

DOI: <https://doi.org/10.70577/jd09wn31>

Usar no es aprender: una lectura crítica de PISA 2025 y de la evidencia internacional sobre IAGen

Melvin Ramírez Bogantes

Profesor investigador | Escuela de Matemática

<https://orcid.org/0000-0001-5516-0085>

meramirez@itcr.ac.cr

Tecnológico de Costa Rica

Heredia - Costa Rica

Marco Gutiérrez Montenegro

Profesor investigador | Escuela de Matemática

<https://orcid.org/0009-0004-0293-1578>

vgutierrez@itcr.ac.cr

Tecnológico de Costa Rica

Cartago - Costa Rica

Cómo citar:	Fecha de recepción: 2026-09-06
	Fecha de aceptación: 2026-09-18
	Fecha de publicación: 2026-10-04

Resumen

La discusión educativa sobre inteligencia artificial generativa suele tratar el “uso de IA” como una exposición homogénea cuyos efectos podrían establecerse mediante una respuesta única. Este ensayo cuestiona esa premisa mediante un análisis documental comparado de un corpus principal de fuentes institucionales y académicas de alta influencia publicadas entre 2023 y 2026, complementado por antecedentes teóricos sobre instrumentación, descarga cognitiva y autorregulación. El corpus comprende PISA 2025, TALIS 2024, documentos normativos de UNESCO, el *OECD Digital Education Outlook 2026*, el *AI Index Report 2026* y dos metaanálisis recientes sobre ChatGPT y resultados de aprendizaje. Los resultados muestran que las aparentes contradicciones entre asociaciones negativas, efectos experimentales favorables y recomendaciones normativas disminuyen cuando se distinguen el tipo de uso observado, el trabajo cognitivo que permanece en el estudiante, la mediación o regulación de la interacción y el alcance inferencial del diseño. PISA 2025 identifica relaciones diferenciadas entre uso de chatbots y desempeño científico según propósito y frecuencia, así como desigualdades en las oportunidades escolares declaradas para evaluar información generada por IA; su diseño no permite convertir esas asociaciones en efectos causales. Como contribución, se propone un criterio de cuatro preguntas para interpretar evidencia sobre inteligencia artificial generativa y aprendizaje. El caso costarricense muestra que el uso relativamente alto de chatbots, las oportunidades declaradas de evaluación crítica y los desafíos persistentes de desempeño pueden coexistir, por lo que la discusión debe desplazarse desde la mera disponibilidad hacia la calidad del uso educativo.

Palabras clave: inteligencia artificial generativa; PISA 2025; aprendizaje; política educativa; alfabetización en IA; evaluación educativa; educación comparada.

Using Is Not Learning: A Critical Reading of PISA 2025 and International Evidence on Generative AI

Abstract

Educational debate on generative artificial intelligence often treats “AI use” as a homogeneous exposure whose effects could be settled by a single answer. This essay challenges that premise through a comparative document analysis of a core corpus of influential institutional and academic sources published between 2023 and 2026, complemented by theoretical work on instrumentation, cognitive offloading, and self-regulated learning. The corpus includes PISA 2025, TALIS 2024, UNESCO normative documents, the *OECD Digital Education Outlook 2026*, the *AI Index Report 2026*, and two recent meta-analyses of ChatGPT and learning outcomes. The findings show that apparent contradictions among negative associations, favorable experimental effects, and normative recommendations diminish when four dimensions are distinguished: the type of use observed, the cognitive work retained by the learner, the mediation or regulation of the interaction, and the inferential scope of the design. PISA 2025 identifies differentiated relationships between chatbot use and science performance by purpose and frequency, as well as inequalities in reported school opportunities to assess AI-generated information; its design does not permit these associations to be interpreted as causal effects. As a contribution, the article proposes a four-question criterion for interpreting evidence on generative artificial intelligence and learning. The Costa Rican case shows that relatively high chatbot use, reported opportunities for critical evaluation, and persistent performance challenges can coexist, shifting the discussion from mere availability to the quality of educational use.

Keywords: generative artificial intelligence; PISA 2025; learning; education policy; AI literacy; educational assessment; comparative education.

Introducción

El dato y el titular

La relación entre inteligencia artificial generativa y aprendizaje ha quedado atrapada con rapidez en una gramática pública demasiado simple. Cuando una intervención apoyada por ChatGPT registra mejores resultados que una condición de comparación, el hallazgo tiende a condensarse en la idea de que la inteligencia artificial mejora el aprendizaje. Cuando una evaluación a gran escala encuentra que ciertos usuarios de chatbots obtienen puntuaciones inferiores a quienes no los usan, la lectura se invierte: la inteligencia artificial perjudica el aprendizaje. En ambos casos desaparecen la definición concreta del uso, la población, la tarea y la arquitectura metodológica que produjo el resultado; una relación situada se convierte en propiedad general de una tecnología.

PISA 2025 permite examinar este problema porque incorporó información sobre el uso de chatbots basados en inteligencia artificial para el trabajo escolar. El informe distingue cuatro propósitos: resumir lecturas asignadas, realizar investigación preliminar, redactar textos para trabajos escritos y utilizar chatbots “para ayudarme a aprender” (OECD, 2026c, Tabla I.B1.4.53). También recoge tareas escolares en las que el alumnado evalúa la calidad de información generada por IA (OECD, 2026c,

Tablas I.B1.4.67-I.B1.4.70). Esta desagregación supera la oposición usuario/no usuario, aunque todavía no describe la actividad cognitiva concreta que ocurre durante la interacción.

Los resultados muestran una relación compleja con el desempeño científico. Para tareas específicas, quienes no utilizan chatbots tienden a superar a quienes sí los emplean; para el propósito general de “ayudarme a aprender”, el uso semanal se asocia con desempeño semejante al de los no usuarios, mientras que el uso muy frecuente o esporádico se relaciona con puntuaciones inferiores (OECD, 2026c, Figura I.4.13; Tabla I.B1.4.61). La propia OCDE advierte que estas relaciones no implican necesariamente un efecto negativo y que pueden reflejar tanto quién adopta la tecnología como la manera en que la utiliza.

En paralelo, metaanálisis recientes de estudios experimentales y cuasiexperimentales encuentran efectos medios positivos de intervenciones que incorporan ChatGPT. Wu et al. (2026), a partir de 35 estudios y 4 193 participantes, estiman Hedges $g = 0,670$, aunque con heterogeneidad muy alta ($I^2 = 91,444\%$). Liu et al. (2025), en 37 estudios, estiman $g = 0,577$. Estas cifras no invalidan PISA ni PISA refuta los metaanálisis: las fuentes responden preguntas distintas, observan poblaciones diferentes y autorizan inferencias diferentes. El experimento pregunta qué ocurre bajo una intervención definida; la evaluación transversal describe cómo se distribuyen prácticas autorreportadas y con qué resultados se relacionan.

Algo semejante ocurre con los documentos normativos. UNESCO propone competencias para estudiantes y docentes centradas en juicio crítico, ética y capacidad de uso responsable (Miao et al., 2024; Miao & Cukurova, 2024), pero esos marcos no son experimentos. El *OECD Digital Education Outlook 2026* añade una distinción decisiva: mejorar la ejecución de una tarea con apoyo generativo no equivale necesariamente a producir aprendizaje duradero; la descarga cognitiva puede elevar el desempeño inmediato sin garantizar adquisición de capacidades (OECD, 2026a).

