancia no intencionada (Ferguson et al., 2016; Tsai et al., 2020).

6. Igualdad¿Se ha mencionado o considerado en el artículo la igualdad de acceso a sus innovaciones? Este punto explora posibles problemas relacionados con el acceso limitado para estudiantes de bajos recursos o de zonas rurales, así como la limitación lingüística de las innovaciones, como su capacidad para analizar diferentes idiomas (Ferguson et al., 2016).

7. Beneficencia¿Se han mencionado o considerado en el artículo posibles problemas que vulneren el principio ético de beneficencia? Dichas vulneraciones pueden incluir los riesgos asociados al etiquetado y la elaboración de perfiles de estudiantes, el uso inadecuado de contenido generado por máquinas para las evaluaciones y los sesgos algorítmicos (Ferguson et al., 2016; Zawacki-Richter et al., 2019).

4 | RESULTADOS

4.1 | El estado actual — RQ1

Identificamos nueve categorías diferentes de tareas educativas que estudios previos han intentado automatizar utilizando LLMs (como se muestra en la Tabla 1). Estudios previos han utilizado LLMs para automatizar el perfilado y etiquetado de 17 tipos de contenidos y conceptos relacionados con la educación (p. ej., publicaciones en foros, sentimiento de los estudiantes y similitud de disciplinas), la detección de seis constructos latentes (p. ej., confusión y urgencia), la calificación de cinco tipos de evaluaciones (p. ej., preguntas de respuesta corta y ensayos), el desarrollo de cinco tipos de apoyo a la enseñanza (p. ej., agente de conversación y respuesta inteligente a preguntas), la predicción de cinco tipos de métricas orientadas al estudiante (p. ej., abandono y participación), la construcción de cuatro tipos de representaciones de conocimiento (p. ej., gráfico de conocimiento y reconocimiento de entidades), la provisión de cuatro formas diferentes de retroalimentación (p. ej., retroalimentación en tiempo real y post-hoc), la generación de cuatro tipos de contenido (p. ej., preguntas de opción múltiple y preguntas abiertas) y la entrega de tres tipos de recomendaciones (p. ej., recurso y curso). De los 118 estudios revisados, 85 estudios apuntaron a automatizar tareas educativas relacionadas con los docentes (por ejemplo, calificación y generación de preguntas), 54 estudios apuntaron a actividades relacionadas con los estudiantes (por ejemplo, retroalimentación y recomendación de recursos), 20 estudios se centraron en apoyar las prácticas institucionales (por ejemplo, recomendaciones de cursos y planificación de disciplinas) y 14 estudios empoderaron a los investigadores con métodos automatizados para investigar constructos latentes (por ejemplo, confusión de estudiantes) y capturar datos verbales (por ejemplo, reconocimiento de voz).

Identificamos cinco categorías de LLM utilizadas en estudios previos para automatizar tareas educativas. BERT y sus variantes (p. ej., RoBERTa, DistilBERT, BERT multilingüe, LaBSE, EstBERT y Sentence-BERT) fueron el modelo predominante utilizado en 109 estudios revisados. Sin embargo, a menudo requerían esfuerzo manual para su ajuste (n = 90). GPT-2 y GPT-3 se han utilizado en cinco y tres estudios, respectivamente. Específicamente, GPT-2 y GPT-3 han tenido un mejor rendimiento que los modelos basados en BERT en tareas de generación y evaluación de contenido, como la generación de problemas de matemáticas universitarios (Drori et al., 2022) y la evaluación de la calidad de las preguntas de respuesta corta generadas por los estudiantes (Moore et al., 2022). Codex de OpenAI se ha utilizado en dos estudios previos, específicamente para tareas de generación de código. T5 también se ha utilizado en dos estudios previos para fines de clasificación y generación. En términos de tareas de aprendizaje automático, 74 estudios utilizaron LLM para realizar tareas de clasificación. Las tareas de generación y predicción se investigaron en 24 y 23 estudios previos, respectivamente. En resumen, las innovaciones basadas en LLM ya se han utilizado para automatizar diversas tareas educativas, pero la mayoría de estas innovaciones se desarrollaron en modelos más antiguos, como BERT y GPT-2. Si bien los modelos de vanguardia, como GPT-3, se han introducido hace más de dos años (Brown et al., 2020), aún no se han aplicado ampliamente para automatizar tareas educativas. Una posible razón para esta falta de adopción podría ser la naturaleza comercial y de código cerrado de estos modelos, lo que aumenta la carga financiera del desarrollo y la operación de innovaciones en tecnología educativa, además de...

tales modelos.

TABLA 1 Tareas educativas en la investigación de maestrías en derecho

Categorías	Tareas educativas
Perfilado y etiquetado	Clasificación de publicaciones en foros, clasificación de actos de diálogo, clasificación de diseños de aprendizaje, análisis de opiniones de revisión, modelado de temas, clasificación pedagógica de MOOC, modelado de resolución colaborativa de problemas, calidad de paráfrasis, etiquetado de discursos, etiquetado de contenido educativo con componentes de conocimiento, extracción de frases clave y palabras clave, análisis de escritura reflexiva, pensamiento representativo multimodal, similitud de disciplinas, clasificación de conceptos, clasificación de niveles cognitivos, segmentación de argumentos de ensayo
Detección	Análisis semántico, detección de mensajes fuera de la tarea, detección de confusión, detección de urgencia, detección de intención conversacional, detección del comportamiento de los docentes
Evaluación y calificación	Calificación de evaluación formativa y sumativa, calificación de respuestas cortas, calificación de ensayos, calificación de preguntas subjetivas, autoexplicación del estudiante
Apoyo a la enseñanza	Enseñanza en el aula, apoyo a la comunidad de aprendizaje, agente de conversación de aprendizaje en línea, preguntas y respuestas inteligentes, reconocimiento de la actividad docente.
Predicción	Predicción del rendimiento estudiantil, predicción de la deserción estudiantil, detección del compromiso emocional y cognitivo, indicadores de crecimiento y desarrollo de estudiantes universitarios, identificación de estudiantes en riesgo.
Representación del conocimiento cción	Construcción de gráficos de conocimiento, reconocimiento de entidades de conocimiento, rastreo de conocimiento, extracción de relaciones causa-efecto
Comentario	Retroalimentación en tiempo real, retroalimentación post-hoc, retroalimentación agregada, retroalimentación sobre la retroalimentación (comentarios de revisión por pares)
Generación de contenido	Generación de preguntas de opción múltiple, generación de preguntas abiertas, generación de código, generación de respuestas (lenguaje natural)
Recomendación	Selección y recomendación de referencias en inglés, recomendación de recursos, recomendación de cursos

4.2 | Desafíos prácticos — RQ2

4.2.1 | Preparación tecnológica

Según la escala de Nivel de Preparación Tecnológica (Ciencia y Grupo), las innovaciones basadas en LLM aún se encuentran en la etapa inicial de desarrollo y prueba. Más de tres cuartas partes de los estudios de LLM (n = 89) se encuentran en la etapa de investigación aplicada (TRL-2), cuyo objetivo es experimentar con la capacidad de los LLM para automatizar diferentes tareas educativas mediante el desarrollo de diferentes modelos y la combinación de LLM con otras técnicas de aprendizaje automático y aprendizaje profundo (p. ej., RCNN [Shang et al., 2022]). Trece estudios han establecido una prueba de concepto y demostrado la viabilidad de utilizar innovaciones basadas en LLM para automatizar ciertos procesos de tareas educativas (TRL-3). Nueve estudios han desarrollado prototipos funcionales y realizado una validación preliminar en entornos de laboratorio controlados (TRL-4), que a menudo implica...

ing stakeholders (p. ej., estudiantes y docentes) para probar y evaluar el resultado de sus innovaciones. Solo siete estudios han dado un paso más y han realizado estudios de validación en entornos de aprendizaje auténticos, con la mayoría de los componentes funcionales integrados en las tareas educativas (TRL-5), como un paciente virtual inteligente estándar para la formación de estudiantes de medicina (Song et al., 2022) y un chatbot inteligente para la admisión universitaria (Nguyen et al., 2021). Sin embargo, ninguna de las innovaciones existentes basadas en LLM se ha verificado mediante operaciones exitosas (TRL-6). En conjunto, estos hallazgos sugieren que, si bien las innovaciones existentes basadas en LLM pueden utilizarse para automatizar ciertas tareas educativas, aún no han mostrado evidencia de mejoras en los procesos de enseñanza, aprendizaje y administración en prácticas educativas auténticas.

