

Influencia de una herramienta interactiva basada en Inteligencia Artificial en el rendimiento académico de la factorización

Influence of an Artificial Intelligence-Based Interactive Tool on Academic Performance in Factoring

Dayana Mishel Rocha Pazmiño

Universidad Técnica de Cotopaxi

dayana.rocha4206@utc.edu.ec

<https://orcid.org/0009-0009-9030-2777>

Latacunga - Ecuador

Jorge Cachuput Gusñay

Universidad Técnica de Cotopaxi

jorge.cachuput6058@utc.edu.ec

<https://orcid.org/0009-0002-7650-8065>

Latacunga - Ecuador

Formato de citación APA

Rocha, D. & Cachuput, J. (2026). Influencia de una herramienta interactiva basada en inteligencia artificial en el rendimiento académico en factorización. Revista REG, Vol. 5 (Nº. 3), p. 5026 - 5070.

CONOCIMIENTO SIN FRONTERAS

Vol. 5 (Nº. 3). julio – septiembre 2026.

ISSN: 3073-1259

Fecha de recepción: 30-07-2026

Fecha de aceptación :31-08-2026

Fecha de publicación:30-09-2026



RESUMEN

La factorización suele generar dificultades cuando el estudiantado memoriza reglas sin reconocer la estructura algebraica que orienta cada procedimiento. El objetivo fue analizar la influencia de una herramienta interactiva basada en inteligencia artificial en el rendimiento académico en factorización de estudiantes de noveno año de Educación General Básica Superior de la Unidad Educativa Particular “Gabriela Mistral”, en Latacunga, durante 2025-2026. Se desarrolló un estudio cuantitativo, aplicado y cuasi-experimental, con pretest y posttest en grupos no equivalentes. Participaron 43 estudiantes: 22 en el grupo control y 21 en el experimental. Durante dos semanas, el grupo experimental utilizó Gemini mediante el Gem Retromate para practicar cinco casos de factorización con orientación docente, retroalimentación inmediata y nuevos intentos. La prueba de conocimientos alcanzó un KR-20 de 0,81. El análisis incluyó estadística descriptiva, Shapiro-Wilk, Mann-Whitney para comparar grupos y Wilcoxon de rangos con signo para las mediciones relacionadas. Los grupos fueron comparables en el pretest ($W = 402,00$; $p = 0,205$; $r = 0,19$). En el posttest, el experimental obtuvo 8,83 y el control 4,67; la diferencia fue significativa y de magnitud amplia ($W = 243,50$; $p < 0,001$; $r = 0,81$). Ambos grupos mejoraron, aunque el cambio descriptivo fue mayor en el experimental. Los resultados respaldan una asociación favorable entre la práctica interactiva y el aprendizaje, dentro de una intervención breve, localizada y sin asignación aleatoria.

PALABRAS CLAVE: inteligencia artificial, rendimiento, factorización, matemática, retroalimentación.

ABSTRACT

Factoring often becomes difficult when students memorize rules without recognizing the algebraic structure that guides each procedure. This study analyzed the influence of an artificial intelligence-based interactive tool on ninth-grade students' academic performance in factoring at the "Gabriela Mistral" Private Educational Unit in Latacunga during the 2025-2026 school year. A quantitative, applied, quasi-experimental pretest-posttest design with nonequivalent groups was used. The participants were 43 students: 22 in the control group and 21 in the experimental group. For two weeks, the experimental group used Gemini through the Retromate Gem to practice five factoring cases with teacher guidance, immediate feedback, and repeated attempts. The knowledge test achieved a KR-20 coefficient of 0.81. Data analysis included descriptive statistics, the Shapiro-Wilk test, Mann-Whitney comparisons between groups, and Wilcoxon signed-rank tests for related measurements. The groups were comparable at pretest ($W = 402.00$; $p = .205$; $r = .19$). At posttest, the experimental group obtained 8.83 and the control group 4.67; the difference was significant and large ($W = 243.50$; $p < .001$; $r = .81$). Both groups improved, although the descriptive change was greater in the experimental group. The findings support a favorable association between interactive practice and factoring learning within a brief, single-site intervention without random assignment.

KEYWORDS: artificial intelligence, achievement, factoring, mathematics, feedback.

INTRODUCCIÓN

El aprendizaje de la matemática continúa planteando retos que no se resuelven únicamente con una mayor cantidad de ejercicios. Los resultados de PISA 2022 mostraron una disminución general del desempeño matemático en numerosos sistemas educativos y confirmaron que una proporción importante del estudiantado no alcanza el nivel básico requerido para resolver situaciones cotidianas (OECD, 2023). En Ecuador, la evaluación Ser Estudiante cumple una función semejante al examinar conocimientos y destrezas en los diferentes subniveles de Educación General Básica y Bachillerato, con el propósito de orientar decisiones frente a las brechas detectadas (Instituto Nacional de Evaluación Educativa [Ineval], 2025). Este escenario vuelve necesario revisar no solo qué contenidos se enseñan, sino también la forma en que se acompaña al estudiante cuando intenta comprenderlos.