El problema central se sitúa, por tanto, antes de decidir si conviene permitir, restringir o promover la IAGen. Consiste en esclarecer qué significa la evidencia que sustenta esas decisiones. Las evaluaciones internacionales describen patrones poblacionales; los experimentos estiman efectos bajo condiciones definidas; los documentos normativos establecen capacidades y principios; las síntesis integran resultados heterogéneos. Todas pueden hablar de “IA y aprendizaje” sin hablar exactamente del mismo objeto.

La pregunta orientadora es: ¿qué evidencia sobre la relación entre IAGen y aprendizaje producen las fuentes internacionales contemporáneas seleccionadas, qué miden efectivamente, qué inferencias permiten y qué criterio puede construirse a partir de su lectura conjunta para orientar decisiones docentes y de política educativa en un sistema como el costarricense? Se defenderá que “usar IAGen” es una categoría insuficiente: su interpretación exige distinguir tipo de uso, trabajo cognitivo conservado, mediación o regulación e inferencia autorizada por el diseño. Bajo esta lectura, PISA 2025 no entrega un veredicto sobre la IAGen; muestra por qué la categoría de uso debe abrirse antes de atribuirle efectos educativos.

Estrategia documental y criterios de contraste

El ensayo se construye como un análisis documental comparado, crítico e interpretativo. No pretende estimar exhaustivamente el conjunto de efectos disponibles ni satisfacer los requisitos de una revisión sistemática. Su propósito es contrastar fuentes internacionales de alta influencia que producen afirmaciones sobre IAGen, educación y aprendizaje desde posiciones institucionales y diseños metodológicos diferentes. La selección del corpus es intencional y teóricamente orientada; por ello, su validez depende de hacer explícitos los criterios de inclusión y la lógica de comparación.

Se utilizaron seis criterios. El primero fue la autoridad institucional o académica de la fuente. El segundo, su capacidad de influir en política educativa, investigación o discurso público. El tercero, la relevancia directa para el uso educativo de IAGen, el aprendizaje o la alfabetización en inteligencia artificial. El cuarto, la diversidad metodológica necesaria para comparar formas distintas de producción de evidencia. El quinto, la actualidad, con prioridad para documentos publicados entre 2023 y 2026, periodo en que la IAGen de acceso masivo se incorpora a las prácticas educativas. El sexto, la trazabilidad: se priorizaron fuentes primarias o publicaciones académicas que permiten reconstruir población, definición operacional, diseño, estimación y limitaciones.

El corpus quedó integrado por *PISA 2025 Results (Volume I): Future-Ready Students* y la nota de país para Costa Rica (OECD, 2026b, 2026c); *Results from TALIS 2024: The State of Teaching* (OECD, 2025); la *Guidance for Generative AI in Education and Research* (Miao & Holmes, 2023); los marcos UNESCO de competencias en IA para estudiantes y docentes (Miao et al., 2024; Miao & Cukurova, 2024); el *Global Education Monitoring Report 2023* sobre tecnología en educación (Global Education Monitoring Report Team, 2023); el *OECD Digital Education Outlook 2026* (OECD, 2026a); el capítulo educativo del *AI Index Report 2026* (Stanford Institute for Human-Centered Artificial Intelligence, 2026); y los metaanálisis de Liu et al. (2025) y Wu et al. (2026).

La inclusión no supone equivalencia epistémica. Cada fuente se comparó mediante dimensiones comunes: finalidad dominante; unidad de análisis; población; definición de uso; temporalidad; tipo de dato; resultado educativo; tratamiento de posibles variables de confusión; alcance inferencial; limitaciones; recomendaciones; y dimensión de equidad. El procedimiento permite separar tres niveles que suelen mezclarse en el debate: el dato observado, la inferencia autorizada por el diseño y la interpretación teórica compatible con ese dato.

La selección no busca representar toda la literatura mundial sobre IAGen. Aquí interesa un problema de educación comparada y política de evidencia: cómo se construye una afirmación cuando fuentes altamente visibles, pero metodológicamente no equivalentes, se colocan en una misma conversación pública. La explicitación del corpus y de los criterios hace discutible y reproducible la lógica de contraste sin simular una exhaustividad que el ensayo no pretende.

La regla de trazabilidad fue restrictiva. Las cifras de PISA y TALIS se verificaron en documentos y tablas oficiales; las magnitudes de efecto, en los artículos de origen; las características de los marcos UNESCO, en las publicaciones institucionales; y los indicadores del *AI Index* se utilizaron únicamente para contextualizar adopción y gobernanza, no como evidencia de aprendizaje. Las fuentes periodísticas se excluyeron como respaldo empírico. El resultado metodológico no es una jerarquía

simple de diseños: una evaluación poblacional posee capacidades que un experimento local no tiene, un experimento puede estimar efectos que una encuesta transversal no identifica y un marco normativo responde preguntas que ninguna regresión resuelve.

El enfoque presenta tres límites que deben conservarse al interpretar sus conclusiones: el corpus es selectivo y no exhaustivo; depende de las definiciones operacionales adoptadas por cada fuente; y examina un campo de rápida transformación, por lo que los indicadores, las herramientas y los marcos de gobernanza pueden modificarse con rapidez. El criterio propuesto debe entenderse, en consecuencia, como una herramienta analítica para disciplinar la lectura de la evidencia y no como una síntesis completa de la literatura mundial.

Marco de lectura

Regímenes de evidencia, tipos de uso y mediación

Tres regímenes de evidencia

Las afirmaciones sobre IAGen y aprendizaje que circulan en el debate educativo se producen bajo regímenes de evidencia distintos. La expresión se utiliza aquí para designar conjuntos de preguntas, diseños y reglas de inferencia que determinan qué puede afirmarse legítimamente. No constituye una jerarquía de calidad. Una evaluación internacional, un experimento y un marco normativo pueden ser rigurosos y, al mismo tiempo, responder a preguntas que no son intercambiables.

El régimen observacional a gran escala pregunta cómo se distribuyen determinadas prácticas y con qué resultados se relacionan en poblaciones amplias. PISA y TALIS pertenecen a este registro. Su fortaleza es comparativa y poblacional; su límite es que la asociación ajustada no identifica por sí sola mecanismos ni causalidad. El régimen experimental o cuasiexperimental pregunta qué ocurre cuando una intervención definida se compara con una condición alternativa. Puede estimar efectos dentro de las condiciones estudiadas, aunque la transferibilidad a usos cotidianos o contextos diferentes debe justificarse. El régimen prescriptivo-normativo pregunta qué capacidades, principios y condiciones deberían orientar la integración de una tecnología. Los marcos UNESCO se ubican aquí: su autoridad descansa en una argumentación normativa y una síntesis experta, no en la estimación de un efecto.

Tabla 1.

Regímenes de evidencia sobre IAGen y aprendizaje

Régimen	Pregunta dominante	Diseño o fuente típica	Inferencia autorizada	Límite característico
Observacional a gran escala	¿Cómo se distribuyen a determinadas prácticas y con qué resultados se relacionan?	Evaluaciones y encuestas internacionales, autorreportes, modelos poblacionales	Asociación, diferencia, distribución y relación ajustada	No identifica por sí solo mecanismos ni efectos causales

Régimen	Pregunta dominante	Diseño o fuente típica	Inferencia autorizada	Límite característico
Experimental o cuasiexperimental	¿Qué ocurre bajo una intervención definida y comparada?	Ensayos y cuasiexperimentos; y metaanálisis de estos estudios	Efecto estimado dentro de las condiciones estudiadas	Transferibilidad limitada a contextos y usos diferentes
Prescriptivo-normativo	¿Qué capacidades, principios y condiciones deberían orientar la adopción?	Guías, marcos de competencia y orientaciones institucionales	Justificación y normativa, curricular y organizativa	No estima efectos educativos por su formulación

Las fuentes de síntesis y monitoreo, como el *OECD Digital Education Outlook* o el *AI Index*, no constituyen un cuarto régimen homogéneo. Integran, resumen o contextualizan evidencia producida por otros diseños y heredan, por tanto, parte de sus límites. Esta distinción importa porque la pregunta por “qué funciona” presupone con facilidad que el objeto de la intervención es estable y que el efecto viaja entre contextos, cuando en educación ambas suposiciones necesitan demostración (Biesta, 2007; Cartwright & Hardie, 2012). La comparación rigurosa no exige elegir una fuente superior, sino reconocer qué pregunta respondió cada una y conservar esa frontera al formular conclusiones.