4.2.2 | Actuación

El rendimiento de las innovaciones basadas en LLM varía según las diferentes tareas educativas y de aprendizaje automático. En cuanto a las tareas de clasificación, las innovaciones basadas en LLM han mostrado un alto rendimiento en tareas educativas sencillas, como modelar los temas a partir de una lista de tareas de programación (mejor F1 = 0,95) (Fonseca et al., 2020), analizar la opinión de los estudiantes (mejor F1 = 0,94) (Truong et al., 2020), construir un grafo de conocimiento de la materia a partir de materiales didácticos (mejor F1 = 0,94) (Su y Zhang, 2020) y clasificar publicaciones en foros educativos (Sha et al., 2022c) (mejor F1 = 0,92). Sin embargo, el rendimiento de clasificación de las innovaciones basadas en LLM disminuye en otras tareas educativas. Por ejemplo, las puntuaciones F1 para detectar la confusión de los estudiantes en el foro del curso (Geller et al., 2021) y los mensajes fuera de la tarea de los estudiantes en el aprendizaje colaborativo basado en juegos (Carpenter et al., 2020) rondan 0,77 y 0,67, respectivamente. Asimismo, la puntuación F1 para clasificar respuestas cortas varía entre 0,61 y 0,82, con el menor rendimiento en preguntas fuera de la muestra (mejor F1 = 0,61) (Condor et al., 2021). También se observaron resultados similares en la clasificación de ensayos argumentativos de los estudiantes (mejor F1 = 0,66) (Ghosh et al., 2020).

Para tareas de predicción, las innovaciones basadas en LLM han demostrado un rendimiento fiable en comparación con la verdad fundamental o con evaluadores humanos. Por ejemplo, las innovaciones basadas en LLM han alcanzado puntuaciones altas de kappa ponderada cuadrática (QWK) en la calificación de ensayos, específicamente para respuestas fuera de tema (QWK = 0,80), sin sentido (QWK = 0,80) y parafraseadas (QWK = 0,94), lo que indica una concordancia sustancial o casi perfecta con los evaluadores humanos (Doewes y Pechenizkiy, 2021). Se han observado resultados similares en la calificación de ensayos en varios otros estudios (p. ej., 0,80 QWK en Beseiso et al., 2021 y 0,81 QWK en Sharma et al., 2021). Asimismo, los resultados de las innovaciones basadas en LLM en la calificación automática de respuestas cortas también estuvieron altamente correlacionados con las calificaciones humanas (correlación de Pearson entre 0,75 y 0,82) (Ahmed et al., 2022; Sawatzki et al., 2022).

En cuanto a las tareas de generación, las innovaciones basadas en LLM demostraron un alto rendimiento en diferentes tareas educativas. Por ejemplo, estas innovaciones alcanzaron una puntuación F1 de 0,92 en la generación de preguntas de opción múltiple con respuestas de una sola palabra (Kumar et al., 2022). Las tecnologías educativas desarrolladas mediante el perfeccionamiento de Codex también demostraron la capacidad de resolver el 81 % de los problemas matemáticos avanzados (Drori et al., 2022). Los resúmenes de texto generados con BERT no presentaron diferencias significativas en comparación con los generados por los estudiantes, y los estudiantes de posgrado no pueden diferenciarlos (Merine y Purkayastha, 2022). De igual manera, los diálogos médico-paciente generados con BERT también resultaron indistinguibles de los diálogos reales, lo que permite crear pacientes virtuales estándar para la formación en prácticas diagnósticas de los estudiantes de medicina (Song et al., 2022). Además, para los cursos introductorios de programación, los LLM de última generación, Codex, podrían generar ejercicios sensatos y novedosos para los estudiantes junto con una solución de muestra apropiada (alrededor de tres de cada cuatro veces) y una explicación precisa del código (67 % de precisión) (Sarsa et al., 2022).

En resumen, aunque el desempeño de clasificación de las innovaciones basadas en LLM en tareas educativas complejas está lejos de ser adecuado para su adopción práctica, las innovaciones basadas en LLM ya han demostrado un alto desempeño en varios aspectos relacionados.

Tareas de clasificación relativamente sencillas que podrían implementarse para generar automáticamente información significativa útil para docentes e instituciones, como la gestión de numerosos comentarios de estudiantes y la revisión de cursos. Asimismo, el rendimiento de predicción y generación de innovaciones basadas en LLM revela un futuro prometedor para la posible automatización de la generación de contenido educativo y la calificación inicial de las evaluaciones de los estudiantes. Sin embargo, deben considerarse cuestiones éticas para tales implementaciones, las cuales abordamos en los hallazgos de la RQ3.

4.2.3 | Replicabilidad

La mayoría de los estudios revisados (n=107) no han revelado suficientes detalles sobre sus metodologías para que otros investigadores y profesionales repliquen sus innovaciones propuestas basadas en LLMs. Entre estos estudios, 12 estudios han publicado el código original para desarrollar las innovaciones, pero no han publicado los datos que utilizaron. Por el contrario, 20 estudios han publicado los datos que utilizaron, pero no han publicado el código real. Alrededor de dos tercios de los estudios revisados (n=75) no han publicado ni el código original ni los datos que utilizaron, dejando solo 11 estudios disponibles públicamente para que otros investigadores y profesionales los repliquen sin necesidad de contactar a los autores originales. Esta falta de replicabilidad podría convertirse en una barrera vital para la adopción, ya que 87 de los 107 estudios no replicables requirieron el ajuste de los LLMs para lograr el rendimiento informado. Este problema de replicación también impide que otros evalúen más a fondo la generalización de las innovaciones propuestas basadas en LLMs en otros conjuntos de datos, lo que limita las posibles utilidades prácticas.

4.3 | Desafíos éticos — RQ3

4.3.1 | Transparencia

Según el índice de transparencia y los tres niveles de transparencia (Chaudhry et al., 2022), la mayoría de los estudios revisados alcanzaron como máximo el Nivel 1 (n=109), que se considera transparente únicamente para investigadores y profesionales de la IA. Si bien estos estudios informaron detalles sobre sus modelos de aprendizaje automático (p. ej., optimización e hiperparámetros), es improbable que dicha información sea interpretable y considerada transparente para personas sin una sólida formación en aprendizaje automático. Los nueve estudios restantes alcanzaron como máximo el Nivel 2, ya que a menudo incluían algún tipo de intervención humana. En concreto, tres estudios han demostrado que las innovaciones de los LLM están disponibles para la evaluación de los estudiantes (Nguyen et al., 2021; Song et al., 2022; Merine y Purkayastha, 2022). Dichas evaluaciones a menudo implicaban que los estudiantes diferenciaran el contenido generado por IA del generado por humanos (Song et al., 2022; Merine y Purkayastha, 2022) y evaluaran su satisfacción con las respuestas generadas por IA (Nguyen et al., 2021). Asimismo, dos estudios han involucrado a expertos en la evaluación de características específicas del contenido generado por las innovaciones basadas en LLM, como el nivel de información (Maheen et al., 2022) y el nivel cognitivo (Moore et al., 2022). Se han utilizado encuestas para evaluar la experiencia de los estudiantes con las innovaciones basadas en LLM desde múltiples perspectivas, como la calidad y la dificultad de las preguntas generadas por IA (Drori et al., 2022; Li y Xing, 2021) y los posibles beneficios de aprendizaje de los sistemas (Jayaraman y Black, 2022). Finalmente, se realizaron entrevistas semiestructuradas para comprender la percepción de los estudiantes sobre el sistema LLM tras su uso en actividades auténticas de aprendizaje colaborativo con soporte informático (Zheng et al., 2022). Si bien estos nueve estudios incluyeron algunos elementos de participación humana, las partes interesadas a menudo participaron en una evaluación posterior en lugar de durante todo el proceso de desarrollo y, por lo tanto, tienen un conocimiento limitado sobre el principio operativo y las posibles debilidades de los sistemas. En consecuencia, ninguna de las innovaciones existentes basadas en LLM puede considerarse de Nivel 3, que describe un sistema de IA transparente para las partes interesadas educativas (p. ej., estudiantes, docentes y padres).

