Qué cuenta como evidencia de aprendizaje con IA

Por José Luis García Boyé[i]
Secretario Técnico de RIED

Palabras clave: RIED, inteligencia artificial generativa, aprendizaje, evidencia, publicación científica, evaluación

Publicado el 09 de octubre de 2026

Hace tres años, que un manuscrito tratara sobre ChatGPT bastaba casi como novedad; ya no. Quienes trabajamos en la edición de revistas de educación recibimos cada vez más trabajos sobre inteligencia artificial generativa (IAG), y cuando uno de ellos afirma que la IA mejora el aprendizaje, la primera pregunta que debemos hacernos es: «¿qué ha medido realmente?»

Bajo la palabra aprendizaje caben resultados muy distintos. Un estudio puede mostrar que los estudiantes resuelven más ejercicios con IA, que escriben mejores textos o que creen haber aprendido más, y cada uno de esos hallazgos responde a una pregunta diferente.

Las revisiones recientes encuentran, en conjunto, efectos favorables de la IAG sobre distintos resultados educativos, aunque con diseños y medidas muy dispares (Deng et al., 2025; Han et al., 2025). Lo que puede concluirse sobre

aprendizaje depende, en cada caso, de qué se haya medido y del alcance del diseño.

¿Y si apagar la herramienta cambia el resultado?

El experimento de Bastani et al. (2025) muestra especialmente bien el problema. Cerca de mil estudiantes de secundaria resolvieron problemas de matemáticas en distintas condiciones, dos de ellas con herramientas basadas en GPT-4: GPT Base y GPT Tutor. Durante la práctica, las calificaciones subieron aproximadamente un 48% con la primera y un 127% con la segunda, frente al grupo de control.

Después, en la fase de examen, la IA dejó de estar disponible.

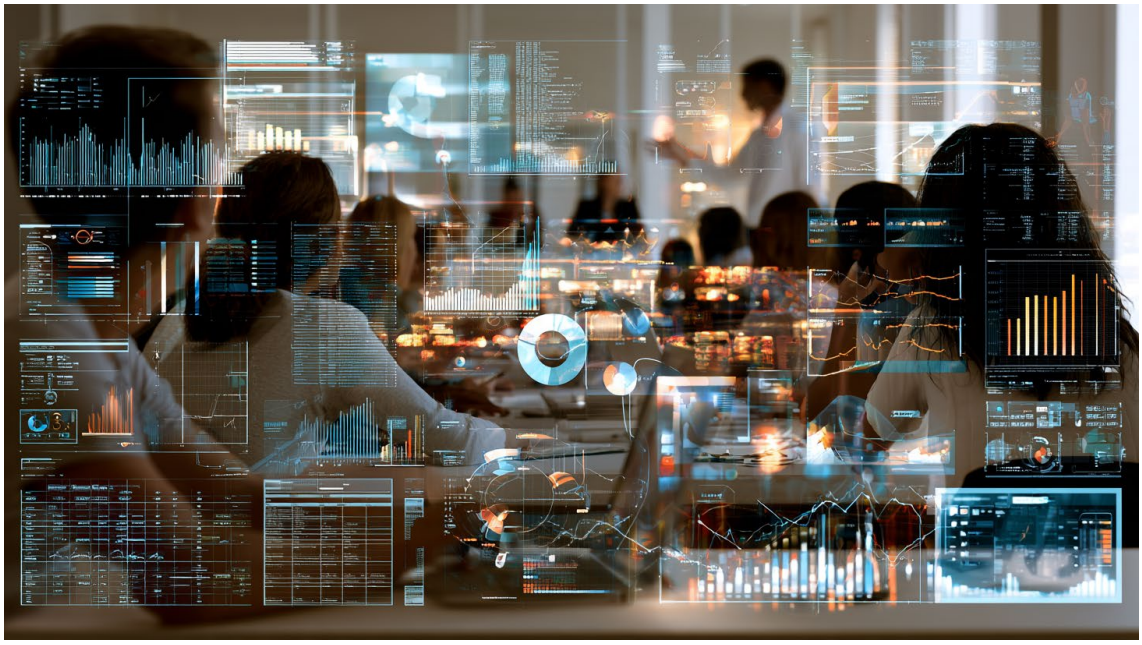
Con GPT Base, los estudiantes obtuvieron resultados aproximadamente un 17% inferiores a los del grupo de control; con GPT Tutor, diseñado para orientar y dificultar el acceso directo a las soluciones, ese perjuicio prácticamente desapareció, aunque tampoco hubo una ventaja significativa frente al control (Bastani et al., 2025).

Si la recogida de datos hubiera terminado antes del examen, habría evidencia sólida de que GPT mejoraba el rendimiento durante la práctica, y habría sido mucho más complicado saber cuánto habían aprendido realmente los estudiantes. Esa diferencia debería notarse en el título, el resumen y las conclusiones.