Tipos de uso y distribución del trabajo cognitivo

El segundo problema conceptual es la expresión “uso de IA”. Aparentemente describe una conducta sencilla: un estudiante usa o no usa una herramienta y lo hace con cierta frecuencia. Sin embargo, la misma etiqueta puede reunir actividades cognitivamente incompatibles. Dos estudiantes pueden declarar que emplean IAGen para redactar y conservar grados muy diferentes de autoría intelectual. Uno puede solicitar un texto completo y asumirlo sin revisión; otro puede elaborar una tesis propia, pedir objeciones, contrastar argumentos, verificar fuentes y reescribir. La categoría nominal coincide; el trabajo cognitivo que permanece en el estudiante no.

La tradición de la génesis instrumental ayuda a precisar este punto. Para Rabardel (1995), un artefacto no se convierte en instrumento por el mero hecho de estar disponible: se vuelve instrumento en la actividad del sujeto mediante procesos de apropiación y construcción de esquemas de utilización. En educación matemática, Artigue (2002) y Trouche (2004) mostraron que esa relación articula dimensiones técnicas y conceptuales y que la intervención docente puede organizarse mediante orquestaciones instrumentales. En este marco, “instrumental” no significa superficial, mecánico o delegativo. Equiparar uso instrumental con cesión del trabajo intelectual sería conceptualmente incorrecto.

Por ello se emplean categorías descriptivas más prudentes. Se denomina propósito potencialmente delegativo a aquel que permite externalizar una parte sustantiva del trabajo que una tarea podría exigir

al estudiante. “Potencialmente” es indispensable: resumir, investigar o redactar pueden servir para sustituir actividad, pero también para contrastar una producción propia. Sin trazas de interacción, productos intermedios u observación del proceso, no puede establecerse cuál de esos recorridos ocurrió.

La noción de descarga cognitiva ofrece un vocabulario útil para describir esta externalización sin prejuzgarla. Delegar operaciones a recursos externos puede reducir demandas internas y ser adaptativo; también puede resultar educativamente problemático cuando se descarga precisamente la operación que se desea desarrollar (Risko & Gilbert, 2016). La pregunta pertinente no es cuánto trabajo realiza la máquina en términos absolutos, sino qué trabajo era epistémicamente relevante para el objetivo de aprendizaje. Automatizar un cálculo auxiliar puede liberar recursos para argumentar; externalizar la argumentación completa cuando esa es la capacidad evaluada cambia el significado del producto final.

Mediación, regulación y alfabetización crítica

Se denomina uso regulado o epistémicamente mediado a una configuración en la que el estudiante conserva funciones de juicio, selección, contraste, verificación o corrección sobre la producción de la IAGen. Tampoco se trata de una categoría binaria. La regulación puede ser parcial, episódica y dependiente de la tarea. Su interés radica en desplazar la atención desde la mera presencia de la herramienta hacia la organización de la actividad. Desde la literatura sobre aprendizaje autorregulado, planificar, monitorear y evaluar la propia actuación son precisamente procesos centrales para sostener autonomía y control del aprendizaje (Zimmerman, 2002).

PISA 2025 aproxima una parte de esta dimensión mediante tareas escolares en las que el alumnado evalúa la calidad de información generada por IA (OECD, 2026c, Tabla I.B1.4.67). El indicador es relevante, pero no debe convertirse en una medida total de mediación pedagógica. No informa cuánto duró la actividad, qué criterios se enseñaron, si existió retroalimentación, con qué frecuencia se practicó ni si la competencia se transfirió a interacciones autónomas. Conviene hablar de oportunidades escolares de evaluación crítica y reservar expresiones más fuertes para estudios que observan la intervención con mayor granularidad.

De este marco se siguen cuatro dimensiones que estructuran el resto del ensayo: tipo de uso, distribución del trabajo cognitivo, mediación o regulación y alcance inferencial. Una asociación negativa puede reflejar selección de usuarios, dificultades previas, dependencia o tipo de tarea; un efecto experimental positivo puede pertenecer a una intervención cuidadosamente diseñada y no al chatbot aislado; una recomendación normativa puede ser defendible sin funcionar como demostración causal. “Uso de IA”, considerado sin estas distinciones, es una variable demasiado gruesa para sostener por sí sola una interpretación pedagógica.

Qué mide PISA 2025 y qué no mide

PISA 2025 fue aplicado a aproximadamente 760 000 estudiantes de 91 países y economías y tuvo ciencias como dominio principal (OECD, 2026c). La información sobre IAGen proviene del cuestionario de contexto, no de una prueba específica de competencia en inteligencia artificial. Esta

distinción es fundamental: PISA no evalúa si el estudiante sabe construir una consulta de alta calidad, detectar una alucinación, comparar una respuesta con evidencia primaria o regular la dependencia de un chatbot. Registra, mediante autorreporte, frecuencia y propósito de uso, junto con determinadas experiencias escolares relacionadas con recursos digitales.

La encuesta pregunta por chatbots basados en IA, con ChatGPT como ejemplo, y diferencia cuatro finalidades: resumir textos de lectura asignada, realizar investigación preliminar sobre un tema nuevo, redactar textos para trabajos escritos y ayudar a aprender en términos generales (OECD, 2026c, Figura I.4.11; Tabla I.B1.4.53). En promedio OCDE, el 45,5% de estudiantes declara usar IA al menos semanalmente para ayudarlo a aprender (OECD, 2026c, Tabla I.6). La nota costarricense resume además los porcentajes promedio de la OCDE para algunos propósitos: 31% en investigación preliminar, 30% en resumen y 29% en redacción (OECD, 2026b).

La granularidad es metodológicamente valiosa porque evita tratar como equivalentes usos con finalidades diferentes. Sin embargo, el propósito informado sigue siendo una categoría amplia. “Investigar” puede significar solicitar una respuesta única o utilizar el chatbot para construir términos de búsqueda y contrastar fuentes. “Resumir” puede consistir en sustituir la lectura o en comparar un resumen propio con uno generado. “Redactar” puede abarcar desde corrección estilística hasta producción casi completa. PISA no dispone de trazas que permitan distinguir esos itinerarios.

La frecuencia añade otra dimensión. El informe diferencia desde ausencia o uso muy ocasional hasta uso casi diario y muestra que la relación con desempeño no sigue una línea simple (OECD, 2026c, Tabla I.B1.4.61). Para el propósito general de ayudar a aprender, los usuarios semanales presentan, en promedio, un desempeño científico similar al de los no usuarios una vez considerado el perfil socioeconómico del estudiante. Quienes reportan uso muy frecuente y quienes lo usan solo ocasionalmente tienden a obtener puntuaciones inferiores. Para tareas específicas, los no usuarios suelen superar a usuarios, aunque en resumen e investigación los usuarios moderados superan a usuarios muy limitados y a usuarios frecuentes. Este patrón ya desaconseja interpretar “más uso” como equivalente a “más beneficio” o “más daño”.