4.3.2 | Privacidad

Los problemas de privacidad relacionados con las innovaciones basadas en LLM rara vez se abordaron o investigaron en los estudios revisados. En concreto, en los estudios que han perfeccionado LLM con datos textuales recopilados de estudiantes, ninguno ha explicado explícitamente sus estrategias de consentimiento (p. ej., si los estudiantes reconocen la recopilación y el uso previsto de sus datos) ni las medidas de protección de datos (p. ej., anonimización y desinfección de datos). Esta falta de atención a la privacidad es especialmente preocupante, ya que las innovaciones basadas en LLM utilizan el lenguaje natural de las partes interesadas, que puede contener información personal y sensible sobre su vida privada e identidad (Brown et al., 2022). Es posible que las partes interesadas desconozcan que sus datos textuales (p. ej., publicaciones o conversaciones en foros) en plataformas digitales (p. ej., MOOC y LMS) se utilizan en innovaciones basadas en LLM para diferentes fines de automatización (p. ej., respuestas automatizadas y chatbots de formación), ya que el proceso de consentimiento suele estar integrado en la inscripción o registro en estas plataformas (Tsai y Gasevic, 2017). Este proceso difícilmente puede considerarse consentimiento informado. En consecuencia, si las partes interesadas compartieran su información personal en estas plataformas en lenguaje natural (p. ej., compartiendo números de teléfono y direcciones con miembros del grupo a través de foros digitales), dicha información podría utilizarse como datos de entrenamiento para el perfeccionamiento de los LLM. Este uso podría exponer información privada, ya que los LLM son incapaces de comprender el contexto y la sensibilidad del texto y, por lo tanto, podrían devolver la información personal de las partes interesadas basándose en relaciones semánticas (Brown et al., 2022).

4.3.3 | Igualdad

Aunque la mayoría de los estudios (n=95) utilizaron LLMs que solo aplican a contenido en inglés, también identificamos escenarios de aplicación de LLMs en la automatización de tareas educativas en otros 12 idiomas. Específicamente, 19 estudios utilizaron LLMs que pueden aplicarse a contenido chino. Diez estudios previos utilizaron LLMs para contenidos en vietnamita (n=3), español (n=3), italiano (n=2) y alemán (n=2). Además, siete estudios aplicaron LLMs a contenido en croata, indonesio, japonés, rumano, ruso, sueco e hindi. Si bien el predominio de innovaciones basadas en inglés sigue siendo un problema de igualdad preocupante, la disponibilidad de innovaciones que admiten una variedad de otros idiomas, específicamente en sociedades no occidentales, educadas, industrializadas, ricas y democráticas (WEIRD) (por ejemplo, Indonesia y Vietnam), puede indicar una señal prometedora de que las innovaciones basadas en LLMs tienen posibles impactos globales y niveles tales problemas de igualdad en el futuro. Sin embargo, las cargas financieras que implica adoptar modelos de última generación (por ejemplo, GPT-3 y Codex de OpenAI) podrían potencialmente exacerbar los problemas de igualdad, haciendo que las innovaciones de mejor desempeño solo sean accesibles y asequibles para las sociedades WEIRD.

4.3.4 | Beneficencia

Un total de siete estudios han analizado posibles problemas relacionados con la violación del principio ético de beneficencia. Por ejemplo, un estudio analizó el riesgo potencial de adoptar modelos de bajo rendimiento, lo que podría afectar negativamente las experiencias de aprendizaje de los estudiantes (Li y Xing, 2021). Estos problemas podrían minimizarse aplazando las decisiones tomadas por dichos modelos (Schneider et al., 2022) y etiquetando el contenido generado por IA con un mensaje de advertencia (p. ej., la revisión del manual del profesorado es obligatoria antes de determinar la corrección real) (Angelone et al., 2022). Además de los problemas que conlleva la adopción de modelos inexactos, dos estudios han sugerido que pueden surgir posibles problemas de sesgo y discriminación si se adopta un modelo preciso pero injusto (Sha et al., 2021; Merine y Purkayastha, 2022). Este problema es particularmente preocupante, ya que la mayoría de los estudios existentes se centraron exclusivamente en el desarrollo de un modelo preciso. Solo nueve estudios revisados publicaron información sobre los datos descriptivos de diferentes grupos de muestra, como el género y la etnia.

(p. ej., (Pugh et al., 2021)). Dos estudios han propuesto posibles enfoques que podrían abordar estos problemas de equidad. Específicamente, el uso de estrategias de muestreo, como equilibrar la distribución demográfica, se ha encontrado como un enfoque eficaz para mejorar tanto la equidad como la precisión del modelo (Sha et al., 2022b,a). Estos enfoques son esenciales para garantizar que las innovaciones basadas en LLM no perpetúen sesgos problemáticos y sistemáticos (p. ej., sesgos de género), especialmente porque los LLM con mejor desempeño a menudo están en caja negra con poca interpretabilidad, trazabilidad y justificación de los resultados (Wu, 2022).

5 | DISCUSIÓN

5.1 | Principales hallazgos

El estudio actual revisó sistemáticamente 118 estudios empíricos revisados por pares que utilizaron LLMs para automatizar tareas educativas. Para la primera pregunta de investigación (RQ1), ilustramos el estado actual de la investigación educativa sobre LLMs. Específicamente, identificamos 53 tipos de escenarios de aplicación de LLMs en la automatización de tareas educativas, resumidos en nueve categorías generales, que incluyen perfilado y etiquetado, detección, evaluación y calificación, apoyo a la enseñanza, predicción, representación del conocimiento, retroalimentación, generación de contenido y recomendación. Si bien algunas de estas categorías resuenan con las utilidades propuestas en trabajos de posicionamiento previos (p. ej., retroalimentación, generación de contenido y recomendación) (Kasneci et al., 2023; Rudolph et al., 2023), direcciones novedosas como el uso de LLMs para automatizar la creación de grafos y entidades de conocimiento indicaron aún más el potencial de las innovaciones basadas en LLMs para respaldar las prácticas institucionales (p. ej., creación de motores de búsqueda basados en el conocimiento en múltiples disciplinas). Estas direcciones identificadas podrían beneficiarse de los LLM de última generación (p. ej., GPT-3 y Codex), ya que la mayoría de los estudios revisados (92 %) se centraron en el uso de modelos basados en BERT, que a menudo requerían esfuerzo manual para su ajuste. Mientras que, los LLM de última generación podrían lograr un rendimiento similar con un enfoque de cero disparos (Bang et al., 2023). Si bien la mayoría de los estudios revisados (63 %) se centraron en el uso de LLM para automatizar tareas de clasificación, podría haber más estudios futuros que apuntaran a abordar la automatización de tareas de predicción y generación con los LLM más capaces (Sallam, 2023). Del mismo modo, aunque el apoyo a los docentes es el enfoque principal (72 %) de las innovaciones existentes basadas en LLM, los estudiantes y las instituciones también podrían beneficiarse de dichas innovaciones, ya que podrían seguir surgiendo nuevas utilidades de la literatura sobre tecnología educativa. En conjunto, los hallazgos de la primera pregunta de investigación podrían inspirar a los investigadores educativos con ideas para explorar el potencial de los LLM de última generación para ampliar las prácticas educativas; específicamente, los 53 tipos de escenarios de aplicación identificados pueden valer la pena volver a explorar a la luz de ChatGPT y otros modelos potentes de IA generativa (Kasneci et al., 2023).