El experimento no demuestra que la IA perjudique el aprendizaje. Kestin et al. (2025) encontraron algo muy distinto en un ensayo aleatorizado con estudiantes universitarios de física, donde un tutor de IA diseñado a partir de principios de las ciencias del aprendizaje produjo mayores ganancias que una clase presencial de aprendizaje activo, y en menos tiempo. La ventaja apareció en el postest realizado inmediatamente después de la lección. El estudio no permite saber si seguía presente semanas o meses más tarde.

Los resultados de ambos estudios no son incompatibles. La IAG puede mejorar mucho el rendimiento durante una actividad sin dejar un beneficio equivalente después, y también puede integrarse en intervenciones que producen mejores resultados posteriores. Han et al. (2025), por su parte, reunieron 68 estudios experimentales y cuasiexperimentales y hallaron un efecto medio favorable, pero con una heterogeneidad extraordinariamente alta (I^2 = 95 %). Decir simplemente

que «la IA mejora el aprendizaje» dice demasiado poco sobre intervenciones tan diferentes.



Fuente de la imagen: imagen generada con Midjourney

Medir una cosa y concluir otra

Antes de discutir si hubo aprendizaje, conviene identificar qué variable se ha estudiado realmente. La calidad de un texto, el rendimiento mientras la IA está disponible, la percepción de aprendizaje, un postest inmediato, la retención y la transferencia no son intercambiables entre sí.

Un estudio basado en percepciones puede ser un excelente estudio de percepciones. Si los participantes dicen que una herramienta les ayudó a aprender, sabemos algo sobre su experiencia; pero para afirmar que aumentaron sus conocimientos necesitamos otra medida. Lo mismo ocurre con los productos: que un texto escrito con ayuda de IA sea mejor no demuestra por sí solo que haya mejorado en la misma medida la competencia de escritura del estudiante.

La evidencia sobre retención y transferencia es todavía mucho menor que la dedicada a efectos inmediatos. El informe *Perspectivas de la educación digital de la OCDE 2026*, de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), reclama investigación longitudinal y métodos de evaluación capaces de distinguir el rendimiento durante una tarea de cambios duraderos en conocimientos y capacidades (OCDE, 2026). Un trabajo breve

puede limitarse a estudiar efectos inmediatos, siempre que sus conclusiones conserven ese mismo alcance.

La dificultad con la IAG es que la herramienta interviene a menudo en la propia tarea que después utilizamos para medir el resultado. Cuando estudiante y sistema producen juntos la respuesta, evaluar solo el producto final puede decirnos mucho sobre esa combinación y poco sobre qué ha adquirido el estudiante.

«Usamos ChatGPT» tampoco describe una intervención

Dos estudios pueden afirmar que utilizaron ChatGPT y, aun así, haber creado situaciones educativas completamente diferentes. En Bastani et al. (2025), GPT Base y GPT Tutor se apoyaban en GPT-4, pero cambiaban las instrucciones, las restricciones y la manera de responder al estudiante. Una configuración facilitaba llegar a la solución; la otra trataba de mantenerlo trabajando sobre el problema. Conocer el modelo empleado es solo una parte de la descripción.

El problema ya ha generado propuestas concretas para mejorar la publicación de estos estudios. Allison (2026) plantea el marco RAISE para aumentar la transparencia con la que se comunican las investigaciones educativas con IA, mientras que Hwang et al. (2026), con el modelo TEP-AIED, ponen el foco en transparencia, ética y pedagogía. En ambos casos se reclama información suficiente para entender qué sistema se utilizó, cómo se integró en la actividad y qué papel desempeñó.

En RIED 29(1), Alier-Forment et al. (2026) formularon siete principios para la adopción segura de la IAG y los concretaron en LAMB (*Learning Assistant Manager and Builder*), un entorno abierto para construir asistentes de aprendizaje contextualizados y bajo control docente. La propuesta explicita el papel de la IA, sus límites y el control pedagógico sobre su funcionamiento.

También importa saber qué parte del trabajo siguió haciendo el estudiante. Una herramienta puede descargar operaciones accesorias y permitir concentrarse en tareas de mayor nivel; la interpretación cambia si descarga precisamente el razonamiento que se pretendía ejercitar. Describir esa división del trabajo forma parte del método.

Hay además una dificultad propia de este campo: el objeto de estudio puede cambiar mientras acumulamos evidencia sobre él. Un resultado obtenido con

una versión concreta de un modelo describe las condiciones disponibles en ese momento. Seis meses después, una actualización puede haber mejorado apreciablemente su capacidad para resolver determinados problemas, seguir instrucciones o proporcionar retroalimentación. El estudio original sigue siendo válido, pero resulta menos seguro extrapolar sus resultados a la versión siguiente. Por eso conviene registrar la versión, cuando pueda identificarse, o al menos la fecha de acceso, además de la configuración y de la información necesaria para saber qué sistema fue realmente evaluado.

Una literatura que crece muy deprisa

La acumulación rápida de estudios complica también las síntesis. En pocos años han aparecido varias revisiones y metaanálisis; cuando cubren preguntas y periodos próximos, hay que comprobar cuánto se solapan sus estudios primarios antes de tratarlos como confirmaciones independientes, además de qué variables están reuniendo bajo etiquetas amplias como *learning outcomes* o *academic performance*.