Dentro del álgebra escolar, la factorización concentra varias dificultades porque exige reconocer la estructura de una expresión antes de elegir un procedimiento. No basta con recordar una fórmula: el estudiante debe observar términos y exponentes, identificar regularidades, distinguir el caso aplicable y comprobar que el producto de los factores reproduce el polinomio inicial. Cuando la enseñanza se limita a repetir modelos, es común que se confundan reglas parecidas, que se aplique un caso por semejanza superficial o que se abandone el ejercicio después del primer error. En noveno año estas dificultades se hacen más visibles, pues el lenguaje simbólico comienza a requerir mayor flexibilidad, control de los pasos y comprensión de las relaciones entre expresiones.

La inteligencia artificial educativa se ha presentado como una alternativa para ampliar la práctica, adaptar actividades y ofrecer información durante el proceso. Sin embargo, la literatura advierte que el campo ha crecido con mayor rapidez que sus fundamentos pedagógicos. Chen et al. (2020) identificaron una distancia entre las aplicaciones disponibles y las teorías que deberían orientar su uso, mientras que Zhang y Aslan (2021) señalaron que el valor de estas tecnologías depende de la articulación entre diseño, objetivos de aprendizaje y contexto. Por esta razón, la presencia de una herramienta inteligente no constituye por sí misma una innovación; su aporte aparece cuando responde a una necesidad concreta y se integra en una secuencia comprensible para el estudiante.

En educación matemática, las investigaciones han descrito a la inteligencia artificial como tutor, sistema de evaluación, apoyo para la resolución de problemas y recurso de personalización. Hwang y Tu (2021) encontraron que buena parte de los estudios se concentra en orientar al estudiante, seguir su desempeño y adaptar las actividades. La revisión de Mohamed et al. (2022) llegó a una conclusión semejante y añadió que los sistemas inteligentes se han utilizado para diagnosticar errores, recomendar contenidos y sostener procesos de práctica diferenciada. Más adelante, Panqueban y

Huincahue (2024) observaron un crecimiento de los estudios empíricos, aunque también señalaron la escasa presencia de marcos propios de la didáctica de la matemática y la necesidad de investigar en contextos escolares diversos.

La evidencia disponible es favorable, pero no uniforme. El metaanálisis de Hwang (2022) encontró un efecto positivo, de magnitud pequeña, de las aplicaciones de inteligencia artificial sobre el rendimiento matemático en estudiantes de educación elemental. En una revisión de sistemas de tutoría inteligente para educación primaria y secundaria, Létourneau et al. (2025) también identificaron resultados generalmente positivos, aunque las ventajas disminuyeron cuando la comparación se realizó con otros apoyos tutoriales bien diseñados. Estos hallazgos son importantes porque impiden atribuir el aprendizaje únicamente a la tecnología: la duración de la intervención, la calidad de las tareas, el acompañamiento docente y la posibilidad de corregir errores modifican los resultados.

Los chatbots educativos ofrecen un antecedente cercano para comprender la función conversacional de las herramientas generativas. Kuhail et al. (2023) mostraron que estos sistemas suelen actuar como agentes de enseñanza o apoyo entre pares y que su efectividad está relacionada con el diseño de la interacción, la personalización y la claridad de las rutas de aprendizaje. En matemática, Lee y Yeo (2022) desarrollaron un chatbot orientado a la práctica de enseñanza responsiva y evidenciaron que el diálogo con el sistema puede favorecer la revisión de decisiones y la consideración de distintas respuestas. Shin (2022), por su parte, subrayó que la integración de sistemas tutoriales requiere conocimiento tecnológico, pedagógico y disciplinar, pues una respuesta automatizada solo es útil cuando se interpreta desde el contenido que se enseña.

La expansión de los modelos generativos amplió estas posibilidades al permitir explicaciones en lenguaje natural, preguntas de seguimiento y retroalimentación inmediata. Lo (2023) advierte, no obstante, que estas herramientas pueden ofrecer información imprecisa, respuestas plausibles pero incorrectas o ayudas excesivas que reduzcan el esfuerzo cognitivo del estudiante. En el campo matemático, Wardat et al. (2023) reconocen su utilidad para proponer ejemplos, aclarar procedimientos y estimular la discusión, pero insisten en que las soluciones deben verificarse. Esta cautela resulta especialmente pertinente en factorización: una explicación verbalmente clara puede contener un error de signo, omitir una condición o llevar al estudiante a copiar pasos sin comprender la estructura algebraica.

La calidad educativa de una respuesta generada tampoco depende únicamente de que el resultado final sea correcto. Schorcht et al. (2024) comprobaron que el modo de formular la instrucción

influye en la estrategia, las representaciones y la claridad de las soluciones matemáticas producidas por la inteligencia artificial. Un estudio posterior mostró que los distintos modelos y técnicas de prompting no ofrecen una calidad pedagógica uniforme, por lo que la selección del recurso y la revisión humana siguen siendo necesarias (Schorcht et al., 2026). En consecuencia, una herramienta para factorización debe orientar la identificación del patrón, pedir al estudiante que justifique su elección, mostrar el punto exacto del error y permitir otro intento antes de entregar la solución completa.