En cuanto a la segunda pregunta de investigación (PI2), identificamos varios desafíos prácticos que deben abordarse para que las innovaciones basadas en LLM generen beneficios educativos reales. El desarrollo y la investigación educativa sobre innovaciones basadas en LLM aún se encuentran en sus etapas iniciales. La mayoría de las innovaciones demostraron un bajo nivel de preparación tecnológica, por lo que aún no se han integrado ni validado plenamente en contextos educativos reales. Este hallazgo coincide con revisiones sistemáticas previas sobre tecnologías educativas relacionadas, como las revisiones sobre generación automatizada de preguntas (Kurdi et al., 2020), retroalimentación (Cavalcanti et al., 2021), calificación de ensayos (Ramesh y Sanampudi, 2022) y sistemas de chatbots (Wollny et al., 2021). Existe una necesidad apremiante de estudios prácticos que proporcionen innovaciones basadas en LLM directamente a los actores educativos para apoyar tareas educativas reales, en lugar de realizar pruebas con diferentes conjuntos de datos o en entornos de laboratorio. Estos estudios auténticos también podrían validar si las innovaciones existentes pueden alcanzar el alto rendimiento del modelo reportado en escenarios reales, específicamente en tareas de predicción y generación, en lugar de limitarse a conjuntos de datos previos. Este proceso de validación es vital para prevenir el uso inadecuado, como la adopción de un modelo de predicción específico para sujetos no previstos. Los investigadores deben ser cuidadosos.

Examinar el grado de generalización de sus innovaciones e informar a las partes interesadas sobre sus limitaciones (Gašević et al., 2016). Sin embargo, abordar estas necesidades podría resultar difícil considerando la escasa replicabilidad de la literatura actual, lo que dificulta la adopción de innovaciones basadas en LLM en contextos educativos auténticos o su validación con diferentes muestras. Se han identificado problemas de replicación similares en otras áreas de la investigación en tecnología educativa (Yan et al., 2022).

Para la tercera pregunta de investigación (PI3), identificamos varios desafíos éticos en relación con las innovaciones basadas en LLM. En particular, la mayoría de las innovaciones existentes basadas en LLM (92%) solo fueron transparentes para investigadores y profesionales de IA (Nivel 1), y solo nueve estudios pueden considerarse transparentes para expertos y entusiastas de la tecnología educativa (Nivel 2). La razón principal de esta baja transparencia puede atribuirse a la falta de componentes de participación humana en estudios previos. Este hallazgo coincide con la demanda de una IA explicable y centrada en el ser humano, que enfatiza el papel vital de las partes interesadas en el desarrollo de tecnología educativa significativa e impactante (Khosravi et al., 2022; Yang et al., 2021). Involucrar a las partes interesadas durante el desarrollo y la evaluación de las innovaciones basadas en LLM es esencial para abordar cuestiones tanto prácticas como éticas. Por ejemplo, como revelan los hallazgos actuales, las innovaciones basadas en LLMs están sujetas a problemas de privacidad de datos, pero rara vez se mencionaron o investigaron en la literatura (Merine y Purkayastha, 2022), lo que podría deberse a la escasa participación de las partes interesadas en investigaciones previas. Las diversas cuestiones preocupantes en torno a la beneficencia también exigen la participación de las partes interesadas, ya que sus perspectivas son vitales para definir las futuras direcciones de las innovaciones basadas en LLMs, como la forma en que se pueden tomar decisiones responsables con estos sistemas de IA (Schneider et al., 2022). Asimismo, la cuestión de la igualdad en cuanto a las cargas financieras que pueden surgir al adoptar innovaciones que aprovechan LLMs comerciales (p. ej., GPT-3 y Codex) también puede estudiarse en mayor profundidad con las partes interesadas institucionales.

5.2 | Trascendencia

Los hallazgos actuales tienen diversas implicaciones para la investigación y la práctica educativa con LLM, las cuales hemos resumido en tres recomendaciones que buscan respaldar estudios futuros para desarrollar innovaciones prácticas y éticas que puedan beneficiar realmente a los actores educativos. En primer lugar, la amplia gama de escenarios de aplicación de las innovaciones basadas en LLM puede beneficiarse aún más de las mejoras en la capacidad de estos. Actualizar las innovaciones existentes con LLM de vanguardia puede reducir aún más el esfuerzo manual requerido para el ajuste fino y lograr resultados similares (Bang et al., 2023). Considerando los 53 casos de uso identificados de LLM en educación, existen múltiples trayectorias de investigación que podrían impulsar el desarrollo de tecnologías educativas prácticas. Estas vías tienen el potencial de abordar algunos de los desafíos más urgentes que afectan al sistema educativo global. En particular, los casos de uso que involucran apoyo docente, evaluación y calificación, retroalimentación y categorías de generación de contenido (Tabla 1) podrían actuar como catalizadores para el desarrollo de tecnologías educativas que podrían aliviar la carga de trabajo y el estrés mental de los docentes al automatizar las laboriosas tareas asociadas con la creación, evaluación y retroalimentación para las evaluaciones de los estudiantes (Carroll et al., 2022). De manera similar, una mayor exploración de los casos de uso en perfilación y etiquetado, detección, predicción y recomendación podría conducir al desarrollo de tecnologías educativas que puedan brindar apoyo de aprendizaje personalizado para cada estudiante en diversas disciplinas (Wollny et al., 2021). Dichas mejoras podrían mejorar el bienestar general de los docentes y aumentar las oportunidades de aprendizaje de los estudiantes, contribuyendo así al logro del ODS 4 para 2030 (Boeren, 2019). No obstante, los investigadores también deben ser conscientes de las posibles cargas financieras y de recursos que podrían imponerse a los actores educativos al innovar con los LLM comerciales (p. ej., GPT-3/4 y ChatGPT).

Las capacidades inigualables de generación de lenguaje natural que exhiben ChatGPT y otros LLM de vanguardia (por ejemplo, LLaMA y PaLM 2) también podrían inspirar estudios futuros para profundizar en un espectro más amplio de direcciones de investigación. Estos

Incluir comparaciones entre la calidad de los escritos generados por los estudiantes y los generados mediante ChatGPT (Li et al., 2023) y evaluar la capacidad de estos LLM para abordar evaluaciones educativas (Gilson et al., 2023). Dichas exploraciones no solo revelarían el potencial de los LLM y los modelos de IA generativa en la generación de contenido educativo y las tareas de evaluación, sino que también expondrían las posibles amenazas que estos modelos representan para la integridad académica, un problema generalizado en todo el sector educativo (Kasneci et al., 2023). Curiosamente, aprovechar los casos de uso de los LLM en tareas como la creación de representaciones del conocimiento (Zheng et al., 2023) y la clasificación de niveles cognitivos (Liu et al., 2022) podría facilitar la transición de evaluaciones centradas en resultados a evaluaciones centradas en procesos. En este caso, los LLM y los modelos de IA generativa podrían emplearse para evaluaciones de aprendizaje de forma similar a las analíticas de aprendizaje (Gašević et al., 2022). En consecuencia, estudios futuros pueden comenzar a explorar métodos para abordar las amenazas potenciales de los LLM con soluciones basadas en ellos.

Para que las innovaciones basadas en LLM alcancen un alto nivel de preparación y rendimiento tecnológico, es necesario mejorar los estándares actuales de presentación de informes. Los estudios futuros deberían apoyar la iniciativa de publicar sus modelos/sistemas en código abierto siempre que sea posible y proporcionar suficientes detalles sobre los conjuntos de datos de prueba, esenciales para que otros repliquen y validen las innovaciones existentes en diferentes contextos, evitando así el riesgo de una nueva crisis de replicación (Maxwell et al., 2015). Esta iniciativa es especialmente crucial en la era de los modelos generativos de IA, ya que la mayoría de estos modelos, especialmente los comerciales (p. ej., ChatGPT y la serie GPT), son propietarios. Por lo tanto, al utilizar estos LLM para mejorar las prácticas educativas, como la calificación de ensayos de estudiantes (Doewes y Pechenizkiy, 2021), la retroalimentación en tiempo real (Zheng et al., 2022) o la generación de preguntas para actividades de aprendizaje (Sarsa et al., 2022), los investigadores deben ser sistemáticos y transparentes en la presentación de informes sobre el uso y las indicaciones del modelo (Wu, 2022). Por ejemplo, al utilizar la API ChatGPT para la generación de preguntas a escala, los investigadores deben al menos informar los modelos exactos, las indicaciones y la temperatura del modelo utilizados en el proceso, ya que los diferentes modelos pueden diferir en su capacidad para generar contenido preciso y confiable, y las indicaciones son esenciales para que otros repliquen los mismos resultados o resultados similares (Kasneci et al., 2023).