PISA también introduce una variable que aproxima oportunidades escolares de alfabetización crítica. En el promedio OCDE, el 62,6% del alumnado reportó que, en sus lecciones, realizó con alguna frecuencia tareas para evaluar la calidad de información generada por chatbots de IA (OECD, 2026c, Tabla I.6). El capítulo 4 precisa que aproximadamente seis de cada diez estudiantes recibieron tareas de este tipo y analiza esas oportunidades según características estudiantiles y escolares (Tablas I.B1.4.67-I.B1.4.68). La formulación importa: se trata de un autorreporte de tareas realizadas, no de haber completado un currículo validado de alfabetización en IA ni de una medición directa de competencia.

El resultado de aprendizaje con el que se relacionan estas variables es, de manera central en el capítulo, el desempeño en ciencias. La Figura I.4.13 presenta puntuaciones medias de ciencias por frecuencia de uso y controla el perfil socioeconómico del estudiante; las tablas anexas permiten desagregaciones adicionales. Tal ajuste aumenta la comparabilidad estadística, pero no transforma el diseño en experimento. Continúan siendo posibles diferencias no observadas entre quienes usan y

no usan la herramienta: rendimiento previo, motivación, dificultades académicas, disponibilidad de apoyo, tipo de centro, cultura familiar, calidad de las tareas o competencia digital.

La temporalidad refuerza la cautela. Los chatbots generativos de acceso masivo se difundieron a finales de 2022 y PISA recogió datos en 2025. Los estudiantes, las escuelas y las políticas estaban todavía en una fase temprana de apropiación. Lo que se mide no es una tecnología estable incorporada durante toda la trayectoria escolar, sino prácticas surgidas en un periodo de rápida transformación. Esto limita cualquier intento de proyectar linealmente los patrones hacia contextos futuros en los que herramientas, normas y capacidades serán diferentes.

También importa la unidad de análisis. PISA vincula respuestas individuales del cuestionario con resultados de evaluación y, en algunos modelos, con características escolares; no reconstruye episodios didácticos particulares. Por ello, una diferencia promedio entre categorías de uso no puede trasladarse automáticamente a una recomendación sobre una tarea concreta sin considerar su diseño y el contenido disciplinar implicado.

El autorreporte plantea, además, problemas habituales de medición. Dos estudiantes pueden interpretar de forma distinta qué significa usar IA “para aprender”; la frecuencia no informa duración ni intensidad; y la categoría “chatbot basado en IA” puede incluir productos con capacidades diferentes. Estas limitaciones no invalidan PISA, pero delimitan el tipo de pregunta que puede responder. Su fortaleza reside en describir la extensión y distribución de prácticas y en relacionarlas con resultados dentro de una población amplia. Su debilidad, para el problema aquí tratado, es que no observa directamente la microestructura de la interacción.

Por ello, PISA 2025 puede sostener afirmaciones sobre prevalencia, frecuencia, propósito, distribución desigual de oportunidades y asociaciones con desempeño. No puede demostrar que la utilización de chatbots cause una disminución o mejora del aprendizaje, ni que una tarea de evaluación de información generada por IA constituya por sí sola mediación pedagógica efectiva. Esa delimitación no reduce el interés del informe; impide pedirle una respuesta que su diseño no fue construido para producir.

Hallazgos de PISA 2025: descripción e interpretación

Una lectura disciplinada de PISA exige mantener separados tres planos: dato observado, inferencia autorizada e interpretación teórica. La distinción permite extraer valor del informe sin convertir una asociación en mecanismo causal. Los resultados sobre IAGen son especialmente adecuados para este ejercicio porque presentan patrones diferentes según finalidad, frecuencia y oportunidad escolar de evaluación crítica.

El primer patrón se refiere a tareas específicas. El capítulo 4 señala que quienes no usan chatbots para resumir textos, realizar investigación preliminar o redactar tienden a obtener mejores resultados en ciencias que quienes los usan (OECD, 2026c, Tablas I.B1.4.60-I.B1.4.61). El dato no autoriza a concluir que esas actividades causen menor rendimiento. Los usuarios pueden recurrir a la herramienta precisamente porque enfrentan dificultades previas o porque realizan tareas distintas.

Tampoco puede saberse si el uso sustituyó una operación cognitiva o funcionó como apoyo a una producción propia.

La interpretación más prudente es que estos propósitos permiten una externalización sustantiva del trabajo cognitivo. Un chatbot puede sintetizar, recopilar información o producir texto sin que el estudiante ejecute todas las operaciones que una tarea convencional exigiría. Esa posibilidad genera una hipótesis didáctica: cuando la externalización coincide con la capacidad que se pretende desarrollar, puede reducir oportunidades de práctica. PISA no prueba la hipótesis; la vuelve pertinente porque los usos se relacionan con el desempeño de manera diferenciada.

El segundo patrón es no lineal. Para el uso general “para ayudarme a aprender”, la OCDE informa que el uso semanal se asocia con desempeño científico similar al de los no usuarios, mientras que el uso casi diario y el uso ocasional se relacionan con puntuaciones inferiores (OECD, 2026c, Figura I.4.13). Este resultado es difícil de conciliar con respuestas binarias. Si la IAGen tuviera un efecto uniforme e independiente de la actividad, esperaríamos una relación más estable con la frecuencia. La curva sugiere, al menos, que intensidad y finalidad importan y que es posible que distintos tipos de estudiantes se concentren en categorías diferentes.

El tercer patrón incorpora las oportunidades escolares para evaluar información generada por IA. El 62,6% del alumnado del promedio OCDE reportó haber realizado en sus lecciones este tipo de tareas (OECD, 2026c, Tabla I.6). Entre quienes disponen de esas oportunidades, los usuarios frecuentes de IA para aprender obtienen resultados científicos ligeramente superiores a los no usuarios o usuarios poco frecuentes; el informe advierte, sin embargo, que no puede determinarse causalidad y que, al considerar los perfiles socioeconómicos de estudiantes y centros, no aparece una asociación positiva general entre realizar tareas de evaluación de contenido de IA y el desempeño (OECD, 2026c, Figura I.4.14; Tablas I.B1.4.68-I.B1.4.70).

Este matiz modifica una lectura demasiado optimista. No sería correcto afirmar que “enseñar a evaluar IA mejora los puntajes”. El resultado más sólido es que el patrón de uso y desempeño cambia cuando se consideran oportunidades escolares de evaluación crítica, lo que es compatible con una hipótesis de mediación, pero no la demuestra. Las escuelas que ofrecen esas experiencias pueden diferir en múltiples dimensiones; además, un único indicador no captura calidad, intensidad ni transferencia de la alfabetización.

La distribución social de esas oportunidades resulta quizás más relevante para política que la pequeña diferencia de desempeño. La OCDE reporta que estudiantes socioeconómicamente aventajados declaran con mayor frecuencia tareas escolares de evaluación de información generada por IA (OECD, 2026c, Tabla I.B1.4.68). Si la capacidad de juzgar información automatizada se convierte en condición de participación académica, la brecha tecnológica puede desplazarse desde el acceso hacia la calidad de la apropiación. La disponibilidad de una herramienta gratuita no iguala automáticamente la capacidad de usarla con criterio.

Este punto se relaciona con un hallazgo más amplio de PISA 2025: el descenso observado en lectura comprende precisamente capacidades especialmente relevantes en contextos saturados de información, como evaluar fuentes, conectar información procedente de varios textos y razonar

críticamente sobre lo leído (OECD, 2026c). El problema de la IAGen no puede separarse de las capacidades fundamentales que permiten interrogar sus respuestas. Una pedagogía de verificación presupone lectura, razonamiento disciplinar y conocimiento suficiente para detectar incoherencias.