Desde una perspectiva educativa, la retroalimentación inmediata tiene sentido cuando ayuda a reorganizar el razonamiento y no cuando reemplaza la actividad del estudiante. Gabriel et al. (2025) sostienen que la inteligencia artificial adquiere valor pedagógico cuando transforma la interacción entre estudiante, contenido y docente, fortalece la percepción de competencia y abre oportunidades para aprender a partir del error. Esta idea coincide con la guía de Miao y Holmes (2023), que propone una integración centrada en las personas, con supervisión docente, protección de datos y tareas que mantengan activo el pensamiento. Por ello, la mediación humana no es un complemento opcional, sino la condición que permite revisar las respuestas del sistema y conectar la ayuda tecnológica con los objetivos de la clase.

En la Unidad Educativa Particular “Gabriela Mistral”, los resultados iniciales mostraron un dominio bajo de los casos de factorización en los dos grupos participantes. La media del pretest fue 3,00 en el grupo control y 3,52 en el experimental. A partir de este diagnóstico se consideró pertinente ampliar las oportunidades de práctica y ofrecer devoluciones en el momento en que aparecía el error, sin retirar la explicación docente ni modificar los contenidos previstos para noveno año. La necesidad no consistía en sustituir la clase habitual, sino en incorporar un apoyo que permitiera ensayar procedimientos, contrastar respuestas y volver sobre los pasos que todavía no estaban claros.

La herramienta seleccionada fue Gemini mediante un Gem denominado Retromate, configurado para trabajar factor común, agrupación de términos, diferencia de cuadrados, suma y diferencia de cubos y trinomio cuadrado perfecto. Su uso se organizó dentro de una secuencia de clases que combinó explicación, práctica guiada, interacción con la herramienta, revisión de errores y ejercicios de cierre. De este modo, el recurso no funcionó como una actividad aislada ni como fuente automática de respuestas, sino como un apoyo para que el estudiante reconociera el caso, explicara el procedimiento y comprobara el resultado.

Con base en esta problemática, la pregunta de investigación fue: ¿qué cambios se observan en el rendimiento académico en factorización cuando una herramienta interactiva basada en inteligencia artificial se incorpora a la práctica de estudiantes de noveno año? El objetivo fue analizar

la influencia del uso de esta herramienta en el rendimiento académico de los estudiantes de Educación General Básica Superior de la Unidad Educativa Particular “Gabriela Mistral”, en Latacunga, durante el año lectivo 2025-2026. Debido al carácter cuasi-experimental y a la ausencia de asignación aleatoria, los resultados se interpretaron como evidencia de asociación dentro de las condiciones estudiadas y no como una demostración causal definitiva.

MÉTODOS MATERIALES

La investigación se desarrolló desde un enfoque cuantitativo, puesto que el rendimiento académico se representó mediante calificaciones y se compararon los resultados obtenidos antes y después de la intervención. El estudio fue aplicado, debido a que utilizó una herramienta interactiva para atender una dificultad concreta del aula, y tuvo un alcance explicativo limitado por las condiciones de un diseño cuasi-experimental. Se trabajó con pretest y posttest en grupos no equivalentes. Los paralelos ya se encontraban conformados por la institución y no hubo asignación aleatoria; uno continuó con la enseñanza habitual y funcionó como grupo control, mientras el otro incorporó la herramienta basada en inteligencia artificial y constituyó el grupo experimental.

La población estuvo integrada por 43 estudiantes de noveno año de Educación General Básica Superior de la Unidad Educativa Particular “Gabriela Mistral”, ubicada en Latacunga. El grupo control reunió 22 estudiantes y el experimental 21. Se trabajó con toda la población disponible, por lo que el procedimiento correspondió a un censo del grupo accesible y no a una selección probabilística. En el posttest del grupo control se registró una ausencia; por esta razón, esa medición se analizó con 21 participantes, mientras los demás conjuntos conservaron el número de casos señalado en los resultados. Esta pérdida se consideró al interpretar las comparaciones y constituye una limitación del estudio.

El instrumento fue una prueba de conocimientos compuesta por diez reactivos sobre los casos de factorización previstos para noveno año. Los reactivos evaluaron el reconocimiento del caso, la aplicación del procedimiento y la obtención del resultado. La validez de contenido fue revisada por tres expertos, quienes valoraron claridad, pertinencia, coherencia y correspondencia entre los reactivos y el contenido evaluado. No se calculó un índice cuantitativo de acuerdo entre jueces; por ello, la validación se reporta como juicio de expertos y no se atribuye un coeficiente que no haya sido obtenido.