Además de los detalles técnicos y metodológicos mencionados, los investigadores y los responsables de las políticas educativas también deberían considerar los posibles impactos más amplios de las soluciones basadas en LLMs en las diferentes partes interesadas. Por ejemplo, en términos de detección e integridad académica, algunas instituciones han adoptado rápidamente herramientas de detección de IA que afirman tener alta precisión y una baja tasa de falsos positivos. Sin embargo, como se revela en un informe reciente de Turnitin, una empresa cuya función de detección de IA se ha utilizado en más de 38,5 millones de entregas de estudiantes, el rendimiento real de su solución resultó en una incidencia significativamente mayor de falsos positivos en comparación con los hallazgos de laboratorio (Chechitelli, 2023). Esta negligencia puede ser devastadora para los estudiantes que han sido falsamente acusados de mala conducta académica, así como para los educadores que deben gestionar las repercusiones. Este ejemplo reforzó la importancia de realizar estudios científicos rigurosos con las partes interesadas clave al adoptar cualquier solución basada en LLMs que tenga impactos directos o indirectos en estudiantes, educadores y otras partes interesadas. Asimismo, la presentación de informes sobre dichos estudios debe cumplir con altos estándares, incorporando tanto detalles metodológicos como descripciones detalladas de los datos. Estos detalles son especialmente pertinentes al considerar la diversidad cultural de los estudiantes y el hecho de que la mayoría de los LLM se forman principalmente con conjuntos de datos en inglés, lo que podría introducir sesgos hacia los estudiantes no nativos de inglés (Liang et al., 2023).

Adoptar un enfoque centrado en el ser humano al desarrollar y evaluar innovaciones basadas en LLM es esencial para garantizar que estas innovaciones sigan siendo éticas en la práctica, especialmente porque los principios éticos pueden no garantizar una IA ética debido a sus modales de arriba hacia abajo (p. ej., desarrollados por organismos reguladores) (Mittelstadt, 2019). Los estudios futuros deben considerar los problemas éticos que pueden surgir de sus escenarios de aplicación específicos e involucrar activamente a las partes interesadas para identificar y abordar dichos problemas. Específicamente, las innovaciones basadas en LLM deben aspirar a alcanzar al menos el Nivel 3 en el índice de transparencia y TRL-7 en preparación tecnológica. Esto implica un sistema completamente funcional que se integra en entornos de aprendizaje auténticos y es validado por estudiantes y educadores en términos de su practicidad y consideraciones éticas. Para cualquier decisión tomada por las innovaciones basadas en LLM, las partes interesadas relevantes deben estar informadas sobre cómo

Se tomó la decisión, así como los posibles riesgos y sesgos involucrados. Por ejemplo, cuando los estudiantes reciben una evaluación con calificación automática, estas calificaciones deben ir acompañadas de un mensaje de advertencia que indique que han sido calificadas por LLM e IA (Angelone et al., 2022). Los estudiantes también deben tener la oportunidad de consultar con su profesor sobre cualquier inquietud.

La participación activa de las partes interesadas debería trascender el sector educativo, involucrando también a los responsables políticos y a las empresas del sector para establecer las directrices que permitan adoptar innovaciones basadas en LLM en las prácticas de aprendizaje y enseñanza, ya que dichas adopciones podrían tener implicaciones más amplias en la sociedad más allá del sector educativo. Por ejemplo, la colaboración entre humanos e IA podría convertirse en una habilidad esencial para que los estudiantes tengan éxito en el mercado laboral, a medida que las soluciones de IA se convierten en un componente integral de la productividad en el sector industrial (Wang et al., 2020). Por lo tanto, las instituciones que pretendan prohibir las herramientas de IA podrían, inadvertidamente, poner a sus estudiantes en desventaja en comparación con otras instituciones que acogen proactivamente dichos cambios. Esto podría lograrse perfeccionando constantemente sus políticas sobre el uso de LLM y soluciones de IA generativa, basándose en la retroalimentación de las partes interesadas y la evidencia empírica.

5.3 | Limitaciones

Los hallazgos actuales deben interpretarse teniendo en cuenta varias limitaciones. En primer lugar, si bien evaluamos la viabilidad y la ética de las innovaciones basadas en LLM con siete ítems diferentes, es posible que omitiéramos otros aspectos de estos conceptos multidimensionales. No obstante, estos ítems de evaluación se seleccionaron directamente de las definiciones correspondientes y se relacionaron con los problemas más urgentes de la literatura (Adams et al., 2021; Weidinger et al., 2021). En segundo lugar, solo incluimos publicaciones en inglés, lo que podría haber sesgado nuestros hallazgos sobre la disponibilidad de innovaciones basadas en LLM en diferentes países. En tercer lugar, dado que seguimos estrictamente el protocolo PRISMA y solo incluimos publicaciones revisadas por pares, es posible que omitiéramos los trabajos emergentes publicados en diferentes archivos de código abierto. Estos estudios pueden contener hallazgos interesantes sobre los LLM más recientes (p. ej., ChatGPT). Además, esta revisión se centró en el potencial de las innovaciones basadas en LLM para la automatización de tareas educativas y, por lo tanto, otros problemas urgentes, como la posible amenaza a la integridad académica, quedaron fuera del alcance de esta revisión sistemática de alcance. Abordamos brevemente estos temas apremiantes en las implicaciones e ilustramos la importancia de los hallazgos actuales para respaldar futuros estudios educativos que aborden estas cuestiones. Además, dado que este estudio es una revisión sistemática de alcance, no evaluamos la calidad de los estudios incluidos y, por lo tanto, los hallazgos, en particular las métricas de rendimiento extraídas de los estudios revisados, podrían necesitar una evaluación adicional. El objetivo de este estudio es proporcionar una visión general de las diferentes tareas educativas que pueden ser aumentadas por los LLM y los modelos generativos de IA, que pueden servir como punto de referencia para futuros estudios que desarrollen aún más el uso de los modelos de vanguardia (p. ej., ChatGPT y PaLM 2). Además, el índice de transparencia que adoptamos para RQ3 no consideró la transparencia para los estudiantes, lo que podría ser una dirección importante para futuros estudios de IA centrados en el ser humano. Finalmente, reconocemos el rápido desarrollo en el campo de la inteligencia artificial en la educación. Cabe mencionar que varios talleres y artículos preliminares recientes, si bien contribuyen a este campo, no se incorporaron en esta revisión de alcance debido a limitaciones de tiempo (Leiker et al., 2023; Ma et al., 2023; Caines et al., 2023). Su exclusión representa una limitación a la amplitud de este estudio, reconociendo el ritmo incesante de los avances académicos en esta área.

6 | CONCLUSIÓN

En este estudio, revisamos sistemáticamente el estado actual de la investigación educativa sobre LLM e identificamos varios desafíos prácticos y éticos que deben abordarse para que las innovaciones basadas en LLM se vuelvan beneficiosas y

Impactante. Con base en los hallazgos, propusimos tres recomendaciones para estudios futuros: actualizar las innovaciones existentes con modelos de vanguardia, adoptar la iniciativa de modelos/sistemas de código abierto y adoptar un enfoque centrado en el ser humano durante todo el proceso de desarrollo. Estas recomendaciones podrían respaldar estudios futuros para desarrollar innovaciones prácticas y éticas que puedan implementarse en contextos reales para automatizar una amplia gama de tareas educativas.

referencias

Adams, C., Pente, P., Lemermeyer, G. y Rockwell, G. (2021) Directrices éticas de inteligencia artificial para la educación K-12: una revisión del panorama global. En*Inteligencia Artificial en Educación: 22.ª Conferencia Internacional, AIED 2021, Utrecht, Países Bajos, 14-18 de junio de 2021, Actas, Parte II*, 24–28. Springer.