Tabla 2.

PISA 2025: propósitos de uso, asociación reportada, límite inferencial e interpretación prudente

Dimensión observada	Asociación reportada por PISA 2025	Límite inferencial	Interpretación prudente
Resumir lecturas asignadas	Los no usuarios tienden a superar a usuarios; los usuarios moderados superan a usuarios muy limitados y frecuentes	El propósito autorreportado no muestra cómo se realizó el resumen	Puede permitir externalización de procesamiento, pero no demuestra delegación cognitiva
Investigación preliminar	Patrón semejante: no usuarios con mejor desempeño y ventaja relativa del uso moderado frente a extremos	“Investigar” agrupa prácticas muy heterogéneas	La calidad del proceso de búsqueda y contraste no es observable en PISA
Redactar trabajos	Los no usuarios tienden a superar a usuarios; la ventaja del uso moderado es menos pronunciada	No se conoce cuánto texto produce la IA ni qué autoría conserva el estudiante	El mismo propósito puede abarcar apoyo lingüístico o sustitución sustantiva de escritura
Ayudar a aprender	Usuarios semanales con desempeño similar a no usuarios; uso muy frecuente u ocasional asociado con menor desempeño	Frecuencia puede ser consecuencia de necesidades previas	La relación no lineal cuestiona lecturas binarias de “más uso = mejor/peor”
Evaluar información generada por IA en clase	Entre quienes tienen esa oportunidad, usuarios frecuentes muestran ligera ventaja; no existe asociación positiva general tras controles socioeconómicos	Un indicador de oportunidad no mide calidad de mediación ni causalidad	Es compatible con una hipótesis de alfabetización crítica, pero no prueba su efecto

La Tabla 2 condensa una regla: el verbo de la conclusión debe conservar la estructura del diseño. PISA permite decir “se asocia”, “se relaciona” o “presenta un patrón diferente”, no “produce” o “causa”. Su contribución, leída de este modo, no es un veredicto favorable o adverso, sino mostrar que la categoría “uso de IA” oculta heterogeneidad según actividad, frecuencia y contexto escolar.

Contraste internacional por régimen de evidencia

La aparente contradicción entre PISA y los resultados experimentales disminuye cuando las fuentes se organizan por régimen de evidencia. Esta reorganización evita comparar cifras que estiman objetos diferentes y permite identificar una convergencia más modesta: la calidad educativa del uso depende de cómo se configura la interacción, no de la mera disponibilidad de la tecnología.

El régimen prescriptivo: capacidades y condiciones deseables

La guía UNESCO sobre IAGen adopta un enfoque centrado en el ser humano y propone acciones regulatorias, protección de datos, validación pedagógica y desarrollo de capacidades para una integración educativa responsable (Miao & Holmes, 2023). Su propósito es normativo. No busca estimar un tamaño de efecto, sino orientar decisiones sobre qué condiciones deberían existir antes y durante la adopción.

Los marcos de competencias de 2024 desarrollan esa orientación. El marco para estudiantes organiza 12 competencias en cuatro dimensiones —mentalidad centrada en lo humano, ética de la IA, técnicas y aplicaciones, y diseño de sistemas— y tres niveles de progresión: comprender, aplicar y crear (Miao et al., 2024). El marco para docentes propone 15 competencias en cinco dimensiones y tres niveles de progresión, con énfasis en mentalidad centrada en lo humano, ética, fundamentos y aplicaciones, pedagogía con IA y desarrollo profesional (Miao & Cukurova, 2024). Estos documentos justifican la idea de alfabetización crítica, pero no constituyen evidencia de que alcanzar una competencia determinada cause un incremento mensurable en PISA.

El *Global Education Monitoring Report 2023* proporciona un antecedente útil para interpretar la velocidad de adopción tecnológica. El informe subraya que la evidencia robusta e imparcial sobre impacto de tecnología educativa es limitada, que la generalización entre contextos es difícil y que la producción de evidencia puede estar condicionada por actores con interés en la expansión de determinadas tecnologías (Global Education Monitoring Report Team, 2023). La advertencia adquiere especial relevancia para la IAGen: la velocidad de innovación puede superar la capacidad de producir evidencia longitudinal sobre aprendizaje.

El régimen observacional a gran escala: distribución y capacidades institucionales

PISA observa estudiantes; TALIS permite incorporar la perspectiva docente. En 2024, el 36% de docentes de secundaria baja del promedio OCDE declaró haber usado IA en su trabajo; entre usuarios, 68% la empleó para aprender o resumir temas y 64% para generar planes o actividades, mientras que 25% la utilizó para revisar datos de desempeño y 26% para evaluar o calificar (OECD, 2025, Tablas 1.59-1.60). La variable de TALIS se refiere a IA en sentido amplio, no exclusivamente a IAGen, por lo que no debe equipararse sin más con los chatbots generativos de PISA.

TALIS también muestra una tensión entre adopción y capacidad profesional. En promedio, el 29% de docentes reportó una necesidad alta de desarrollo profesional en el uso de IA para enseñar y aprender, la proporción más elevada entre las áreas de formación examinadas, y 38% había participado en formación profesional sobre IA (OECD, 2025, Tablas 4.27-4.29). Estos datos no prueban que la formación mejore resultados estudiantiles. Sí muestran que la expansión tecnológica ocurre dentro de sistemas donde una parte importante del profesorado identifica necesidades de conocimiento y preparación.

El régimen experimental: efectos bajo intervenciones definidas

Los metaanálisis permiten una comparación distinta. Wu et al. (2026) sintetizaron 35 estudios experimentales o cuasiexperimentales publicados entre 2022 y 2024, con 4 193 participantes y 134 tamaños de efecto. El efecto agregado sobre resultados de aprendizaje fue $g = 0,670$, IC 95% [0,495, 0,844], con efectos diferenciados en resultados cognitivos y no cognitivos. El resultado, sin embargo, estuvo acompañado por heterogeneidad muy elevada: $Q = 409,067$ e $I^2 = 91,444\%$, razón por la que los autores emplearon un modelo de efectos aleatorios. Además, disciplina, duración de la intervención y modalidad instructiva funcionaron como moderadores significativos.

El dato experimental autoriza hablar de un efecto medio positivo en el conjunto de intervenciones incluidas, no de un efecto universal de ChatGPT. La heterogeneidad indica precisamente que las intervenciones no producen resultados intercambiables. En algunos contextos el efecto fue mayor; en otros, menor o negativo. La pregunta pedagógicamente productiva no es si $g = 0,670$ demuestra que ChatGPT “funciona”, sino qué configuraciones de tarea, duración, disciplina y modalidad explican parte de la variación.

Liu et al. (2025) ofrecen una segunda síntesis vigente. Su metaanálisis incluyó 37 estudios publicados entre 2022 y 2025 y estimó un efecto medio de $g = 0,577$, IC 95% [0,395, 0,759], sobre rendimiento académico. Los autores examinaron moderadores como nivel educativo, disciplina, duración, tamaño de muestra, tipo de conocimiento, modalidad instructiva, rol asignado a ChatGPT y tipo de contenido generado. La convergencia en una magnitud media positiva entre ambos metaanálisis resulta relevante, pero no elimina la diversidad de condiciones ni autoriza a trasladar directamente esos efectos al uso cotidiano registrado por PISA.