Cada pregunta requiere un diseño distinto. Una investigación cualitativa puede ser la mejor opción para estudiar cómo experimentan los docentes una tecnología nueva, y una encuesta resulta más adecuada para analizar actitudes o confianza. Lo decisivo, en cualquier caso, es si el diseño permite sostener la afirmación que finalmente hace el artículo.

Al preparar RIED 30(1) intentamos ampliar el tipo de evidencia más allá de las percepciones y pedimos, cuando la pregunta de investigación lo permitía, resultados contrastables. Entre los trabajos ya aceptados para el número, que se publicarán próximamente en *Online First*, encontramos, por ejemplo, un ensayo aleatorizado por conglomerados que evalúa desempeño, una intervención que compara retroalimentación generada por IA con la proporcionada por un experto, un estudio basado en *prompts* reales de estudiantes junto con medidas de autorregulación y otro que contrasta una clasificación automatizada con valoración humana. No responden a la misma pregunta ni aportan el mismo tipo de evidencia. Muestran distintas formas de incorporar comportamientos, producciones o contrastes externos al estudio de la IAG.

Ese avance convive con límites habituales del campo, como intervenciones breves y pocos seguimientos capaces de observar retención o transferencia. No invalidan los resultados, pero obligan a acotar las conclusiones y señalan qué evidencia falta por acumular.

Esa experiencia está también detrás de la convocatoria del monográfico RIED 30(2) sobre inteligencia artificial generativa. En ella damos por superada la fase inicial de «asombro tecnológico» y pedimos investigaciones que aporten evidencias sólidas, contrastables y relevantes. La convocatoria contempla diseños cuantitativos, cualitativos y mixtos bien justificados, y presta especial atención a estudios experimentales o cuasiexperimentales, longitudinales y de evaluación de intervenciones, entre otros (RIED, 2026).

Sabemos bastante sobre usos, percepciones y efectos inmediatos de la IAG. Sabemos mucho menos sobre cuánto duran esos efectos, hasta dónde se transfieren y qué ocurre tras un uso sostenido.

Al revisar estos trabajos solemos empezar por dos cuestiones básicas: qué se midió y qué papel tuvo la IA en la actividad. A partir de ahí puede valorarse si las conclusiones guardan proporción con los datos. Cuando el artículo sostiene que los estudiantes adquirieron una competencia, conviene saber además en qué condiciones fueron capaces de demostrarla y qué ocurre cuando la ayuda cambia o desaparece.

Una mejora del producto o del rendimiento asistido ya es un hallazgo valioso. Si se quiere hablar de aprendizaje, hace falta medir algo más.

Referencias

Alier-Forment, M., Casañ-Guerrero, M. J., Pereira-Varela, J. A., García-Peñalvo, F. J., & Llorens-Largo, F. (2026). Generative artificial intelligence and educational autonomy: Historical metaphors and ethical principles for pedagogical transformation [Inteligencia artificial generativa y autonomía educativa: metáforas históricas y principios éticos para la transformación pedagógica]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), 9–28. https://doi.org/10.5944/ried.45536

Allison, J. (2026). RAISE the standard: A framework for transparent reporting of artificial intelligence studies in education. *Journal of Educational*

Computing Research, 64(1), 3–15. https://doi.org/10.1177/07356331251377430

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Han, X., Peng, H., & Liu, M. (2025). The impact of GenAI on learning outcomes: A systematic review and meta-analysis of experimental studies. *Educational Research Review, 48*, 100714. https://doi.org/10.1016/j.edurev.2025.100714

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2026). Presenting your AI in education research with rigor: The TEP-AIED model. *Computers and Education: Artificial Intelligence, 10*, Article 100598. https://doi.org/10.1016/j.caeai.2026.100598

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Organización para la Cooperación y el Desarrollo Económicos. (2026). *OECD digital education outlook 2026: Exploring effective uses of generative AI in education*. https://doi.org/10.1787/062a7394-en

RIED. (2026, 30 de junio). *Nuevo call for papers RIED 30(2): Monográfico - IAG en educación digital: Integridad académica, evaluación, personalización y gobernanza*. Blog RIED. https://blogderied.blogspot.com/2026/06/nuevo-call-for-papers-ried-302-iag-en.html

Cómo citar esta entrada:

García Boyé, José Luis (2026). Qué cuenta como evidencia de aprendizaje con IA. *Aula Magna* [Blog]. https://cuedespyd.hypotheses.org/22908

[i] Ingeniero Técnico Superior de Telecomunicación por la Universidad Politécnica de Madrid. Colaborador de la Cátedra UNESCO de Educación a Distancia y administrador de la Lista Cuedistancia desde 2002, administrador técnico y tutor de los cursos del Máster de Educación a Distancia de la UNED (2000–2006) y Secretario Técnico de RIED en la actualidad.