

Mathematics assessment in the era of large language models

Quiroz-Chavil, Helga¹; Capuñay-Uceda, Oscar², Capuñay-Uceda, Carlos²

¹Universidad Tecnológica del Perú, Perú, c22682@utp.edu.pe

²Universidad Nacional Pedro Ruiz Gallo, Perú, ocapunayu@unprg.edu.pe, ccapunay@unprg.edu.pe

Abstract– *The rapid adoption of large language models (LLMs), such as ChatGPT, has introduced significant tensions in the assessment of mathematics in higher education, especially in engineering and science programs. Although these models demonstrate a high capacity for solving mathematical problems and generating coherent explanations, their integration challenges the validity of traditional assessment practices focused on the final product. This study presents a critical review of the literature on the use of LLMs in university mathematics assessment, based on a corpus of 27 articles indexed in Scopus, published between 2023 and 2026. A critical analysis methodology was employed, based on categories of pedagogical alignment, validity of evidence, assessment transformation, academic risks, and future projections. The results show that most studies use LLMs primarily as solution generators, while assessment instruments remain largely unchanged. Exam- and task-based assessments predominate, with little attention paid to process assessment, argumentation, or transfer. Likewise, methodological limitations in the empirical evidence and a predominantly declarative treatment of risks such as academic integrity and cognitive dependency are identified. Overall, the review highlights the need for a redesign of mathematics assessment that prioritizes evidence of reasoning, explanation, and verification, integrating LLMs in a critical and pedagogically aligned manner to preserve rigor, equity, and validity in higher education.*

Keywords– *Generative artificial intelligence; large language models (LLMs); mathematics assessment; higher education; deep learning.*

Evaluación en matemáticas en la era de los modelos de lenguaje grandes

Quiroz-Chavil, Helga¹; Capuñay-Uceda, Oscar², Capuñay-Uceda, Carlos²

¹Universidad Tecnológica del Perú, Perú, c22682@utp.edu.pe

²Universidad Nacional Pedro Ruiz Gallo, Perú, ocapunayu@unprg.edu.pe, ccapunay@unprg.edu.pe

Resumen— La rápida adopción de modelos de lenguaje grandes (LLMs), como ChatGPT, ha introducido tensiones significativas en la evaluación de las matemáticas en la educación superior, especialmente en programas de ingeniería y ciencias. Aunque estos modelos demuestran una alta capacidad para resolver problemas matemáticos y generar explicaciones coherentes, su integración desafía la validez de las prácticas evaluativas tradicionales centradas en el producto final. Este estudio presenta una revisión crítica de la literatura sobre el uso de LLMs en la evaluación de matemáticas universitarias, basada en un corpus de 27 artículos indexados en Scopus, publicados entre 2023 y 2026. Se empleó una metodología de análisis crítico sustentada en categorías de alineamiento pedagógico, validez de la evidencia, transformación evaluativa, riesgos académicos y proyección futura. Los resultados muestran que la mayoría de los estudios utilizan los LLMs principalmente como generadores de soluciones, mientras que los instrumentos de evaluación permanecen mayoritariamente inalterados. Predominan evaluaciones basadas en exámenes y tareas, con escasa atención a la evaluación del proceso, la argumentación o la transferencia. Asimismo, se identifican limitaciones metodológicas en la evidencia empírica y un tratamiento predominantemente declarativo de riesgos como la integridad académica y la dependencia cognitiva. En conjunto, la revisión evidencia la necesidad de un rediseño de la evaluación matemática que priorice evidencias de razonamiento, explicación y verificación, integrando los LLMs de forma crítica y pedagógicamente alineada para preservar el rigor, la equidad y la validez en la educación superior.

Palabras clave— Inteligencia artificial generativa; modelos de lenguaje grandes (LLM); evaluación en matemáticas; educación superior; aprendizaje profundo.

I. INTRODUCCIÓN

La irrupción de los modelos de lenguaje grandes (LLMs) ha generado una transformación profunda en la educación superior, particularmente en las disciplinas matemáticas, donde la evaluación cumple un rol central en la validación del razonamiento, la comprensión conceptual y la transferencia del conocimiento. Herramientas como ChatGPT han demostrado una capacidad notable para resolver problemas matemáticos, generar explicaciones paso a paso y producir respuestas coherentes en contextos universitarios, lo que ha suscitado un debate creciente sobre su impacto en las prácticas evaluativas tradicionales [1], [2].

Diversos estudios recientes han documentado que los LLMs pueden igualar o incluso superar el desempeño promedio de los estudiantes en evaluaciones matemáticas formales, incluyendo pruebas de admisión universitaria y

tareas de ingeniería [3], [4]. Sin embargo, esta capacidad plantea interrogantes críticos sobre la validez de la evaluación matemática, ya que la obtención de respuestas correctas no garantiza necesariamente la comprensión conceptual ni el desarrollo del pensamiento matemático profundo [5].

La literatura emergente muestra que, en muchos contextos, los LLMs se han integrado sin un rediseño sustantivo de los instrumentos de evaluación, manteniendo exámenes y tareas centradas en el producto final [1]. Este enfoque ha sido señalado como problemático, dado que facilita la delegación cognitiva y puede erosionar la capacidad del estudiante para justificar, verificar y transferir procedimientos matemáticos a nuevos contextos [6], [7].