Ambas síntesis requieren, además, dos cautelas. Wu et al. (2026) integraron 134 tamaños de efecto procedentes de 35 estudios y los propios autores señalan que futuras investigaciones deberían emplear técnicas como el metaanálisis de tres niveles; el promedio agregado depende, por tanto, de las decisiones analíticas adoptadas. Asimismo, las ventanas temporales de Wu et al. (2022-2024) y Liu et al. (2022-2025) se solapan ampliamente. La similitud entre $g = 0,670$ y $g = 0,577$ constituye una concordancia descriptiva entre dos síntesis, pero no debe interpretarse como una replicación completamente independiente sin contrastar la coincidencia de estudios primarios.

En términos comparados, la contradicción se disuelve. PISA pregunta cómo se relaciona el uso autorreportado con desempeño científico en una población internacional; los metaanálisis preguntan cuánto difieren resultados bajo intervenciones definidas respecto de comparadores. Un estudiante que

usa un chatbot por su cuenta para redactar una tarea no constituye la misma “exposición” que un participante de un estudio donde el uso está guiado, delimitado y alineado con un objetivo pedagógico.

Fuentes transversales: síntesis y monitoreo

El OECD Digital Education Outlook 2026 sintetiza literatura emergente y formula una distinción que resulta central para este ensayo: ejecutar mejor una tarea con IAGen no garantiza aprendizaje. La OCDE señala que herramientas generativas de propósito general pueden elevar la calidad del producto mientras están disponibles, pero que las ventajas pueden desaparecer o incluso invertirse cuando el apoyo se retira; también advierte sobre riesgos de descarga cognitiva y “pereza metacognitiva” cuando la tecnología sustituye esfuerzo que debería contribuir a la adquisición de habilidades (OECD, 2026a). En cambio, usos diseñados con intención pedagógica y alineados con principios de aprendizaje muestran resultados más prometedores.

El AI Index Report 2026 aporta otra dimensión: escala de adopción y rezago de gobernanza. El capítulo de educación informa que cuatro de cada cinco estudiantes estadounidenses de secundaria y universidad utilizan IA para trabajo escolar, mientras que solo alrededor de la mitad de las escuelas medias y secundarias dispone de políticas y apenas 6% de docentes considera claras esas políticas (Stanford Institute for Human-Centered Artificial Intelligence, 2026). Estas cifras son útiles para mostrar la velocidad del fenómeno, pero están fuertemente ancladas en Estados Unidos y no constituyen evidencia de efectos sobre aprendizaje. Su función en el ensayo es contextual, no causal.

Convergencia de orientación, no demostración

Tabla 3.

Matriz comparativa de fuentes internacionales seleccionadas

Fuente	Función principal	Qué observa o prescribe	Inferencia legítima	Límite principal
PISA 2025 Vol. I	Evaluación a gran escala	Frecuencia y propósito de uso de chatbots, oportunidades escolares de evaluación de IA, desempeño	Asociación poblacional de distribución de IA,	No identifica causalidad ni microprocesos de interacción
TALIS 2024	Encuesta docente internacional	Uso profesional de IA, finalidades, barreras y necesidades de formación	Distribución y asociación	IA se mide en y sentido amplio; datos autorreportados

Fuente	Función principal		Qué observa o prescribe	Inferencia legítima	Límite principal
UNESCO 2023-2024	Orientación normativa		Principios, competencias y condiciones para uso responsable	Prescripción fundamentada	No estima efectos de aprendizaje
OECD Digital Education Outlook 2026	Síntesis evidencia	de	Desempeño asistido, aprendizaje, diseño pedagógico y políticas	Integración condicionada por fuentes primarias	Hereda heterogeneidad del corpus sintetizado
AI Index 2026	Monitoreo		Adopción, políticas y tendencias	Descripción contextual	Fuerte sesgo geográfico en varios indicadores; no mide aprendizaje
Wu et al. (2026)	Metaanálisis experimental/cuasi		Resultados de aprendizaje bajo intervenciones con ChatGPT	Efecto agregado condicionado	Heterogeneidad muy alta; 134 efectos de 35 estudios; diversidad de intervenciones y posible estructura anidada
Liu et al. (2025)	Metaanálisis		Rendimiento académico bajo intervenciones con ChatGPT	Efecto agregado condicionado	Variación por diseño, disciplina, duración y otras características

La convergencia que emerge no es una demostración causal de que “la mediación mejora el aprendizaje”. Es una convergencia de orientación. UNESCO coloca capacidades humanas y juicio crítico en el centro; PISA muestra heterogeneidad según finalidad, frecuencia y oportunidad escolar; TALIS documenta adopción docente junto con necesidades de preparación; el Digital Education Outlook distingue ejecución asistida de aprendizaje; y los metaanálisis encuentran efectos positivos bajo intervenciones definidas, pero variables. Fuentes metodológicamente distintas coinciden, por vías diferentes, en que la configuración pedagógica importa.

Esa convergencia es más útil para política que una media única. Sugiere que regular el acceso sin intervenir sobre diseño de tareas, alfabetización, evaluación y formación profesional deja intacta la variable pedagógicamente más relevante: qué se hace con la IAGen y qué actividad intelectual se conserva en quien aprende.

Costa Rica: adopción, desempeño y brecha de capacidades

El caso costarricense obliga a matizar una narrativa inicial plausible: la de una adopción intensa de IAGen acompañada por ausencia total de respuesta educativa. PISA 2025 muestra un panorama más complejo. En Costa Rica participaron 6 881 estudiantes de 202 centros educativos, representativos de aproximadamente 54 300 jóvenes de 15 años, con una cobertura estimada del 73% de esa población (OECD, 2026b). El 50% declaró utilizar chatbots al menos semanalmente para ayudarlo a aprender, frente al 46% del promedio OCDE. Para investigación preliminar, la proporción fue del 40% frente al 31%; para resumir textos, del 37% frente al 30%; y para redactar trabajos, del 34% frente al 29%. Solo el 10% indicó no usar nunca o casi nunca chatbots en las tareas examinadas, frente al 14% del promedio OCDE (OECD, 2026b).

La utilización estudiantil es, por tanto, relativamente alta. También fue ligeramente superior el indicador de oportunidad escolar: el 66,1% del alumnado costarricense reportó que en sus lecciones evaluó, al menos algunas veces, la calidad de información generada por chatbots de IA, frente al 62,6% del promedio OCDE (OECD, 2026c, Tabla I.6). El dato indica que no es adecuado describir el sistema simplemente como un contexto de uso elevado sin respuesta educativa declarada. Sin embargo, el indicador no informa la calidad, profundidad, frecuencia ni transferencia de esas actividades y, por ello, no permite concluir que Costa Rica disponga de una alfabetización crítica consolidada.

El desempeño ofrece el tercer componente del cuadro. En 2025, Costa Rica obtuvo 424 puntos en ciencias, 387 en matemática y 416 en lectura, por debajo de los promedios OCDE correspondientes. Frente a 2022, el desempeño aumentó en ciencias, mientras que matemática y lectura se mantuvieron aproximadamente estables; las diferencias en estas dos últimas áreas no fueron estadísticamente significativas (OECD, 2026b, 2026c). Aproximadamente el 56% del alumnado alcanzó al menos el nivel 2 en ciencias, el 30% en matemática y el 53% en lectura. Estas dificultades no pueden atribuirse a la IAGen: los problemas de rendimiento anteceden a la expansión de los chatbots y responden a múltiples factores escolares y sociales.

TALIS 2024 añade una dimensión docente que vuelve el caso más interesante. El 52% de docentes costarricenses declaró haber utilizado IA en su trabajo, frente a 36% del promedio OCDE. Entre usuarios costarricenses, 77% la empleó para aprender o resumir temas, 71% para generar planes o actividades y 56% para apoyar estudiantes con necesidades educativas especiales; usos como evaluar o calificar (36%) y revisar datos de participación o desempeño (32%) fueron menos frecuentes (OECD, 2025). A la vez, 55% de docentes reportó una necesidad alta de desarrollo profesional en habilidades para utilizar IA en enseñanza y aprendizaje, una proporción muy superior al promedio OCDE de 29% (OECD, 2025).