La salida de R mostró un alfa de Cronbach de 0,81 y un coeficiente KR-20 de 0,8091533. Debido a que los diez reactivos tuvieron puntuación dicotómica, se priorizó el KR-20 y se reportó redondeado

a dos decimales ($KR-20 = 0,81$). Este resultado respalda una consistencia interna adecuada para el uso previsto. Foster (2021) explica que la elección del estimador debe corresponder a la estructura de los datos y que coeficientes prácticamente equivalentes no deben interpretarse como evidencias independientes.

El pretest se aplicó a los dos grupos antes de iniciar la secuencia didáctica. Durante dos semanas, el grupo experimental utilizó Gemini mediante el Gem Retromate. Las sesiones abordaron de forma progresiva el factor común, el factor común por agrupación de términos, la diferencia de cuadrados perfectos, la suma y diferencia de cubos perfectos y el trinomio cuadrado perfecto. Cada clase combinó explicación docente, resolución guiada, práctica con la herramienta, revisión de errores y ejercicios de cierre. El grupo control trabajó los mismos contenidos mediante la metodología habitual, sin utilizar el recurso interactivo. Al finalizar el período se aplicó el postest en ambos grupos.

La función prevista para Retromate fue acompañar la práctica y no resolver automáticamente la tarea. El estudiante debía identificar el caso, proponer un procedimiento y revisar la devolución recibida. Cuando aparecía un error, la interacción permitía volver sobre el paso problemático y realizar otro intento. El docente mantuvo la responsabilidad de explicar los contenidos y verificar las respuestas. Esta forma de incorporación coincide con la literatura que recomienda utilizar los chatbots como apoyo tutorial y mantener supervisión humana sobre la precisión y pertinencia de la retroalimentación (Kuhail et al., 2023; Lo, 2023).

El procesamiento se realizó en R, versión 4.5.1 (R Core Team, 2025). Primero se calcularon el número de casos válidos, la media, la desviación estándar, la mediana y los valores mínimo y máximo. La normalidad se examinó mediante Shapiro-Wilk. Como las mediciones del grupo experimental no se ajustaron a una distribución normal, se adoptó una estrategia no paramétrica y se fijó el nivel de significancia en $\alpha = 0,05$.

Las comparaciones entre grupos se realizaron con la prueba U de Mann-Whitney: primero en el pretest, para valorar la comparabilidad inicial, y después en el postest, para examinar la diferencia al finalizar la intervención. Los cambios dentro de cada grupo se analizaron mediante la prueba de rangos con signo de Wilcoxon, adecuada para mediciones relacionadas. Además de los valores p, se calculó el tamaño del efecto r a partir del estadístico Z mediante $r = |Z|/\sqrt{N}$; en las comparaciones pareadas, N correspondió al número de pares completos. Esta combinación permitió distinguir las diferencias entre grupos de los cambios producidos dentro de cada grupo.

ANÁLISIS DE RESULTADOS

El análisis comenzó con la descripción de las calificaciones para reconocer el comportamiento de los grupos antes de interpretar los contrastes. La Tabla 1 presenta el número de casos válidos, la media, la desviación estándar, la mediana y los valores extremos de cada medición.

Tabla 1.

Estadísticos descriptivos del rendimiento académico en factorización

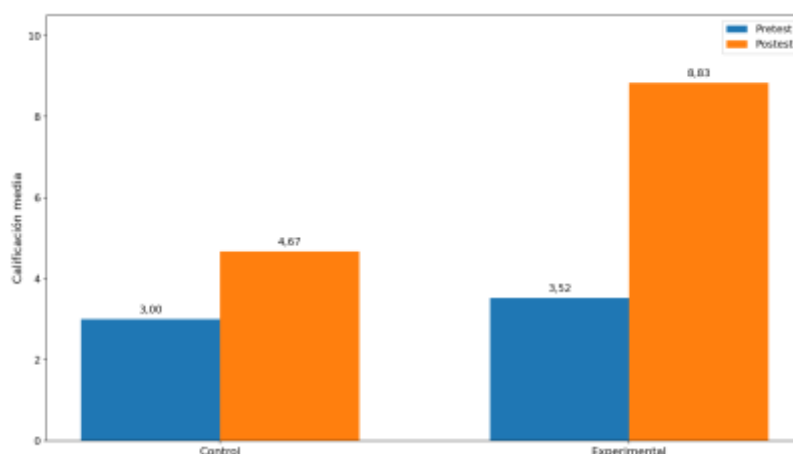
Grupo	Momento	n	Media	DE	Mediana	Mínimo	Máximo
Control	Pretest	22	3,00	0,90	3,00	0,50	4,50
Control	Posttest	21	4,67	0,89	4,50	3,00	6,50
Experimental	Pretest	21	3,52	1,06	3,50	2,00	6,50
Experimental	Posttest	21	8,83	1,28	9,00	4,50	10,00

Nota. DE = desviación estándar. Datos procesados en R a partir de las calificaciones proporcionadas.