Algunos trabajos comienzan a explorar alternativas, como el uso de LLMs para apoyar procesos de argumentación, resolución colaborativa de problemas y andamiaje cognitivo, especialmente en cursos de matemáticas e ingeniería [2], [7]. No obstante, incluso en estos estudios, la evaluación suele permanecer anclada a métricas tradicionales, como calificaciones numéricas o percepciones estudiantiles, con escasa atención a la evaluación del proceso, la explicación o la demostración matemática [8].

Desde una perspectiva más amplia, investigaciones sobre evaluación educativa en la era de la IA generativa advierten que los marcos evaluativos actuales no están preparados para distinguir entre el razonamiento producido por el estudiante y aquel mediado por sistemas generativos [8]. Esta limitación se vuelve especialmente crítica en matemáticas, disciplina en la que la evaluación históricamente ha funcionado como garante de rigor epistemológico y formación del pensamiento lógico-deductivo [5].

Asimismo, varios estudios reconocen riesgos asociados al uso de LLMs en evaluación matemática, como la dependencia cognitiva, la sobreconfianza en respuestas generadas automáticamente y la dificultad para asegurar la autoría académica, aunque dichos riesgos rara vez se traducen en estrategias evaluativas concretas o políticas institucionales claras [1], [4].

En este contexto, pese al rápido crecimiento de investigaciones sobre LLMs en educación matemática, la evidencia se encuentra fragmentada, con enfoques heterogéneos y una marcada ausencia de análisis sistemáticos

centrados específicamente en la evaluación en educación superior. La mayoría de los estudios se concentra en el potencial instructivo o en el desempeño inmediato, dejando en segundo plano cuestiones fundamentales como la validez evaluativa, el alineamiento pedagógico y la transformación de las prácticas de evaluación [6].

Por ello, el presente estudio desarrolla una revisión crítica de la literatura reciente, basada en un corpus de 27 artículos indexados en Scopus, con el objetivo de analizar cómo los LLMs están siendo incorporados en la evaluación de matemáticas en educación superior, qué tipos de evidencias se utilizan para justificar su impacto y cuáles son los principales vacíos conceptuales, metodológicos y normativos que persisten. A partir de este análisis, se busca contribuir a la formulación de principios para el rediseño de la evaluación matemática en un escenario educativo cada vez más mediado por inteligencia artificial generativa.

II. ESTADO DEL ARTE

La literatura reciente sobre LLMs y evaluación en matemáticas universitarias muestra un campo emergente, altamente dinámico y aún metodológicamente heterogéneo, donde la principal tensión se sitúa entre (i) el potencial de los LLMs para apoyar aprendizaje y resolución de problemas y (ii) la crisis de validez de evaluaciones tradicionales frente a la capacidad de estos modelos de producir soluciones plausibles y explicaciones completas.

A. Desempeño de LLMs en pruebas y tareas matemáticas

Una línea visible del estado del arte examina el desempeño de LLMs en contextos evaluativos de alta exigencia. Por ejemplo, se reportan análisis del rendimiento de ChatGPT en pruebas de admisión universitaria, lo que posiciona a los LLMs como actores capaces de influir directamente en instrumentos de evaluación estandarizada y decisiones académicas [3]. En paralelo, estudios en educación de ingeniería exploran la capacidad de versiones recientes de ChatGPT (p. ej., 3.5 y 4.0) para responder problemas técnico-matemáticos, resaltando su potencial y, simultáneamente, la necesidad de cautela al interpretar “aciertos” como sinónimo de comprensión [4].

B. Integración de LLMs en cursos universitarios y efectos sobre evaluación

Otra línea se centra en experiencias de integración en cursos (por ejemplo, cálculo), donde ChatGPT se utiliza como apoyo para resolver ejercicios, explicar procedimientos o acompañar actividades de aprendizaje. Estas experiencias evidencian oportunidades para reforzar el estudio autónomo, pero también muestran que, sin rediseño evaluativo, la herramienta puede facilitar la obtención de productos correctos sin trazabilidad del razonamiento del estudiante [2]. En la misma dirección, análisis comparativos sobre el uso de ChatGPT en educación

superior advierten que el “diálogo fluido” del modelo no reemplaza un diseño evaluativo que capture comprensión, justificación y verificación, reforzando la necesidad de evaluación centrada en procesos [1].

C. Evaluación, integridad y redefinición del “evidenciar aprendizaje” en la era GenAI

Un aporte clave del estado del arte es el desplazamiento del foco desde “si el estudiante puede resolver” hacia “cómo demostrar autoría, proceso y comprensión” en un entorno donde la generación automática es ubicua. En particular, trabajos que discuten evaluación educativa en la era de la IA generativa enfatizan que las prácticas evaluativas vigentes tienen dificultades para discriminar entre razonamiento propio y producción mediada, y que esta dificultad exige rediseños que incorporen evidencias multimodales (explicación, defensa, iteración, justificación) [8].