El perfil nacional, por tanto, no es de ausencia de contacto tecnológico. El estudiantado reporta un uso de chatbots para trabajo escolar superior al promedio OCDE, mientras que el profesorado declara una utilización de IA en sentido amplio también relativamente elevada; estas medidas no son equivalentes. El alumnado informa, además, oportunidades escolares de evaluación de contenido generado por IA ligeramente superiores al promedio. La tensión se encuentra en otra parte: la adopción y la exposición educativa pueden avanzar más rápido que la consolidación de capacidades profundas para decidir cuándo utilizar la herramienta, cómo verificarla y qué actividad no debe externalizarse.

En conjunto, Costa Rica puede describirse como un contexto de uso estudiantil relativamente alto de chatbots de IA y de uso profesional docente de IA también elevado, sobre una base de desempeño educativo desafiante. La cuestión analítica es determinar si las oportunidades de alfabetización crítica y mediación educativa poseen la calidad suficiente para acompañar esos usos, no atribuir a la IA los resultados de desempeño. PISA permite verificar la existencia declarada de una oportunidad escolar; no permite evaluar la profundidad de esa experiencia.

Esta distinción abre una agenda de equidad. La brecha de IA no debe reducirse al acceso. Si el recurso está ampliamente disponible, la desigualdad puede desplazarse hacia capacidades para evaluar evidencia, reconocer errores, formular problemas, contrastar fuentes y conservar control sobre decisiones epistémicas. PISA muestra a escala internacional que las oportunidades de evaluar información generada por IA se reportan más entre estudiantes aventajados. El promedio nacional costarricense, aunque superior al promedio OCDE, puede ocultar diferencias internas que requieren análisis específico.

La política educativa enfrenta así un problema diferente del que sugeriría una estrategia centrada en prohibición o disponibilidad. La pregunta no es solo cuántos estudiantes usan IAGen, sino qué aprenden a hacer con ella y cómo se distribuyen esas oportunidades entre centros, regiones y grupos sociales. Para el profesorado, la elevada necesidad declarada de formación indica que la apropiación profesional no puede asumirse por el simple hecho de que más de la mitad ya utilice IA. Acceso y competencia son dimensiones distintas.

Costa Rica resulta, en este sentido, un caso especialmente útil para el argumento general del ensayo. La coexistencia de uso intenso, oportunidades declaradas de evaluación crítica y desafíos persistentes de rendimiento impide convertir cualquiera de esas variables en explicación suficiente. Obliga a pasar de indicadores de presencia hacia indicadores de calidad de interacción, diseño de tareas, regulación y verificación.

Discusión: un criterio de cuatro preguntas para interpretar evidencia sobre IAGen y aprendizaje

El recorrido comparativo permite formular la contribución central del ensayo. Antes de decidir si un hallazgo apoya o cuestiona la integración de IAGen, conviene someterlo a cuatro preguntas. El criterio no pretende estimar causalidad ni reemplazar marcos de evaluación de calidad. Su función es disciplinar la interpretación y hacer visible qué información falta.

Primera pregunta: ¿qué tipo de uso de IAGen se observa y con qué granularidad se define? La categoría “uso” solo adquiere sentido pedagógico cuando se especifican finalidad, frecuencia, herramienta, tarea y, cuando sea posible, secuencia de interacción. PISA mejora la granularidad al distinguir propósitos, pero todavía agrupa prácticas diversas. Un experimento puede definir con mayor precisión qué hace el participante. Un marco normativo, en cambio, puede no observar usos empíricos, sino proponer usos deseables.

Segunda pregunta: ¿qué trabajo cognitivo permanece en el estudiante y qué parte puede externalizarse? Esta pregunta impide identificar tecnología con aprendizaje. El producto final puede mejorar porque el sistema realizó operaciones que antes correspondían al estudiante. La externalización puede ser legítima si libera recursos para una actividad de mayor nivel, o problemática si sustituye la habilidad que se pretendía desarrollar. La respuesta exige analizar la tarea, no solo la herramienta.

Tercera pregunta: ¿qué mediación, alfabetización, regulación o verificación acompaña la interacción? La misma herramienta puede integrarse en una actividad guiada por criterios de contraste, recibir un rol tutorial delimitado o utilizarse sin acompañamiento. La mediación incluye diseño de tareas, instrucciones, reglas de verificación, retroalimentación y competencias previas. PISA aproxima una parte de esta dimensión mediante el indicador de evaluación de información generada por IA, pero no captura la arquitectura pedagógica completa.

Cuarta pregunta: ¿qué tipo de inferencia autoriza el diseño que produjo la evidencia? Esta pregunta funciona como límite epistemológico. Un estudio transversal autoriza asociación; un experimento bien diseñado puede estimar un efecto dentro de condiciones específicas; un metaanálisis agrega efectos de estudios que cumplen criterios determinados; un marco UNESCO justifica competencias y principios, pero no estima impacto.

Las cuatro preguntas no son una lista de cautelas añadida después del análisis; condensan problemas distintos de validez. La primera interroga si la categoría “uso” representa un constructo suficientemente definido. La segunda examina la distribución de operaciones entre persona y artefacto, cuestión central cuando la descarga cognitiva puede ser funcional o sustitutiva (Risko & Gilbert, 2016). La tercera pregunta por las condiciones de autorregulación y mediación que permiten planificar, monitorear y evaluar la propia actividad (Zimmerman, 2002). La cuarta preserva la frontera entre evidencia y recomendación: que una intervención produzca un resultado bajo ciertas condiciones no garantiza que ese resultado viaje intacto a otro sistema, población o configuración pedagógica (Biesta, 2007; Cartwright & Hardie, 2012).

Aplicar el criterio a PISA modifica de inmediato su lectura. La primera pregunta muestra que se observan cuatro propósitos de uso con categorías de frecuencia, no una exposición única. La segunda revela que PISA no observa directamente qué trabajo cognitivo permanece: resumir puede ser delegación o contraste. La tercera identifica una medida parcial de oportunidad escolar de evaluación crítica, sin información suficiente sobre intensidad o calidad. La cuarta restringe el verbo: las diferencias son asociaciones poblacionales. Así, “usar IA reduce el rendimiento” se convierte en una afirmación más precisa: determinados patrones autorreportados de uso se relacionan con diferencias

de desempeño científico y esas relaciones varían por propósito, frecuencia y oportunidad escolar, sin que el diseño identifique causalidad.

El mismo criterio aplicado a Wu et al. (2026) produce otra interpretación. La primera pregunta obliga a reconocer que los 35 estudios no representan una sola forma de uso de ChatGPT; varían por disciplina, duración y modalidad instructiva. La segunda pregunta exige examinar qué operaciones realiza el estudiante en cada intervención, información que una magnitud agregada no conserva plenamente. La tercera adquiere mayor relevancia porque el tratamiento experimental puede incluir instrucciones, guías, retroalimentación o integración con enseñanza convencional. La cuarta permite utilizar la palabra “efecto”, pero circunscrita a las intervenciones sintetizadas. El resultado $g = 0,670$ deja de ser una propiedad de ChatGPT y se interpreta como promedio de configuraciones heterogéneas, hecho reforzado por un I^2 superior al 90%.