En el pretest, la diferencia entre las medias fue de 0,52 puntos: 3,00 en el grupo control y 3,52 en el experimental. Después de las dos semanas de enseñanza, ambos grupos aumentaron su calificación. El control llegó a 4,67, lo que representa un incremento descriptivo de 1,67 puntos, mientras el experimental alcanzó 8,83 y aumentó 5,31 puntos. La diferencia entre ambas ganancias medias fue de 3,64 puntos a favor del grupo que utilizó Retromate. La mediana experimental pasó de 3,50 a 9,00; en el control cambió de 3,00 a 4,50. Estas cifras muestran un desplazamiento más amplio en el grupo experimental, aunque no sustituyen el análisis de la variabilidad individual.

Figura 1.

Comparación de las medias del pretest y posttest por grupo



Nota. La figura representa las medias informadas en el análisis descriptivo; las etiquetas usan coma decimal para mantener consistencia con las tablas.

Comprobación de la normalidad

La inspección de histogramas, curvas de densidad, diagramas de caja y gráficos Q-Q mostró un ajuste desigual a la distribución normal. Esta revisión gráfica se complementó con la prueba de Shapiro-Wilk, cuyos resultados se presentan en la Tabla 2.

Tabla 2.

Resultados de la prueba de normalidad de Shapiro-Wilk

Grupo	Momento	p	Decisión
Control	Pretest	0,144	Distribución compatible con normalidad
Control	Postest	0,695	Distribución compatible con normalidad
Experimental	Pretest	0,017	Distribución no normal
Experimental	Postest	< 0,001	Distribución no normal

Nota. Se consideró $\alpha = 0,05$.

Los puntajes del grupo control fueron compatibles con una distribución normal en los dos momentos. En el grupo experimental, en cambio, el pretest y el postest presentaron valores inferiores a 0,05. Debido a que el supuesto no se cumplió en todas las mediciones y a que los tamaños de los grupos fueron reducidos, se continuó con procedimientos no paramétricos.

Comparaciones entre grupos

La prueba de Mann-Whitney permitió comparar a los grupos sin asumir normalidad. Se aplicó en el pretest para examinar la equivalencia inicial y en el postest para valorar la diferencia al finalizar la secuencia didáctica.

Tabla 3.

Comparaciones entre grupos mediante la prueba de Mann-Whitney

Momento	W	Z	p	r	Resultado
Pretest	402,00	-1,269	0,205	0,19	No significativa
Postest	243,50	-5,258	< 0,001	0,81	Significativa

Nota. W corresponde a la suma de rangos reportada por el análisis. El tamaño del efecto se calculó mediante $r = |Z|/\sqrt{N}$; para el pretest $N = 43$ y para el postest $N = 42$.

En el pretest no se observó una diferencia estadísticamente significativa entre el grupo control y el experimental ($W = 402,00$; $Z = -1,269$; $p = 0,205$; $r = 0,19$). El tamaño del efecto fue pequeño, lo que respalda la comparabilidad inicial de los grupos sin afirmar que fueran idénticos. En el postest, la diferencia fue significativa y de magnitud amplia ($W = 243,50$; $Z = -5,258$; $p < 0,001$; $r = 0,81$). Este resultado coincide con las medias finales: 4,67 en el grupo control y 8,83 en el experimental.

Cambios dentro de cada grupo

Para examinar la evolución entre pretest y posttest se aplicó la prueba de rangos con signo de Wilcoxon a los estudiantes que contaban con ambas mediciones. La Tabla 4 resume los contrastes intragrupo.

Tabla 4.

Cambios intragrupo mediante la prueba de rangos con signo de Wilcoxon

Grupo	n de pares	Z	p	r	Resultado
Control	21	-3,677	< 0,001	0,80	Significativa
Experimental	21	-4,024	< 0,001	0,88	Significativa

Nota. El tamaño del efecto se calculó mediante $r = |Z|/\sqrt{n}$, donde n fue el número de pares completos de cada grupo.

Los dos grupos presentaron cambios significativos entre el pretest y el posttest. En el control, la media aumentó de 3,00 a 4,67 ($Z = -3,677$; $p < 0,001$; $r = 0,80$); en el experimental pasó de 3,52 a 8,83 ($Z = -4,024$; $p < 0,001$; $r = 0,88$). Las ganancias medias fueron de 1,67 y 5,31 puntos, respectivamente. Esta diferencia de ganancias es descriptiva, porque los resultados disponibles no incluyen un contraste inferencial directo de los puntajes de cambio. En conjunto, la equivalencia inicial, la diferencia amplia en el posttest y la evolución intragrupo forman un patrón consistente con un aporte favorable de la herramienta, sin convertir el diseño cuasi-experimental en una demostración causal absoluta.