D. De tutor/andamiaje a muleta: dependencia cognitiva y calidad del aprendizaje

En el plano pedagógico-cognitivo, emergen estudios que problematizan el rol del LLM como andamiaje frente al riesgo de convertirse en “muleta” que reduzca esfuerzo cognitivo y autonomía. Esta preocupación aparece explícitamente en trabajos que analizan el soporte de LLMs en estudiantes universitarios, describiendo tensiones entre mejora inmediata y desarrollo de habilidades matemáticas duraderas (Wang et al., 2025). En paralelo, revisiones sobre dificultades en matemáticas universitarias resaltan que la comprensión profunda requiere intervención pedagógica orientada a conceptualización, lo que refuerza la idea de que el valor del LLM depende críticamente de cómo se diseñen tareas y evaluación [5].

E. Nuevas prácticas: argumentación, colaboración y generación automática de ítems

El estado del arte también registra intentos de expansión de prácticas evaluativas hacia dimensiones menos automatizables: argumentación, discusión conceptual y trabajo colaborativo. Se reportan experiencias centradas en argumentación sobre conceptos matemáticos, que abren una vía para evaluar razonamiento y comunicación matemática más allá del resultado final [6]. Asimismo, aparecen propuestas de resolución de problemas mediante pair-programming apoyado por IA, que sugieren un giro hacia evaluación del proceso en actividades de carácter aplicado [7]. Finalmente, en el plano técnico-evaluativo, la generación automática de ítems en áreas STEM (incluida matemática) emerge como un frente de investigación con potencial para diversificar bancos de preguntas, aunque plantea retos de validez, sesgo y control de calidad [9].

F. Panorama general: impacto sistémico y necesidad de marcos específicos para matemática

La literatura empieza a consolidar perspectivas más amplias sobre el impacto de los LLMs en la educación matemática, subrayando que el cambio no es únicamente instrumental, sino estructural: afecta qué se considera evidencia de aprendizaje, cómo se certifican competencias y cómo se sostienen la confianza y la justicia evaluativa [10]. En síntesis, el estado del arte converge en una conclusión crítica: la evaluación matemática universitaria requiere rediseño para preservar validez, equidad y rigor, y para aprovechar LLMs sin degradar el aprendizaje profundo [1], [8], [11].

III. METODOLOGÍA

A. Diseño del estudio

Se realizó una revisión de análisis crítico (critical analysis review) orientada a: (i) identificar patrones de uso de LLMs en evaluación matemática universitaria, (ii) evaluar el alineamiento pedagógico y la validez de la evidencia reportada, y (iii) proponer orientaciones para el rediseño de prácticas evaluativas. Este enfoque se eligió por el carácter reciente, disperso y heterogéneo del campo, donde una síntesis interpretativa y evaluativa resulta más adecuada que un meta-análisis cuantitativo.

B. Fuente de datos y estrategia de búsqueda

El corpus se obtuvo desde Scopus (Elsevier), utilizando la siguiente cadena de búsqueda diseñada para recuperar estudios que intersectan LLMs / IA generativa, evaluación, matemáticas y educación superior:

TITLE-ABS-KEY (("large language model*" OR ChatGPT OR "generative AI") AND (assessment OR evaluation OR exam* OR grading OR rubric*) AND ("mathematics education" OR "engineering mathematics" OR calculus OR "linear algebra") AND ("higher education" OR universit* OR undergraduate))

El periodo de interés fue 2023–2026, por corresponder a la fase de adopción temprana de LLMs en educación universitaria.

C. Proceso de selección

La búsqueda arrojó 28 registros exportados a CSV. Se aplicaron criterios de inclusión/exclusión definidos a priori:

Inclusión: educación superior; matemáticas explícitas (p. ej., cálculo, álgebra, evaluación matemática universitaria); presencia de LLM/IA generativa (p. ej., ChatGPT); enfoque en evaluación (exámenes, tareas, instrumentos, retroalimentación, integridad).

Exclusión: ausencia de alguno de los cuatro ejes (LLM, evaluación, matemáticas, educación superior), duplicados o documentos sin relevancia evaluativa.

Con este procedimiento, se incluyeron 27 artículos para el análisis crítico.

D. Unidad de análisis y alcance (basado en resúmenes)

Dado que el usuario solicitó explícitamente trabajar con el resumen de cada artículo, la unidad principal de análisis fue el abstract (más título y palabras clave cuando estuvieron disponibles). En consecuencia, la codificación se considera preliminar-analítica, adecuada para mapear el campo y producir síntesis crítica; cuando el resumen no ofreció suficiente detalle (p. ej., tipo exacto de evaluación, métricas), se registró como “no especificado” para evitar inferencias no sustentadas.

E. Marco de categorías y codificación

Se empleó una matriz de extracción estructurada en seis categorías críticas:

- 1) Alineamiento pedagógico (competencias matemáticas evaluadas vs. tipo de evaluación vs. rol del LLM).
- 2) Validez y calidad de evidencia (diseño, indicadores, presencia de comparación, medidas de transferencia/explicación).
- 3) Transformación evaluativa (replicación de prácticas tradicionales vs. rediseño centrado en proceso, explicación y trazabilidad).
- 4) Riesgos académicos y socio-técnicos (integridad, dependencia cognitiva, alucinaciones, equidad).
- 5) Posicionamiento normativo-ético (políticas de uso, transparencia, responsabilidad).
- 6) Agenda futura (vacíos, recomendaciones, líneas de investigación).