Ahmed, A., Joorabchi, A. y Hayes, MJ (2022) Sobre la aplicación de transformadores de oraciones a respuestas cortas automáticas Calificación en la evaluación combinada. En*33.ª Conferencia Irlandesa de Señales y Sistemas (ISSC) de 2022*, 1–6. IEEE.

Angelone, AM, Galassi, A. y Vittorini, P. (2022) Clasificación automatizada mejorada de oraciones en ejercicios de ciencia de datos. En *Metodologías y Sistemas Inteligentes para el Aprendizaje Potenciado por la Tecnología, 11ª Conferencia Internacional 11*, 12–21. Springer.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W. et al. (2023) Un multitarea, multi-Evaluación lingual y multimodal del chatgpt sobre razonamiento, alucinación e interactividad.*preimpresión de arXiv arXiv:2302.04023*.

Becker, HJ (2000) Resultados de la encuesta sobre enseñanza, aprendizaje y computación.*Archivos de análisis de políticas educativas*,**8**, 51–51.

Beseiso, M., Alzubi, OA y Rashaideh, H. (2021) Un nuevo enfoque automatizado de calificación de ensayos para una educación superior confiable evaluaciones.*Revista de Computación en Educación Superior*,**33**, 727–746.

Boeren, E. (2019) Comprender el objetivo de desarrollo sostenible (ODS) 4 sobre "educación de calidad" desde las perspectivas micro, meso y macro perspectivas.*Revista internacional de educación*,**65**, 277–294.

Brown, H., Lee, K., Mireshghallah, F., Shokri, R. y Tramèr, F. (2022) ¿Qué significa que un modelo de lenguaje preserve ¿Privacidad? En*Conferencia de la ACM 2022 sobre equidad, rendición de cuentas y transparencia*, 2280–2292.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, JD, Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020) Los modelos lingüísticos son aprendices con pocas oportunidades de aprender.*Avances en los sistemas de procesamiento de información neuronal*,**33**, 1877–1901.

Bulut, O. y Yildirim-Erbasli, SN (2022) Generación automática de historias y elementos para evaluaciones de comprensión lectora con transformadores.*Revista Internacional de Herramientas de Evaluación en Educación*,**9**, 72–87.

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Mullooly, A., Nicholls, D. y Buttery, P. (2023) Sobre la aplicación de grandes modelos lingüísticos para la enseñanza de idiomas y la tecnología de evaluación. En*Talleres de AIED*, en prensa.

Carpenter, D., Emerson, A., Mott, BW, Saleh, A., Glazewski, KD, Hmelo-Silver, CE y Lester, JC (2020) Detección Comportamiento fuera de la tarea a partir del diálogo estudiantil en el aprendizaje colaborativo basado en juegos. En*Inteligencia Artificial en la Educación: XXI Conferencia Internacional, AIED 2020, Ifrane, Marruecos, 6-10 de julio de 2020, Actas, Parte I 21*, 55–66. Springer.

Carroll, A., Forrest, K., Sanders-O'Connor, E., Flynn, L., Bower, JM, Fynes-Clinton, S., York, A. y Ziaei, M. (2022) Profesor Estrés y agotamiento en Australia: examen del papel de los factores intrapersonales y ambientales.*Psicología Social de la Educación*,**25**, 441–469.

Cavalcanti, AP, Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D. y Mello, RF (2021) Retroalimentación automática en Entornos de aprendizaje en línea: una revisión sistemática de la literatura.*Computadoras y educación: Inteligencia artificial*,**2**, 100027.

Chaudhry, MA, Cukurova, M. y Luckin, R. (2022) Un marco de índice de transparencia para la IA en la educación. En*Inteligencia artificial en Educación. Pósteres y Resultados de Última Hora, Talleres y Tutoriales, Sectores de Industria e Innovación, Consorcio de Profesionales y Doctorados: 23.ª Conferencia Internacional, AIED 2022, Durham, Reino Unido, 27-31 de julio de 2022, Actas, Parte II*, 195– 198. Springer.

Chechitelli, A. (2023) Actualización sobre detección de escritura con inteligencia artificial del director de productos de Turnitin.https://www.turnitin.com/blog/ai-Actualización de detección de escritura del director de productos de Turnitin.Consultado: 12-06-2023.

Condor, A., Litster, M. y Pardos, Z. (2021) Calificación automática de respuestas cortas con sbert en preguntas fuera de la muestra.*Internacional-Sociedad Nacional de Minería de Datos Educativos*.

Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2018) Bert: Preentrenamiento de transformadores bidireccionales profundos para el lenguaje comprensión.*preimpresión de arXiv arXiv:1810.04805*.

Doewes, A. y Pechenizkiy, M. (2021) Sobre las limitaciones del acuerdo humano-computadora en la calificación automatizada de ensayos.*Enterrar-Sociedad Nacional de Minería de Datos Educativos*.

Doyle, W. y Ponder, GA (1977) La ética práctica en la toma de decisiones docentes.*Intercambio*,**8**, 1–12.

Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., Liu, K., Chen, L., Tran, S., Cheng, N. et al. (2022) Una red neuronal Resuelve, explica y genera problemas matemáticos universitarios mediante síntesis de programas y aprendizaje de pocos intentos a nivel humano. *Actas de la Academia Nacional de Ciencias*,**119**, e2123433119.

Ertmer, PA (1999) Abordar las barreras de primer y segundo orden al cambio: estrategias para la integración de tecnología.*Educativo investigación y desarrollo de tecnología*,**47**, 47–61.

Ferguson, R., Hoel, T., Scheffel, M. y Drachsler, H. (2016) Editorial invitado: Ética y privacidad en el análisis de aprendizaje.*Diario de análisis de aprendizaje*,**3**, 5–15.

Fonseca, SC, Pereira, FD, Oliveira, EH, Oliveira, DB, Carvalho, LS y Cristea, AI (2020) Automático basado en sujetos Contextualización de listas de tareas de programación.*Sociedad Internacional de Minería de Datos Educativos*.

Gašević, D., Dawson, S., Rogers, T. y Gasevic, D. (2016) El análisis de aprendizaje no debería promover una solución única para todos: los efectos de las condiciones de instrucción en la predicción del éxito académico.*Internet y la educación superior*,**28**, 68–84.

Gašević, D., Greiff, S. y Shaffer, DW (2022) Hacia el fortalecimiento de los vínculos entre el análisis del aprendizaje y la evaluación: Desafíos y potencialidades de un nuevo bono prometedor.*Las computadoras en el comportamiento humano*,**134**, 107304. URL:https://www.sciencedirect.com/science/article/pii/S0747563222001261.

Geller, SA, Gal, K., Segal, A., Sripathi, K., Kim, HG, Facciotti, MT, Igo, M., Hoernle, N. y Karger, D. (2021) Nuevos métodos para detectar confusiones en foros del curso: Estudiante, profesor y máquina.*Transacciones IEEE sobre tecnologías de aprendizaje*,**14**, 665–679.

Ghosh, D., Klebanov, BB y Song, Y. (2020) Un estudio exploratorio de la escritura argumentativa por parte de estudiantes jóvenes: Enfoque basado en transformadores. En*Actas del Decimoquinto Taller sobre el Uso Innovador de la PNL para la Creación de Aplicaciones Educativas*, 145–150.

Gilson, A., Safranek, CW, Huang, T., Socrates, V., Chi, L., Taylor, RA, Chartash, D. et al. (2023) ¿Cómo funciona chatgpt? ¿En el examen de licencia médica en Estados Unidos? Implicaciones de los grandes modelos lingüísticos para la educación médica y la evaluación del conocimiento.*Educación Médica JMIR*,**9**, e45312.

Holmes, W. y Porayska-Pomsta, K. (2022)*La ética de la inteligencia artificial en la educación: prácticas, desafíos y debates*. Taylor y Francis.