DISCUSIÓN

Los resultados muestran un patrón favorable al grupo que utilizó la herramienta interactiva. En el pretest no se encontraron diferencias significativas entre control y experimental ($p = 0,205$; $r = 0,19$), mientras que en el posttest la diferencia fue significativa y de magnitud amplia ($p < 0,001$; $r = 0,81$). El grupo experimental pasó de una media de 3,52 a 8,83 y el control avanzó de 3,00 a 4,67. Las pruebas de Wilcoxon confirmaron cambios dentro de los dos grupos, lo que indica que ambos aprendieron durante el período de enseñanza; aun así, la amplitud descriptiva del cambio y la separación final fueron mayores en el experimental. Esta lectura es más precisa que atribuir toda mejora a la inteligencia artificial, porque distingue el efecto de la enseñanza compartida del posible aporte adicional de Retromate.

El hallazgo coincide con las revisiones que atribuyen a la inteligencia artificial una función tutorial, de seguimiento y adaptación de la práctica matemática (Hwang & Tu, 2021; Mohamed et al., 2022). También guarda relación con el metaanálisis de Hwang (2022), que identificó un efecto positivo, aunque pequeño, sobre el rendimiento matemático. La ganancia descriptiva observada en esta investigación fue mayor que la tendencia promedio descrita en ese metaanálisis; sin embargo, la

comparación debe realizarse con cautela porque aquí se trabajó con una población pequeña, una intervención corta y un contenido específico.

Una explicación posible se encuentra en el momento de la retroalimentación. En factorización, un error temprano puede alterar todo el procedimiento: si el estudiante identifica de manera equivocada el caso, los pasos posteriores dejan de responder a la estructura del polinomio. La posibilidad de recibir una indicación durante la práctica, revisar el punto problemático y volver a intentarlo pudo reducir la repetición de procedimientos incorrectos. Esta función coincide con la evidencia sobre tutores inteligentes y chatbots educativos, donde los mejores resultados aparecen cuando el apoyo responde al progreso del estudiante y no se limita a presentar contenido digital (Kuhail et al., 2023; Létourneau et al., 2025).

La interacción conversacional también pudo contribuir a que el estudiante explicara qué estaba haciendo. Cuando una herramienta solicita reconocer el patrón, justificar la elección o comprobar la multiplicación final, el procedimiento deja de ser una secuencia silenciosa de símbolos y se convierte en una actividad que puede revisarse. Lee y Yeo (2022) mostraron que un chatbot puede favorecer el análisis de decisiones cuando propone situaciones abiertas y obliga a considerar respuestas alternativas. En este estudio no se registraron las conversaciones individuales, por lo que no es posible afirmar qué tipo de ayuda fue la más efectiva; aun así, la estructura prevista de Retromate buscó sostener ese diálogo antes de entregar una solución completa.

La calidad de la ayuda generada requiere una lectura crítica. Wardat et al. (2023) reconocen que los modelos conversacionales pueden aclarar procedimientos matemáticos, pero también producir errores o explicaciones que parecen convincentes sin ser correctas. Schorcht et al. (2024) demostraron que la formulación de los prompts influye en la estrategia y en la calidad educativa de las soluciones; su estudio posterior confirmó que no existe una combinación de modelo e instrucción que funcione de la misma manera para todas las tareas (Schorcht et al., 2026). Por ello, el resultado favorable de una herramienta configurada para factorización no debería generalizarse a cualquier uso libre de Gemini ni a cualquier contenido matemático.

La intervención no sustituyó la explicación del docente. Retromate se utilizó dentro de sesiones que combinaron orientación, práctica, revisión de errores y evaluación formativa. Este punto es relevante porque una respuesta inmediata puede convertirse en una ayuda superficial si el estudiante se limita a copiarla. Shin (2022) sostiene que la integración de sistemas inteligentes en matemática exige articular tecnología, pedagogía y conocimiento disciplinar. De manera semejante, Gabriel et al. (2025) plantean que la inteligencia artificial es pedagógicamente valiosa cuando

transforma la interacción con el contenido y fortalece la percepción de competencia, no cuando desplaza la responsabilidad de pensar.

El aumento del grupo control merece una interpretación propia. La prueba de Wilcoxon mostró un cambio significativo también en este grupo ($Z = -3,677$; $p < 0,001$; $r = 0,80$), de modo que la enseñanza habitual produjo aprendizaje. En el experimental, el cambio fue igualmente significativo y presentó un tamaño del efecto algo mayor ($Z = -4,024$; $p < 0,001$; $r = 0,88$). Por tanto, el estudio no enfrenta una intervención contra la ausencia de enseñanza, sino dos formas de trabajar el mismo contenido. La comparación posttest mediante Mann-Whitney aporta la evidencia más clara a favor del grupo experimental, aunque el aporte específico de la herramienta debe interpretarse junto con la mediación docente y la planificación desarrollada.

La distribución de los puntajes muestra además que el avance no fue homogéneo. En el posttest experimental se registró un mínimo de 4,50, mientras la mediana alcanzó 9,00 y el máximo 10,00. Esto significa que, aunque la mayoría obtuvo calificaciones altas, al menos un estudiante mantuvo dificultades importantes. También puede existir un efecto de techo, porque varios puntajes se acercaron al límite de la escala. La personalización tecnológica no elimina por sí sola las diferencias individuales ni garantiza que todos comprendan el procedimiento con la misma profundidad. Los estudiantes con bajo rendimiento final requieren una revisión específica de errores y apoyos adicionales.