La codificación se realizó en dos pasadas: primera pasada: identificación de pertinencia y extracción de rasgos explícitos del resumen (p. ej., curso, instrumento, rol del LLM), y segunda pasada: codificación analítica por categorías (p. ej., “transformación nula/parcial/alta”), manteniendo criterios conservadores para no sobre-interpretar la evidencia.

IV. RESULTADOS

A. Caracterización general del corpus analizado

El corpus final estuvo conformado por 27 estudios publicados entre 2023 y 2026, lo que confirma el carácter reciente y emergente de la investigación sobre LLMs y evaluación en matemáticas en educación superior. La producción presenta una tendencia creciente, con mayor concentración de publicaciones en los años 2024 y 2025, lo

que sugiere un interés académico en rápida expansión tras la adopción masiva de modelos como ChatGPT [1], [8]. En cuanto al tipo de documento, predominan los artículos de revista y ponencias en conferencias, mientras que las revisiones formales son escasas. Este patrón indica que el campo aún se encuentra en una fase de exploración conceptual y empírica, con limitada consolidación teórica [5], [10].

B. Rol asignado a los LLMs en contextos de evaluación matemática

El análisis revela que el rol más frecuente de los LLMs en los estudios revisados es el de generador de soluciones o apoyo directo a la resolución de problemas, especialmente en tareas y exámenes matemáticos universitarios [1], [4]. En más del 80 % de los artículos, ChatGPT es utilizado como herramienta para producir respuestas completas, explicaciones paso a paso o verificaciones automáticas. Un subconjunto menor de estudios explora roles alternativos, como tutor o herramienta de andamiaje, orientados a promover la reflexión y el aprendizaje autónomo [2], [11]. Sin embargo, incluso en estos casos, la función evaluativa permanece mayoritariamente centrada en el producto final, sin mecanismos sistemáticos para capturar el proceso de razonamiento del estudiante.

Estos resultados evidencian una asimetría entre el potencial pedagógico atribuido a los LLMs y su uso efectivo en prácticas evaluativas, lo que refuerza la necesidad de rediseñar los instrumentos de evaluación matemática en educación superior [6].

C. Tipos de evaluación matemática predominantes

Respecto a los formatos de evaluación, los resultados muestran un claro predominio de exámenes escritos y tareas domiciliarias, presentes en la mayoría de los estudios analizados [2], [3]. Estas evaluaciones se caracterizan por centrarse en la obtención de respuestas correctas, lo que las hace particularmente vulnerables al uso directo de LLMs. La evaluación oral, la defensa de soluciones y la evaluación basada en procesos aparecen de forma marginal o anecdótica. Solo un número muy reducido de estudios menciona explícitamente estrategias orientadas a evaluar la argumentación, la justificación o la comunicación matemática [6], [7]. Este patrón sugiere que, pese a la presencia de LLMs capaces de automatizar procedimientos, las prácticas evaluativas matemáticas continúan ancladas en modelos tradicionales, con escasa transformación estructural [8].

D. Evidencia empírica y validez de las conclusiones reportadas

En relación con la calidad de la evidencia empírica, la mayoría de los estudios basa sus conclusiones en indicadores de desempeño inmediato (calificaciones, tasas de acierto) y en percepciones estudiantiles sobre utilidad o satisfacción [1],

[2]. Los diseños cuasi-experimentales y experimentales son minoritarios, y rara vez incluyen medidas de transferencia a nuevos problemas o evaluación longitudinal del aprendizaje.

Solo un pequeño número de artículos emplea rúbricas explícitas para evaluar razonamiento, comprensión conceptual o argumentación matemática [6]. En consecuencia, gran parte de la literatura tiende a inferir mejoras en el aprendizaje a partir de evidencias limitadas, lo que debilita la validez de sus afirmaciones [5], [11].

E. Riesgos académicos y socio-técnicos identificados

Los riesgos asociados al uso de LLMs en evaluación matemática son ampliamente reconocidos, aunque escasamente operacionalizados. La integridad académica y la autoría de las soluciones son mencionadas de forma recurrente, pero pocos estudios proponen mecanismos concretos para abordarlas [1], [8].

Asimismo, varios trabajos advierten sobre la dependencia cognitiva y la posibilidad de que los LLMs reduzcan el esfuerzo intelectual requerido en la resolución de problemas matemáticos [4], [11]. Las alucinaciones matemáticas y los problemas de equidad de acceso reciben una atención marginal, a pesar de su relevancia para la justicia y validez evaluativa [10].

F. Síntesis de hallazgos

En conjunto, los resultados muestran que la literatura sobre LLMs y evaluación en matemáticas universitarias se caracteriza por: un uso predominante de los LLMs como generadores de soluciones, más que como herramientas evaluativas estructurales; la persistencia de formatos de evaluación tradicionales, poco alineados con la nueva realidad tecnológica; evidencia empírica limitada, centrada en rendimiento inmediato y percepción; y reconocimiento declarativo de riesgos, sin traducción sistemática en diseño evaluativo. Estos hallazgos configuran un escenario en el que la validez de la evaluación matemática se encuentra en tensión, y donde la necesidad de rediseño pedagógico y evaluativo emerge como una conclusión central del análisis [1], [8].