La aplicación a UNESCO es todavía más reveladora. La primera pregunta no busca un uso observado, sino los usos y capacidades que el marco considera educativamente deseables. La segunda aparece como propósito normativo: preservar juicio humano y agencia. La tercera es central, porque el marco se construye precisamente alrededor de alfabetización, ética y competencia profesional. La cuarta impide confundir prescripción con demostración: que una competencia sea razonable y necesaria no implica que su implementación produzca automáticamente una magnitud determinada de aprendizaje. El valor del marco es normativo y curricular, no experimental.



Figura 1. Criterio de cuatro preguntas para interpretar evidencia sobre IAGen y aprendizaje.

Nota. La numeración indica una secuencia analítica de lectura y no representa un mecanismo causal.

El criterio también permite precisar la conexión con educación matemática universitaria sin extrapolar resultados de PISA. En matemática, la distinción entre respuesta correcta y actividad epistémica es especialmente visible. Una IAGen puede producir una derivación algebraica, una demostración o una explicación correcta y mejorar el producto observable sin que ello muestre qué decisiones comprendió, verificó o pudo reconstruir el estudiante. Desde la génesis instrumental, la cuestión relevante no es simplemente si se empleó un artefacto, sino cómo se integró en esquemas de actividad y cómo la orquestación docente organizó esa relación (Rabardel, 1995; Trouche, 2004).

La extrapolación desde estudiantes de 15 años sería injustificada; la extensión teórica, en cambio, es legítima porque el problema de atribución de aprendizaje atraviesa niveles educativos. En educación superior se requieren registros de interacción, análisis de tareas, medidas de aprendizaje sin asistencia y diseños que distingan desempeño aumentado de adquisición de capacidades. Esa agenda confirma además la utilidad del criterio más allá de una herramienta concreta: frente a nuevos modelos, seguirá siendo necesario reconstruir qué uso se estudia, qué trabajo conserva la persona, qué regulación existe y qué puede afirmarse legítimamente.

Conclusiones e implicaciones

La evidencia internacional reciente no sostiene una conclusión unívoca según la cual la IAGen, considerada en abstracto, mejore o perjudique el aprendizaje. PISA 2025 identifica asociaciones diferentes según finalidad y frecuencia de uso; TALIS documenta una adopción docente ya extendida en 2024, junto con necesidades importantes de formación; UNESCO propone capacidades críticas y condiciones de gobernanza; las síntesis de la OCDE distinguen desempeño asistido de aprendizaje; y los metaanálisis encuentran efectos medios positivos bajo intervenciones definidas, aunque heterogéneas. Estas afirmaciones pueden coexistir porque pertenecen a regímenes de evidencia distintos y porque “uso de IA” no designa una intervención única.

Para política educativa, la prioridad debería desplazarse desde la oposición acceso/prohibición hacia la calidad del uso, la mediación y la verificación. La expansión de la IAGen no vuelve prescindibles las capacidades fundamentales; aumenta el valor de aquellas que permiten juzgar información, reconocer límites, comparar evidencias y conservar responsabilidad sobre las decisiones. La equidad debe analizarse, por ello, como distribución de oportunidades para desarrollar esas capacidades y no únicamente como disponibilidad tecnológica. El caso costarricense es ilustrativo: un uso estudiantil de chatbots y un uso docente de IA en sentido amplio relativamente altos, oportunidades declaradas de evaluación crítica superiores al promedio OCDE y un desempeño académico desafiante coexisten sin que ninguno de esos indicadores pueda explicar por sí solo a los otros.

Para la docencia, el problema es de diseño. Una tarea robusta frente a IAGen no es necesariamente aquella que prohíbe utilizarla, sino aquella en la que su utilización no elimina el trabajo cognitivo que se desea desarrollar. Esto exige hacer visibles decisiones, justificar verificaciones, distinguir apoyo de sustitución y, cuando resulte pertinente, observar también desempeño sin asistencia. En matemática, lectura o escritura, la pregunta pedagógica es qué debe seguir haciendo el estudiante para que el producto pueda interpretarse como evidencia de aprendizaje.

Para la investigación, el desafío es aumentar la granularidad. Frecuencia y propósito son útiles, pero insuficientes para explicar mecanismos. Se requieren trazas de interacción, medidas de regulación, análisis de estrategias de verificación, estudios longitudinales y diseños que permitan estimar efectos bajo configuraciones pedagógicas explícitas. También conviene distinguir con mayor precisión aprendizaje duradero, desempeño aumentado durante el uso y percepción subjetiva de utilidad.

Esa agenda exige, además, reportar con mayor transparencia qué hace exactamente la persona con la herramienta, de modo que futuras síntesis puedan comparar configuraciones de actividad y no únicamente nombres de aplicaciones.

El criterio propuesto resume esta posición mediante cuatro preguntas: qué uso se observa, qué trabajo cognitivo permanece, qué mediación acompaña la interacción y qué inferencia autoriza el diseño. Su función no es resolver la controversia, sino evitar que se formule mal. Usar no es aprender; pero tampoco dejar de usar garantiza aprendizaje. Entre ambas afirmaciones se encuentra el objeto que la investigación educativa debe explicar: cómo se distribuye la actividad intelectual cuando parte del trabajo puede ser ejecutada por un sistema generativo y bajo qué condiciones esa relación produce evidencia defendible de aprendizaje.

Referencias Bibliográficas

- Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning*, 7(3), 245–274.
<https://doi.org/10.1023/A:1022103903080>
- Biesta, G. (2007). Why “what works” won’t work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory*, 57(1), 1–22. <https://doi.org/10.1111/j.1741-5446.2006.00241.x>
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.
- Global Education Monitoring Report Team. (2023). *Global education monitoring report 2023: Technology in education: A tool on whose terms?* UNESCO. <https://doi.org/10.54676/UZQV8501>
- Liu, Z., Zuo, H., & Lu, Y. (2025). The impact of ChatGPT on students’ academic achievement: A meta-analysis. *Journal of Computer Assisted Learning*, 41(4), e70096.
<https://doi.org/10.1111/jcal.70096>
- Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO.
<https://www.unesco.org/en/articles/ai-competency-framework-teachers>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.
<https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research>
- Miao, F., Shiohira, K., & Lao, N. (2024). *AI competency framework for students*. UNESCO.
<https://www.unesco.org/en/articles/ai-competency-framework-students>
- OECD. (2025). *Results from TALIS 2024: The state of teaching*. OECD Publishing.
<https://doi.org/10.1787/90df6235-en>
- OECD. (2026a). *OECD Digital Education Outlook 2026: Exploring effective uses of generative AI in education*. OECD Publishing. <https://doi.org/10.1787/062a7394-en>

- OECD. (2026b). *PISA 2025 results (Volume I): Costa Rica*. OECD Publishing.
https://www.oecd.org/en/publications/pisa-2025-results-volume-i-country-notes_2d4ff9ea-en/costa-rica_53f42f10-en.html
- OECD. (2026c). *PISA 2025 results (Volume I): Future-ready students*. OECD Publishing.
<https://doi.org/10.1787/73451bc5-en>
- Rabardel, P. (1995). *Les hommes et les technologies: Approche cognitive des instruments contemporains*. Armand Colin.
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- Stanford Institute for Human-Centered Artificial Intelligence. (2026). *The AI Index report 2026*. Stanford University. <https://hai.stanford.edu/ai-index/2026-ai-index-report>
- Trouche, L. (2004). Managing the complexity of human/machine interactions in computerized learning environments: Guiding students' command process through instrumental orchestrations. *International Journal of Computers for Mathematical Learning*, 9(3), 281–307. <https://doi.org/10.1007/s10758-004-3468-5>
- Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications*, 13, 684. <https://doi.org/10.1057/s41599-026-07019-z>
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2