Las limitaciones metodológicas condicionan el alcance de las conclusiones. La investigación se realizó en una sola institución, con 43 estudiantes y grupos previamente conformados. No se controlaron mediante asignación aleatoria las diferencias que podían existir entre los paralelos, la intervención duró únicamente dos semanas y no se aplicó una medición diferida que permitiera comprobar la permanencia del aprendizaje. Una ausencia redujo el posttest control a 21 casos. Aunque se estimaron tamaños del efecto r , no se calcularon intervalos de confianza ni se dispuso de información sobre frecuencia de uso, calidad de los prompts, conversaciones generadas o errores producidos por el sistema.

La estrategia inferencial se fortaleció al diferenciar las comparaciones entre grupos de las mediciones relacionadas: Mann-Whitney respetó la independencia entre control y experimental, mientras Wilcoxon conservó el emparejamiento pretest-posttest. No obstante, los resultados disponibles no incluyen una comparación directa de las ganancias individuales ni un modelo que estime la interacción entre grupo y momento. Esta ausencia impide cuantificar con total precisión cuánto del cambio diferencial corresponde a la intervención. Una nueva investigación debería

contrastar los puntajes de ganancia o utilizar un modelo de medidas repetidas, además de informar intervalos de confianza y conservar la identificación anónima de cada par de observaciones.

Desde la práctica educativa, los resultados respaldan una integración gradual y supervisada. Una herramienta de este tipo puede utilizarse para ampliar ejercicios, ofrecer pistas, preguntar por el caso elegido y pedir comprobaciones; no debería convertirse en una fuente incuestionable de respuestas. El docente necesita revisar las indicaciones, establecer reglas de uso y evitar que se compartan datos personales innecesarios. La guía de Miao y Holmes (2023) insiste en una adopción centrada en las personas y en la protección de datos, mientras Lo (2023) advierte sobre respuestas inexactas y dependencia. En estudiantes menores de edad, estas precauciones adquieren especial importancia.

Los estudios posteriores deberían ampliar el tiempo de aplicación, incorporar más instituciones y registrar el proceso de interacción, no solo la calificación final. También sería útil comparar tipos de retroalimentación, analizar qué errores se corrigen con mayor frecuencia y valorar comprensión conceptual, autonomía y transferencia a ejercicios nuevos. De esta manera se podría determinar si el rendimiento alto responde a una mejor identificación de patrones o a una práctica intensa de ejercicios semejantes. La solidez de la evidencia aumentará cuando los resultados cuantitativos se acompañen de información sobre cómo razonan los estudiantes y cómo interviene el docente durante el uso de la herramienta.

CONCLUSIONES

La incorporación de Retromate estuvo asociada con un rendimiento final mayor en factorización dentro de la experiencia analizada. Los grupos presentaron resultados iniciales comparables ($p = 0,205$; $r = 0,19$) y, al finalizar, el grupo experimental obtuvo una media de 8,83 frente a 4,67 en el control. La comparación posttest fue significativa y mostró un tamaño del efecto amplio ($p < 0,001$; $r = 0,81$).

Los dos grupos mejoraron significativamente entre el pretest y el posttest, lo que confirma que hubo aprendizaje durante las dos semanas. Sin embargo, el cambio descriptivo fue mayor en el experimental: 5,31 puntos frente a 1,67 en el control. El posible aporte de la herramienta se relaciona con la retroalimentación inmediata, la oportunidad de realizar nuevos intentos y la orientación para reconocer el caso de factorización, siempre dentro de una secuencia dirigida por el docente.

Los hallazgos deben leerse dentro de una intervención breve, desarrollada en una sola institución y con grupos no aleatorizados. Mann-Whitney y Wilcoxon mejoran la correspondencia entre el análisis y el diseño, pero no reemplazan una comparación directa de ganancias o un modelo grupo por momento. Por ello, la conclusión más sólida es que los datos son consistentes con una asociación favorable entre Retromate y el rendimiento en factorización, no que la herramienta explique por sí sola todos los cambios observados.