V. DISCUSIÓN

A. Interpretación general: la evaluación matemática entra en una “zona de inestabilidad” epistemológica

Los resultados sugieren que el impacto de los LLMs en matemáticas universitarias no es únicamente instrumental (una herramienta más), sino estructural: altera la relación entre evidencia, autoría y validez en evaluación. Cuando un modelo puede producir soluciones completas, explicaciones paso a paso y justificaciones plausibles, la evaluación centrada en el

producto (respuesta final) pierde capacidad discriminante para inferir aprendizaje. Esta tensión aparece de forma consistente en el corpus: los estudios describen usos frecuentes de LLMs para resolver ejercicios y apoyar tareas/exámenes, mientras que el rediseño evaluativo es comparativamente incipiente [1], [2], [8]. En consecuencia, la “crisis” no es que los estudiantes usen LLMs, sino que los instrumentos tradicionales ya no garantizan que lo evaluado represente razonamiento propio, comprensión o transferencia.

Un hallazgo particularmente relevante es que el corpus parece moverse entre dos polos: (i) estudios que evidencian capacidad de desempeño de LLMs en pruebas formales y contextos exigentes (por ejemplo, admisión universitaria o tareas técnicas), y (ii) estudios que problematizan la interpretación pedagógica de ese desempeño [3], [4], [5]. Esta coexistencia alimenta un riesgo frecuente: confundir “respuesta correcta” con “aprendizaje profundo”, especialmente cuando la evidencia empírica se limita a rendimiento inmediato o percepciones.

B. Alineamiento pedagógico: el principal desajuste es “competencias profundas vs. evidencias superficiales”

La discusión central del corpus puede sintetizarse como un desalineamiento entre lo que matemáticas universitarias declara formar (razonamiento, modelamiento, justificación, prueba y transferencia) y lo que muchas evaluaciones siguen capturando (producto final y ejecución procedimental). En varios estudios, el rol del LLM se aproxima a “resolver” o “producir solución”, mientras que el instrumento evaluativo permanece inalterado: exámenes o tareas con baja trazabilidad del proceso. Esta combinación tiende a debilitar la validez interpretativa de la evaluación: ya no está claro si la calificación representa competencia del estudiante o competencia del sistema utilizado [1], [2].

Los trabajos que exploran argumentación y discusión conceptual ofrecen una pista valiosa: desplazar la evaluación hacia la explicación y la argumentación puede recuperar evidencias más cercanas a pensamiento matemático [6]. De modo similar, propuestas asociadas a resolución colaborativa (p. ej., pair-programming con apoyo de IA) sugieren que evaluar procesos y decisiones puede ser más robusto que evaluar solo resultados [7]. No obstante, el corpus muestra que estas alternativas siguen siendo minoritarias; por ello, el aporte más consistente de la revisión es conceptual: si la evaluación no cambia, la validez se erosiona.

C. Transformación evaluativa: predominio de adaptación táctica, no rediseño estratégico

Una lectura fina de la literatura sugiere que muchas iniciativas se ubican en lo que podría llamarse “adaptación táctica”: permitir, prohibir o tolerar el uso de LLMs sin rediseñar la tarea evaluativa. En cambio, el rediseño

estratégico implicaría reconfigurar (i) qué se pide, (ii) qué evidencias se exigen, y (iii) cómo se justifica la inferencia de aprendizaje. En el corpus, la evaluación oral, la defensa de soluciones, la trazabilidad (bitácoras/versiones) y la evaluación de transferencia aparecen como caminos, pero aún con baja presencia explícita [6], [8].

Este patrón es coherente con la fase de madurez del campo: una porción importante de publicaciones son aportes descriptivos o conceptuales, y los estudios empíricos con rediseños evaluativos sostenidos son menos frecuentes. Esto sugiere un campo que todavía está “probando” enfoques, más que consolidando modelos evaluativos robustos. En ese sentido, la literatura comienza a señalar que integrar LLMs en cursos (por ejemplo, cálculo) puede favorecer aprendizaje autónomo, pero sin un marco de evaluación auténtica el impacto pedagógico queda difícil de sostener [2], [11].

D. Validez y calidad de evidencia: abundan afirmaciones, faltan métricas de transferencia y comprensión

Uno de los hallazgos más críticos es la brecha entre las preguntas evaluativas relevantes (“¿comprende?”, “¿transfiere?”, “¿justifica?”) y los indicadores usados para responderlas. En el corpus, se observa dependencia de calificaciones o desempeño inmediato, junto con percepciones de utilidad. Estas métricas son informativas, pero insuficientes para sostener conclusiones fuertes sobre aprendizaje profundo, especialmente en matemáticas universitarias [1], [2], [5].

En contraste, los estudios que se aproximan a evaluación de razonamiento mediante rúbricas, argumentación o experiencias estructuradas ofrecen un camino metodológico más compatible con las competencias matemáticas [6]. Sin embargo, la baja frecuencia de evaluaciones de transferencia (resolver problemas nuevos sin plantillas) y de análisis longitudinal implica que el campo aún no cuenta con evidencia suficiente para afirmar que el uso de LLMs “mejora el aprendizaje” en sentido duradero. En este contexto, la revisión crítica aporta una recomendación metodológica: avanzar hacia diseños que distingan (i) desempeño asistido, (ii) desempeño independiente, y (iii) capacidad de explicar/defender decisiones.