Jayaraman, J. y Black, J. (2022) Eficacia de un sistema inteligente de preguntas y respuestas para la enseñanza de la alfabetización financiera: un estudio piloto. En*Innovaciones en aprendizaje y tecnología para el trabajo y la educación superior: Actas de la Conferencia de Ideas de Aprendizaje de 2021*, 133–140. Springer.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E. et al. (2023) ¿Chatgpt para bien? sobre las oportunidades y desafíos de los grandes modelos lingüísticos para la educación. *Aprendizaje y diferencias individuales*,**103**, 102274.

Khosravi, H., Shum, SB, Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S. y Gašević, D. (2022) Inteligencia artificial explicable en educación.*Computadoras y educación: Inteligencia artificial*,**3**, 100074.

Kumar, N., Mali, R., Ratnam, A., Kurpad, V. y Magapu, H. (2022) Identificación y tratamiento de las lagunas de conocimiento en los estudiantes. En*3.ª Conferencia Internacional de Tecnologías Emergentes (INCET) 2022*, 1–6. IEEE.

Kurdi, G., Leo, J., Parsia, B., Sattler, U. y Al-Emari, S. (2020) Una revisión sistemática de la generación automática de preguntas para fines educativos.*Revista Internacional de Inteligencia Artificial en Educación*,**30**, 121–204.

Leiker, D., Finnigan, S., Gyllen, AR y Cukurova, M. (2023) Creación de prototipos del uso de modelos lingüísticos grandes (LLMS) para adultos Creación de contenido de aprendizaje a escala. En*Talleres de AIED*, en prensa.

Li, C. y Xing, W. (2021) Generación de lenguaje natural mediante aprendizaje profundo para apoyar a los estudiantes de MOOC.*Revista Internacional de Inteligencia Artificial en la Educación*,**31**, 186–214.

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D. y Chen, G. (2023) ¿Pueden los modelos de lenguaje grandes escribir reflexivamente*Computadoras y educación: Inteligencia artificial*, 100140.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. y Zou, J. (2023) Los detectores de Gpt están sesgados contra los escritores ingleses no nativos. *preimpresión de arXiv arXiv:2304.02819*.

Liu, S., Liu, S., Liu, Z., Peng, X. y Yang, Z. (2022) Detección automatizada del compromiso emocional y cognitivo en mooc Discusiones para predecir el logro del aprendizaje.*Computadoras y educación*,**181**, 104461.

Liu, Z., He, X., Liu, L., Liu, T. y Zhai, X. (2023) El contexto importa: una estrategia para preentrenar el modelo de lenguaje para la educación científica. *preimpresión de arXiv arXiv:2301.12031*.

Ma, Q., Wu, S. y Koedinger, K. (2023) ¿Es llm el mejor compañero de programación? En*Talleres de AIED*, en prensa.

Maheen, F., Asif, M., Ahmad, H., Ahmad, S., Alturise, F., Asiry, O. y Ghadi, YY (2022) Dominio de la informática automática Generación de preguntas de opción múltiple basadas en oraciones informativas.*PeerJ Ciencias de la Computación*,**8**, e1010.

Maxwell, SE, Lau, MY y Howard, GS (2015) ¿Está la psicología sufriendo una crisis de replicación? ¿Qué significa "no poder"? ¿Qué significa realmente "replicar"?*Psicólogo estadounidense*,**70**, 487.

Merine, R. y Purkayastha, S. (2022) Riesgos y beneficios del resumen de texto generado por IA para contenido de nivel experto en posgrado. Informática sanitaria avanzada. En*10.ª Conferencia Internacional sobre Informática Sanitaria (ICHI) del IEEE 2022*, 567–574. IEEE.

Min, B., Ross, H., Sulem, E., Veyseh, APB, Nguyen, TH, Sainz, O., Agirre, E., Heinz, I. y Roth, D. (2021) Avances recientes en el procesamiento del lenguaje natural a través de grandes modelos de lenguaje pre-entrenados: una encuesta.*preimpresión de arXiv arXiv:2111.01243*.

Mittelstadt, B. (2019) Los principios por sí solos no pueden garantizar una IA ética.*Inteligencia de la máquina de la naturaleza*,**1**, 501–507.

Moore, S., Nguyen, HA, Bier, N., Domadia, T. y Stamper, J. (2022) Evaluación de la calidad de los textos cortos generados por los estudiantes Responder preguntas usando gpt-3. En*Educar para un nuevo futuro: Entendiendo la adopción del aprendizaje mejorado con tecnología: 17.ª Conferencia Europea sobre Aprendizaje Mejorado con Tecnología, EC-TEL 2022, Toulouse, Francia, 12-16 de septiembre de 2022, Actas*, 243–257. Springer.

Munn, Z., Peters, MD, Stern, C., Tufanaru, C., McArthur, A. y Aromataris, E. (2018) ¿Revisión sistemática o revisión de alcance? Orientación para los autores a la hora de elegir entre un enfoque de revisión sistemática o de alcance.*Metodología de investigación médica de BMC*, **18**, 1–7.

Nguyen, TT, Le, AD, Hoang, HT y Nguyen, T. (2021) Neu-chatbot: Chatbot para la admisión de economía nacional universidad.*Computadoras y educación: Inteligencia artificial*,**2**, 100036.

Nye, B., Mee, D. y Core, MG (2023) Modelos generativos de lenguaje grande para tutoría basada en diálogo: una consideración temprana de oportunidades y preocupaciones. En*Talleres de AIED*, en prensa.

Oleny, A. (2023) Generación de preguntas de opción múltiple a partir de un libro de texto: LIms iguala el desempeño humano en la mayoría de las métricas. En *Talleres de AIED*, en prensa.

OpenAI (2023) Presentamos chatgpt.https://openai.com/blog/chatgpt.Consultado: 25-02-2023.

Page, MJ, McKenzie, JE, Bossuyt, PM, Boutron, I., Hoffmann, TC, Mulrow, CD, Shamseer, L., Tetzlaff, JM, Akl, EA, Brennan, SE et al. (2021) La declaración prisma 2020: una guía actualizada para informar revisiones sistemáticas. *Revista internacional de cirugía*,**88**, 105906.

Pardo, A. y Siemens, G. (2014) Principios éticos y de privacidad para el análisis de aprendizaje.*Br J Educ Technol*,**45**, 438–450.

Pugh, SL, Subburaj, SK, Rao, AR, Stewart, AE, Andrews-Todd, J. y D'Mello, SK (2021) ¿Qué dices? Modelado automático de habilidades de resolución colaborativa de problemas a partir del habla estudiantil en la naturaleza.*Sociedad Internacional de Minería de Datos Educativos*.

Ramesh, D. y Sanampudi, SK (2022) Un sistema automatizado de puntuación de ensayos: una revisión sistemática de la literatura.*Inteligencia Artificial Revisión de la agencia*,**55**, 2495–2527.

Rudolph, J., Tan, S. y Tan, S. (2023) Chatgpt: ¿Un generador de tonterías o el fin de las evaluaciones tradicionales en la educación superior? *Revista de aprendizaje y enseñanza aplicados*,**6**.

Sallam, M. (2023) La utilidad de chatgpt como ejemplo de grandes modelos lingüísticos en la educación, la investigación y la práctica sanitaria: Revisión sistemática sobre las perspectivas futuras y potenciales limitaciones.*medRxiv*, 2023–02.

Sarsa, S., Denny, P., Hellas, A. y Leinonen, J. (2022) Generación automática de ejercicios de programación y explicaciones de código utilizando modelos de lenguaje grandes. En*Actas de la Conferencia ACM de 2022 sobre Investigación Internacional en Educación en Computación - Volumen 1*, 27–43.

Sawatzki, J., Schlippe, T. y Benner-Wickner, M. (2022) Técnicas de aprendizaje profundo para la calificación automática de respuestas cortas: Pre-Dictando puntuaciones para respuestas en inglés y alemán. En*Inteligencia artificial en la educación: tecnologías emergentes, modelos y aplicaciones: Actas de la 2.ª Conferencia internacional sobre inteligencia artificial en tecnología educativa de 2021*, 65–75. Springer.