REFERENCIAS BIBLIOGRÁFICAS

- Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, Article 100002. <https://doi.org/10.1016/j.caeai.2020.100002>
- Foster, R. C. (2021). KR20 and KR21 for some nondichotomous data (It's not just Cronbach's alpha). *Educational and Psychological Measurement*, 81(6), 1172–1202. <https://doi.org/10.1177/0013164421992535>
- Gabriel, F., Kennedy, J., Marrone, R., & Leonard, S. (2025). Pragmatic AI in education and its role in mathematics learning and teaching. *npj Science of Learning*, 10, Article 26. <https://doi.org/10.1038/s41539-025-00315-4>
- Hwang, G.-J., & Tu, Y.-F. (2021). Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Mathematics*, 9(6), Article 584. <https://doi.org/10.3390/math9060584>
- Hwang, S. (2022). Examining the effects of artificial intelligence on elementary students' mathematics achievement: A meta-analysis. *Sustainability*, 14(20), Article 13185. <https://doi.org/10.3390/su142013185>
- Instituto Nacional de Evaluación Educativa. (2025, 24 de abril). *Ineval presentó los resultados de la evaluación nacional Ser Estudiante 2024*. <https://www.evaluacion.gob.ec/ineval-presento-los-resultados-de-la-evaluacion-nacional-ser-estudiante-2024/>
- Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28, 973–1018. <https://doi.org/10.1007/s10639-022-11177-3>
- Lee, D., & Yeo, S. (2022). Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers & Education*, 191, Article 104646. <https://doi.org/10.1016/j.compedu.2022.104646>
- Létourneau, A., Deslandes Martineau, M., Charland, P., Karan, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10, Article 29. <https://doi.org/10.1038/s41539-025-00320-7>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), Article 410. <https://doi.org/10.3390/educsci13040410>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>

- Mohamed, M. Z. B., Hidayat, R., Suhaizi, N. N. B., Sabri, N. B. M., Mahmud, M. K. H. B., & Baharuddin, S. N. B. (2022). Artificial intelligence in mathematics education: A systematic literature review. *International Electronic Journal of Mathematics Education*, 17(3), Article e0694. <https://doi.org/10.29333/iejme/12132>
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- Panqueban, D., & Huincahue, J. (2024). Artificial intelligence in mathematics education: A systematic review. *Uniciencia*, 38(1), 1–17. <https://doi.org/10.15359/ru.38-1.20>
- R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem: Investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, 9, Article 1386075. <https://doi.org/10.3389/feduc.2024.1386075>
- Schorcht, S., Müller, F. A., & Buchholtz, N. (2026). No one-size-fits-all: A study of prompt techniques and large language models to enhance AI's mathematics educational quality. *ZDM—Mathematics Education*. Advance online publication. <https://doi.org/10.1007/s11858-026-01784-6>
- Shin, D. (2022). Teaching mathematics integrating intelligent tutoring systems: Investigating prospective teachers' concerns and TPACK. *International Journal of Science and Mathematics Education*, 20, 1659–1676. <https://doi.org/10.1007/s10763-021-10221-x>
- Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), Article em2286. <https://doi.org/10.29333/ejmste/13272>
- Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research and future directions. *Computers and Education: Artificial Intelligence*, 2, Article 100025. <https://doi.org/10.1016/j.caeai.2021.100025>
- Chías, M., & Zurita, J. (2009). *EmocionArte con los niños: el arte de acompañar a los niños en su emoción* (2a ed.). Editorial Desclée de Brouwer.
- Crespin Tomala, B. F. (2026). *La dramatización en el desarrollo de la conciencia emocional en niños de 4 a 5 años* (Trabajo de titulación, 75p.). Universidad Estatal Península de Santa Elena, Sede Playas.

- Gonzabay Tomalá, E. L., & Loor Salinas, A. S. (2024). El juego dramático en el desarrollo de la expresión oral en niños de 4 a 5 años. Universidad Estatal Península de Santa Elena.
- Pallarés, M. (2010). Emociones y sentimientos: dónde se forman y cómo se transforman. Marge Books.
- Valdez Banchon, C. S. (2026). Estrategias lúdicas para la autorregulación emocional en niños de 5 a 6 años. Universidad Estatal Península de Santa Elena.
- Yagual, R., & Stephany, A. (2022). El arte: un factor de resiliencia en niños de 4 a 5 años. Universidad Estatal Península de Santa Elena.
- Bisquerra, R. (2009). Educación emocional y bienestar
Bisquerra, R. (2009). Educación emocional y bienestar. Wolters Kluwer España.
- Goleman, D. (1996). Inteligencia emocional
Goleman, D. (1996). Inteligencia emocional. Kairós.
- Gardner, H. (2011). Estructuras de la mente: La teoría de las inteligencias múltiples
Gardner, H. (2011). Estructuras de la mente: La teoría de las inteligencias múltiples (3.ª ed.). Fondo de Cultura Económica.
- Motos, T., & Navarro, A. (2003). El teatro en la educación
Motos, T., & Navarro, A. (2003). El teatro en la educación. Octaedro.
- Denham, S. A. (2007). Dealing with Feelings: How Children Negotiate the Worlds of Emotions and Social Relationships
- Denham, S. A. (2007). Dealing with Feelings: How Children Negotiate the Worlds of Emotions and Social Relationships. Psychology Press.

CONFLICTO DE INTERÉS:

Los autores declaran que no existen conflicto de interés posibles

FINANCIAMIENTO

No existió asistencia de financiamiento de parte de pares externos al presente artículo.

NOTA:

El artículo no es producto de una publicación anterior.