E. Integridad académica: el problema no es solo plagio, es “indeterminación de autoría cognitiva”

La literatura reconoce de manera recurrente preocupaciones de integridad académica; sin embargo, una lectura crítica sugiere que el problema es más profundo que plagio textual. En matemáticas, el reto principal es la indeterminación de autoría cognitiva: incluso con respuestas originales, no queda claro quién realizó el razonamiento. Esto obliga a diferenciar entre “texto producido” y “capacidad demostrada”. Esta preocupación aparece en discusiones sobre evaluación en la era de la IA generativa, donde se advierte que

los marcos tradicionales de evaluación y aseguramiento de aprendizaje requieren actualización para mantener justicia y credibilidad [8].

Una consecuencia directa para matemáticas universitarias es que políticas de “permitido/prohibido” son insuficientes si no se acompañan de rediseño. En otras palabras, la integridad no se resuelve solo con vigilancia, sino con instrumentos evaluativos que produzcan evidencias difíciles de externalizar: explicaciones situadas, defensas orales breves, variaciones personalizadas, análisis de errores, y evaluación del proceso. El corpus aporta ejemplos parciales de rutas alternativas (argumentación, colaboración), pero aún no consolida un estándar [6], [7].

F. Riesgos cognitivos: “andamiaje o muleta” como dilema central

Un aporte conceptual importante proviene del dilema “andamiaje vs. muleta” (scaffold vs. crutch). En la literatura revisada, el LLM puede actuar como tutor que promueve reflexión, pero también como sustituto de esfuerzo, reduciendo práctica deliberada y autorregulación. Este dilema es especialmente agudo en matemáticas, donde la automatización temprana puede degradar el desarrollo de intuiciones y esquemas de resolución [5], [11].

Aquí emerge una repercusión pedagógica para la evaluación: si el LLM se usa como apoyo, la evaluación debe diseñarse para detectar comprensión, no solo ejecución. Por ejemplo, pedir al estudiante que justifique por qué un método es apropiado, comparar alternativas, analizar errores deliberados o explicar condiciones de validez puede disminuir la dependencia de soluciones generadas y aumentar evidencia de comprensión. En la misma dirección, experiencias en cálculo con integración de ChatGPT muestran potencial para apoyo, pero también resaltan la necesidad de guiar usos y mantener exigencias evaluativas que midan algo más que el resultado [2].

G. Alucinaciones y fiabilidad matemática: un riesgo subatendido en evaluación

Aunque parte del corpus considera desempeño de modelos avanzados en contextos de ingeniería, la revisión sugiere que la discusión sobre fiabilidad matemática (errores plausibles, pasos incorrectos convincentes) aparece de modo menos operacionalizado de lo deseable [3], [4]. Esto es crítico en evaluación: si se permite el uso de LLMs sin mecanismos de verificación, se corre el riesgo de evaluar productos con errores sutiles o razonamientos inválidos presentados con alta fluidez.

De aquí se deriva una recomendación para la agenda: incorporar evaluación explícita de verificación (comprobación de resultados, contrastación por métodos alternativos) como

competencia evaluativa. Esta orientación puede transformar el riesgo (alucinación) en oportunidad formativa: entrenar a los estudiantes a auditar respuestas generadas, desarrollar criterio matemático y mejorar alfabetización crítica en IA [8].

H. Equidad y justicia evaluativa: el “elefante en la habitación” con baja explicitación

La revisión crítica también sugiere que la dimensión de equidad recibe atención limitada en comparación con integridad o rendimiento. Sin embargo, en evaluación matemática universitaria, la equidad es crucial: acceso desigual a modelos (versiones pagas vs. gratuitas), conectividad, idioma, y habilidades de prompting pueden generar ventajas no académicas. Si la evaluación acepta herramientas sin provisión institucional y sin criterios transparentes, la justicia evaluativa se ve comprometida. Algunas lecturas más amplias sobre el impacto de LLMs en educación matemática apuntan a implicaciones sistémicas de este tipo, aunque el tratamiento aún es incipiente en el corpus [8], [10].

I. Consideraciones para el rediseño de evaluación matemática en educación superior

A partir de los patrones observados, la discusión sostiene que el futuro de la evaluación matemática con LLMs debe moverse desde “control de uso” hacia “diseño de evidencias”. En términos operativos, el corpus sugiere cuatro direcciones de rediseño:

- 1) Evaluación centrada en proceso: incorporar evidencia del razonamiento (versiones, bitácoras, decisiones intermedias), especialmente en tareas donde el producto es fácilmente generable [7], [8].
- 2) Evaluación de explicación y argumentación: pedir justificación, interpretación, defensa conceptual y análisis de errores; experiencias de argumentación muestran que es una vía factible para capturar competencias profundas [5], [6].
- 3) Evaluación de transferencia: priorizar problemas nuevos, variaciones contextualizadas y tareas no rutinarias; la evidencia del corpus sugiere que medir transferencia es el gran vacío que debe cerrarse para evitar inferencias infladas de aprendizaje [1], [11].
- 4) Alfabetización evaluativa en IA: convertir el uso del LLM en objeto de evaluación, exigiendo verificación, crítica y transparencia del proceso, alineado con las preocupaciones sobre evaluación en la era GenAI [2], [8].