Schneider, J., Richner, R. y Riser, M. (2022) Hacia una autocalificación confiable de respuestas breves, multilingües y de múltiples tipos. *Revista Internacional de Inteligencia Artificial en Educación*, 1–31.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, CA y Kersting, K. (2022) Los modelos de lenguaje preentrenados de gran tamaño contienen Sesgos similares a los humanos sobre lo que es correcto o incorrecto hacer.*Inteligencia de la máquina de la naturaleza*,**4**, 258–268.

Ciencia, D. y Grupo, T. () Definiciones y descripciones de los niveles de preparación tecnológica.https://www.dst.defence.gov.au/sitios/predeterminado/archivos/páginas_básicas/documentos/TRL%20Explicaciones_1.pdf.Consultado: 20-01-2023.

Selwyn, N. (2019) ¿Cuál es el problema con el análisis del aprendizaje?*Liga de la Justicia*,**6**, 11–19.

Sha, L., Li, Y., Gasevic, D. y Chen, G. (2022a) ¿Datos más grandes o datos más justos? Ampliación de BERT mediante muestreo activo para fines educativos. Clasificación de textos. En*Actas de la 29ª Conferencia Internacional sobre Lingüística Computacional*, 1275–1285.

Sha, L., Raković, M., Das, A., Gašević, D. y Chen, G. (2022b) Aprovechamiento de técnicas de equilibrio de clases para aliviar problemas algorítmicos Sesgo en las tareas predictivas en educación.*Transacciones IEEE sobre tecnologías de aprendizaje*,**15**, 481–492.

Sha, L., Raković, M., Lin, J., Guan, Q., Whitelock-Wainwright, A., Gašević, D. y Chen, G. (2022c) ¿Es el último el gran-¿Es este un estudio comparativo de enfoques automáticos para clasificar publicaciones en foros educativos?*Transacciones IEEE sobre tecnologías de aprendizaje*.

Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, VM, Gasevic, D. y Chen, G. (2021) Evaluación de algoritmos Equidad mic en clasificadores automáticos de publicaciones en foros educativos. En *Inteligencia Artificial en la Educación: 22.ª Conferencia Internacional, AIED 2021, Utrecht, Países Bajos, 14-18 de junio de 2021, Actas, Parte I 22*, 381–394. Springer.

Shang, J., Huang, J., Zeng, S., Zhang, J. y Wang, H. (2022) Representación y extracción de conocimiento de física basada en Gráfico de conocimiento y clasificación de texto combinada con incrustación para el aprendizaje cooperativo. En *25.ª Conferencia Internacional IEEE 2022 sobre Trabajo Cooperativo con Apoyo de Computadora en Diseño (CSCWD)*, 1053–1058. IEEE.

Sharma, A., Kabra, A. y Kapoor, R. (2021) Redes de cápsulas mejoradas para una calificación automática y robusta de ensayos. En *Máquina Aprendizaje y descubrimiento de conocimiento en bases de datos. Tema de Ciencia de Datos Aplicada: Conferencia Europea, ECML PKDD 2021, Bilbao, España, 13-17 de septiembre de 2021, Actas, Parte V 21*, 365–380. Springer.

Song, W., Hou, X., Li, S., Chen, C., Gao, D., Sun, Y., Hou, J., Hao, A. et al. (2022) Un paciente estándar virtual inteligente para Formación de estudiantes de medicina basada en grafo de conocimiento oral. *Transacciones IEEE sobre multimedia*.

Sridhar, P., Doyle, A., Agarwal, A., Bogart, C., Savelka, J. y Sakr, M. (2023) Aprovechamiento de las películas en el diseño curricular: uso de gpt-4 para apoyar la creación de objetivos de aprendizaje. En *Talleres de AIED*, en prensa.

Su, Y. y Zhang, Y. (2020) Construcción automática de un gráfico de conocimiento temático basado en big data educativo. En *Actas de la 3ª Conferencia Internacional sobre Big Data y Educación de 2020*, 30–36.

Truong, T.-L., Le, H.-L. y Le-Dang, T.-P. (2020) Análisis de sentimientos implementando un modelo de lenguaje preentrenado basado en bert para vietnamita. En *7.ª Conferencia NAFOSTED sobre Ciencias de la Información y la Computación (NICS) 2020*, 362–367. IEEE.

Tsai, Y.-S. y Gasevic, D. (2017) Análisis del aprendizaje en la educación superior: desafíos y políticas: una revisión de ocho estrategias de aprendizaje políticas de análisis. En *Actas de la séptima conferencia internacional sobre análisis del aprendizaje y el conocimiento*, 233–242.

Tsai, Y.-S., Whitelock-Wainwright, A. y Gašević, D. (2020) La paradoja de la privacidad y sus implicaciones para el análisis del aprendizaje. En *Actas de la décima conferencia internacional sobre análisis del aprendizaje y el conocimiento*, 230–239.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN, Kaiser, Ł. y Polosukhin, I. (2017) La atención lo es todo. Necesitas. *Avances en los sistemas de procesamiento de información neuronal*, **30**.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y. y Wang, Q. (2020) De la colaboración entre humanos a Colaboración humano-IA: Diseño de sistemas de IA que puedan trabajar en conjunto con las personas. *Resúmenes ampliados de la conferencia CHI 2020 sobre factores humanos en sistemas informáticos*, 1–6.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A. et al. (2021) Riesgos éticos y sociales del daño de los modelos lingüísticos. *preimpresión de arXiv arXiv:2112.04359*.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M. y Drachsler, H. (2021) ¿Ya llegamos a ese punto? Revisión de la literatura sobre chatbots en educación. *Fronteras en inteligencia artificial*, **4**, 654924.

Wu, J. (2022) Análisis y evaluación del impacto de la integración de la educación en salud mental en la enseñanza universitaria. Cursos de educación cívica en el contexto de la inteligencia artificial. *Comunicaciones inalámbricas y computación móvil*, **2022**.

Wu, X., He, X., Li, T., Liu, N. y Zhai, X. (2023) Ejemplo coincidente como predicción de la siguiente oración (mensp): indicación de disparo cero Aprendizaje para la calificación automática en la enseñanza de las ciencias. *preimpresión de arXiv arXiv:2301.08771*.

Yan, L., Zhao, L., Gasevic, D. y Martinez-Maldonado, R. (2022) Escalabilidad, sostenibilidad y ética del aprendizaje multimodal análisis. En *LAK22: 12ª Conferencia Internacional sobre Análisis del Aprendizaje y Conocimiento*, 13–23.

Yang, SJ, Ogata, H., Matsui, T. y Chen, N.-S. (2021) Inteligencia artificial centrada en el ser humano en la educación: ver lo invisible a través de lo visible. *Computadoras y educación: Inteligencia artificial*, **2**, 100008.

Zawacki-Richter, O., Marín, VI, Bond, M. y Gouverneur, F. (2019) Revisión sistemática de la investigación sobre inteligencia artificial aplicada Aplicaciones en la educación superior: ¿dónde están los educadores?*Revista Internacional de Tecnología Educativa en la Educación Superior*, **16**, 1–27.

Zeng, Z., Gašević, D. y Chen, G. (2023) Sobre la eficacia del aprendizaje curricular en la calificación de textos educativos. En*Actas de la Conferencia AAAI sobre Inteligencia Artificial*.

Zheng, L., Niu, J., Long, M. y Fan, Y. (2023) Un enfoque de construcción automática de gráficos de conocimiento para promover la colaboración Construcción de conocimiento participativo, desempeño grupal, interacción social y regulación socialmente compartida en cscl.*Revista británica de tecnología educativa*,**54**, 686–711.

Zheng, L., Niu, J. y Zhong, L. (2022) Efectos de un enfoque de retroalimentación en tiempo real basado en análisis de aprendizaje sobre el desarrollo de conocimientos. Oración, convergencia de conocimientos, relaciones interactivas y desempeño grupal en cscl.*Revista británica de tecnología educativa*,**53**, 130–149.