J. Limitaciones de la revisión y robustez del enfoque crítico

Esta discusión se fundamenta en el análisis sistemático de títulos, resúmenes y palabras clave de los 27 estudios incluidos, lo que habilita una síntesis crítica sólida de

tendencias y vacíos, pero limita el detalle fino sobre instrumentos, rúbricas específicas, muestras y procedimientos cuando el resumen no lo explicita. Por ello, algunas categorías se codifican conservadoramente como “no especificado” cuando corresponde. Aun así, el corpus es lo suficientemente focalizado como para sustentar conclusiones robustas sobre: (i) predominio de evaluaciones tradicionales, (ii) rol resolutorio de LLMs, (iii) déficit de métricas de transferencia, y (iv) necesidad de rediseño evaluativo [1], [6], [8].

K. Agenda de investigación: hacia una evaluación matemática “a prueba de LLMs” y orientada a futuro

Finalmente, la revisión crítica sugiere una agenda de investigación con alto potencial de impacto: Diseños cuasi-experimentales que comparen evaluación tradicional vs. evaluación centrada en explicación/proceso, midiendo transferencia y retención [2], [11]; Modelos de rúbricas para razonamiento y verificación, especialmente en cálculo e ingeniería, con evidencia de validez [5], [6]; Estudios de equidad sobre acceso y competencia de uso de LLMs en evaluación matemática, con enfoques institucionales [8], [10]; e Investigación sobre generación de ítems y control de calidad (sesgo, dificultad, validez), como línea emergente para diversificar evaluaciones en STEM [9].

En conjunto, el corpus no solo documenta un cambio tecnológico, sino que evidencia una transición epistemológica: la evaluación matemática debe evolucionar para seguir siendo un indicador confiable de competencia en un mundo donde el cálculo simbólico, la explicación y la producción textual pueden ser generados automáticamente. La oportunidad —con visión de futuro— es convertir a los LLMs en catalizadores de una evaluación más auténtica, centrada en razonamiento, verificación y transferencia, en lugar de limitarse a reproducir prácticas centradas en el producto final [1], [6], [8].

VI. CONCLUSIONES

La presente revisión crítica analizó de manera sistemática 27 estudios recientes (2023–2026) sobre el uso de modelos de lenguaje grandes (LLMs) en la evaluación de matemáticas en educación superior, revelando un campo emergente caracterizado por avances tecnológicos rápidos, pero con importantes vacíos pedagógicos, evaluativos y metodológicos. Los resultados muestran que la principal transformación inducida por los LLMs no se limita al apoyo instruccional, sino que cuestiona directamente la validez epistemológica de las prácticas evaluativas tradicionales en matemáticas universitarias.

Una primera conclusión central es que la evaluación matemática basada en el producto final (respuestas correctas en exámenes o tareas) ha perdido capacidad para funcionar como evidencia confiable de aprendizaje en un contexto donde

los LLMs pueden generar soluciones completas, coherentes y formalmente plausibles. La literatura revisada evidencia que, en la mayoría de los estudios, los LLMs son utilizados como generadores de soluciones o apoyo directo a la resolución de problemas, mientras que los instrumentos evaluativos permanecen inalterados, produciendo un desalineamiento pedagógico entre las competencias matemáticas declaradas (razonamiento, prueba, transferencia) y las evidencias realmente recogidas [1], [2], [8].

En segundo lugar, el análisis muestra que la evidencia empírica utilizada para sostener afirmaciones sobre mejora del aprendizaje es, en general, limitada. Predominan indicadores de desempeño inmediato y medidas de percepción estudiantil, con escasa presencia de evaluaciones de transferencia, argumentación o verificación matemática. Esta situación refuerza el riesgo de inferencias infladas, donde se asume aprendizaje profundo a partir de evidencias que no capturan procesos cognitivos de alto nivel [5], [11].

Asimismo, la revisión pone de manifiesto que los riesgos académicos y socio-técnicos asociados al uso de LLMs — como la integridad académica, la dependencia cognitiva, las alucinaciones matemáticas y la equidad de acceso — son ampliamente reconocidos, pero rara vez se traducen en estrategias evaluativas concretas. En particular, la problemática de la autoría cognitiva emerge como un desafío clave para la evaluación matemática: aun en ausencia de plagio textual, resulta difícil determinar quién realizó el razonamiento matemático evaluado [8].

Desde una perspectiva pedagógica, los hallazgos sugieren que el dilema no es si los LLMs deben o no utilizarse en matemáticas universitarias, sino cómo rediseñar la evaluación para que continúe siendo un mecanismo válido, justo y formativo. En este sentido, la literatura revisada apunta, de manera aún incipiente, hacia enfoques de evaluación centrados en el proceso, la explicación, la argumentación y la transferencia, los cuales aparecen como vías prometedoras para preservar el rigor matemático en entornos mediador por IA generativa [6], [7].

En conjunto, esta revisión crítica concluye que la integración de LLMs en la evaluación matemática universitaria exige un cambio de paradigma: pasar de evaluar productos estáticos a evaluar evidencias ricas de razonamiento, donde la interacción con la IA no se oculte ni se prohíba de forma superficial, sino que se haga explícita, verificable y pedagógicamente significativa. Lejos de representar una amenaza insalvable, los LLMs pueden convertirse en catalizadores para una evaluación matemática más auténtica, reflexiva y alineada con las competencias que la educación en ingeniería y ciencias demanda en el siglo XXI [10].

VII. LÍNEAS FUTURAS DE INVESTIGACIÓN

A partir de los hallazgos y limitaciones identificadas, se proponen las siguientes líneas futuras de investigación, orientadas a fortalecer el rigor y la pertinencia de la evaluación matemática en contextos mediados por LLMs: se requieren estudios que desarrollen y validen rúbricas explícitas para evaluar razonamiento matemático, explicación, verificación y transferencia a problemas nuevos, comparando su desempeño frente a evaluaciones tradicionales en entornos con y sin LLMs [5], [11]. La literatura futura debería priorizar diseños cuasi-experimentales y longitudinales que permitan analizar no solo el desempeño inmediato, sino la retención del aprendizaje y la evolución del pensamiento matemático cuando los LLMs se integran de forma sistemática en la evaluación [2]. Una línea crítica consiste en investigar mecanismos evaluativos que permitan capturar la trazabilidad del razonamiento, como bitácoras, versiones sucesivas, análisis de errores y defensas orales breves, con el fin de reducir la indeterminación de autoría cognitiva en matemáticas [8]. Futuros estudios deberían explorar modelos de andamiaje adaptativo, donde el LLM ofrezca apoyo graduado y la evaluación mida explícitamente la autonomía progresiva del estudiante, abordando el dilema “andamiaje vs. muleta” de forma empírica [11]. Se identifica la necesidad de investigaciones que analicen de manera sistemática las alucinaciones matemáticas de los LLMs y que incorporen la verificación de soluciones como competencia evaluativa clave, transformando un riesgo técnico en una oportunidad formativa [4].

Finalmente, se requieren estudios que aborden explícitamente la equidad evaluativa, considerando diferencias de acceso a LLMs, versiones del modelo, idioma y habilidades de uso, así como el rol de las instituciones en garantizar condiciones justas y transparentes de evaluación [8], [10].

REFERENCES

- [1] D. Santandreu Calonge, L. Smail, y F. Kamalov, «Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics», *J. Appl. Learn. Teach.*, vol. 6, n.º 2, pp. 346-357, ago. 2023, doi: 10.37074/jalt.2023.6.2.22.
- [2] D. Serhan y N. Welcome, «Integrating ChatGPT in the Calculus Classroom: Student Perceptions», *Int. J. Technol. Educ. Sci.*, vol. 8, n.º 2, pp. 325-335, abr. 2024, doi: 10.46328/ijtes.559.
- [3] A. Udias *et al.*, «ChatGPT's performance in university admissions tests in mathematics», *Int. Electron. J. Math. Educ.*, vol. 19, n.º 4, p. em0795, oct. 2024, doi: 10.29333/iejme/15517.
- [4] M. E. Frenkel y H. Emara, «ChatGPT-3.5 and -4.0 and mechanical engineering: Examining performance on the FE mechanical engineering and undergraduate exams», *Comput. Appl. Eng. Educ.*, vol. 32, n.º 6, p. e22781, 2024, doi: 10.1002/cae.22781.
- [5] N. Karjanto, «Investigating difficulties and enhancing understanding in linear algebra: Leveraging SageMath and ChatGPT for (orthogonal) diagonalization and singular value decomposition», *Math. Biosci. Eng.*, vol. 20, n.º 9, pp. 16551-16595, 2023, doi: 10.3934/mbe.2023738.
- [6] S. Urhan, O. Gençaslan, y Ş. Dost, «An argumentation experience regarding concepts of calculus with ChatGPT», *Interact. Learn.*

- Environ.*, vol. 32, n.º 10, pp. 7186-7211, nov. 2024, doi: 10.1080/10494820.2024.2308093.
- [7] S. Urhan y S. A. Kocadere, «Problem-Solving Through Pair-Programming: The Mediatonal Role of ChatGPT», *2024 5th Int. Conf. Electron. Eng. Inf. Technol. Amp Educ. EEITE*, pp. 1-10, may 2024, doi: 10.1109/EEITE61750.2024.10654434.
- [8] P. W. Wachira, X. Liu, y S. Koc, *Educational Assessments in the Age of Generative AI*. IGI Global Scientific Publishing, 2024. Accedido: 31 de enero de 2026. [En línea]. Disponible en: <https://www.igi-global.com/gateway/book/343544>
- [9] K. W. Chan *et al.*, «Automatic item generation in various STEM subjects using large language model prompting», *Comput. Educ. Artif. Intell.*, vol. 8, p. 100344, jun. 2025, doi: 10.1016/j.caeai.2024.100344.
- [10] I. Rizos y N. Gkrekas, «The impact of LLMs on mathematics education and research at the university», *Soc. Sci. Humanit. Open*, vol. 12, p. 101969, ene. 2025, doi: 10.1016/j.ssaho.2025.101969.
- [11] K. D. Wang, Z. Wu, L. Tufts, C. Wieman, S. Salehi, y N. Haber, «Scaffold or Crutch? Examining College Students' Use and Views of Generative AI Tools for STEM Education», en *2025 IEEE Global Engineering Education Conference (EDUCON)*, abr. 2025, pp. 1-10. doi: 10.1109/EDUCON62633.2025.11016406.