Education in the Knowledge Society

journal homepage http://revistas.usal.es/index.php/eks/

Ediciones Universidad Salamanca

Tres escenarios para la IA en educación: del apoyo responsable a la cocreación

Three Scenarios for AI in Education: From Responsible Assistance to Co-Creation

Francisco José García-Peñalvo

Departamento de Informática y Automática, Instituto de Ciencias de la Educación, Grupo GRIAL, Universidad de Salamanca, España
http://orcid.org/0000-0001-9987-5584 fgarcia@usal.es
Director Científico / Editor-In-Chief Education in the Knowledge Society Journal

INFORMACIÓN DEL ARTÍCULO

Palabras clave
Inteligencia artificial generativa en educación superior; Alfabetización crítica en inteligencia artificial; Transparencia y trazabilidad; Evaluación auténtica y agencia humana; Gobernanza y marco normativo.

RESUMEN

Este artículo propone una vía pragmática y proporcionada para integrar la inteligencia artificial generativa en la educación superior mediante tres escenarios graduados por autonomía, agencia y riesgo (apoyo responsable, colaboración guiada y cocreación con declaración reforzada) que convierten principios amplios en decisiones docentes verificables y trazables a lo largo del ciclo docente (planificación, creación de materiales, apoyo y evaluación). El hilo conductor es la inteligencia artificial como complemento bajo juicio académico, nunca sustituto, con transparencia (declaración de uso y marcado de contenido sintético), verificación externa de hechos y citas, y equidad e inclusión por diseño, en coherencia con la guía de la UNESCO (visión centrada en las personas, acciones inmediatas y refuerzo de capacidades), el AI Act (Artículo 50 sobre obligaciones de transparencia y marcado), el *Safe AI in Education Manifesto* (supervisión humana, privacidad, precisión, explicabilidad, transparencia) y el marco SAFE (Seguridad, Rendición de cuentas, Justicia y Eficacia) como puente operativo entre política y aula. En el Escenario 1 se priorizan bajo riesgo y alta transparencia; en el 2, la iteración trazable con post-edición humana significativa; en el 3, evidencias robustas y auditoría (*prompts*, versiones, verificación, sesgos/idiomas, revisión humana/pares), con controles reforzados por su mayor impacto. Este gradiente se alinea con la orientación sectorial, que promueve autenticidad, agencia y propiedad del proceso y desaconseja depender de detectores, reforzando diseños que comprueban agencia y trazabilidad. Dos instrumentos facilitan la adopción y la evaluación homogénea. Por un lado, una rúbrica transversal (veracidad y actualidad, trazabilidad, corrección de alucinaciones, equidad e idioma, calidad de interacción) y, por otro lado, listas de verificación por tipo de tarea. El resultado es un mapa operativo para marcar, verificar y documentar con proporcionalidad al riesgo, que permite convertir la inteligencia artificial en oportunidad pedagógica sin ceder en rigor, justicia y responsabilidad.

Keywords
Generative artificial intelligence in higher education; Critical artificial intelligence literacy; Transparency and traceability; Authentic assessment and human agency; Governance and regulatory frameworks.

ABSTRACT

This article proposes a pragmatic and proportionate pathway for integrating generative artificial intelligence into higher education through three scenarios graduated by autonomy, agency, and risk (responsible support, guided collaboration, and co-creation with a strengthened declaration). These scenarios convert broad principles into verifiable and traceable teaching decisions throughout the teaching cycle (planning, material creation, support, and assessment). The common thread is the use of artificial intelligence as a supplement to academic judgment, never as a substitute, with transparency (including disclosure and marking of synthetic content), external verification of facts and citations, and equity and inclusion by design. This is consistent with UNESCO's guidance (a human-centred vision, immediate actions, and capacity building), the AI Act (Article 50 on transparency and marking obligations), the Safe AI in Education Manifesto (human supervision, privacy, accuracy, explainability, transparency), and the SAFE framework (Safety, Accountability, Fairness, and Efficacy) as an operational bridge between policy and the

Recibido, 27/10/2025. Revisado, 27/10/2025. Aceptado, 27/10/2025. Publicado, 13/11/2025.
e-ISSN: 2444-8729

classroom. Scenario 1 prioritises low risk and high transparency; Scenario 2 focuses on traceable iteration with significant human post-editing; and Scenario 3 demands robust evidence and auditing (prompts, versions, verification, bias/language checks, human/peer review), with strengthened controls due to its higher impact. This gradient aligns with sector guidance, which promotes authenticity, agency, and ownership of the process and advises against relying on detectors, thereby reinforcing designs that verify agency and traceability. Two instruments facilitate adoption and consistent evaluation: firstly, a cross-cutting rubric (veracity and currency, traceability, correction of hallucinations, equity and language, and quality of interaction); and secondly, checklists for each type of task. The result is an operational map for marking, verifying, and documenting in proportion to risk, which enables artificial intelligence to be leveraged as a pedagogical opportunity without compromising on rigour, fairness, and responsibility.

1. Introducción

A punto de cumplirse el tercer aniversario del lanzamiento público de ChatGPT (30 de noviembre de 2022), el sistema universitario se enfrenta a una madurez acelerada de la inteligencia artificial generativa (IAGen) (Jovanović & Campbell, 2022): modelos más capaces, mayor integración en herramientas de autoría y programación, y un uso social y académico que ya no es excepcional, sino cotidiano (Chatterji et al., 2025). Este punto de inflexión invita a revisar lo aprendido y, sobre todo, a proponer marcos operativos que permitan pasar de la reacción a la gobernanza pedagógica.

En 2023, el primer editorial de esta serie tomó el pulso a la academia en caliente: el aterrizaje de ChatGPT precipitó un debate polarizado entre la promesa de disrupción y el pánico ante riesgos reales (plagio, opacidad, sesgos, desplazamiento de competencias) (García-Peñalvo, 2023). Allí se afirmó que el diagnóstico debía ir más allá del entusiasmo o la alarma, y pidió situar la conversación en términos de alfabetización crítica, trazabilidad y responsabilidad humana en el uso de IAGen. Ese marco de interpretación, consciente de la novedad, pero exigente con la evidencia y la ética, permitió nombrar con claridad la naturaleza del fenómeno y sus límites.

En 2024, el segundo editorial se dirigió hacia una mirada multiperspectiva, analizando la IAGen desde los roles que conviven en la educación superior: profesorado, estudiantado, responsables académicos y equipos de desarrollo/soporte (García-Peñalvo, 2024a). Este mapa de actores puso de relieve que las decisiones no son homogéneas ni simétricas: difieren los objetivos, la exposición al riesgo, las capacidades de verificación y las obligaciones de transparencia. A partir de esta cartografía, se enfatizó que las políticas institucionales debían alinearse con prácticas de aula y con un diseño instruccional que hiciera explícitas las agencias (humana y de la herramienta), la citación de fuentes y el registro de las contribuciones de la inteligencia artificial (IA).

Esta tercera entrega, realizada en el último trimestre de 2025, trae consigo, además, un contexto regulatorio y de orientaciones internacionales más definido. El Reglamento de Inteligencia Artificial de la Unión Europea 2024/1689 (AI Act) (European Parliament & The Council of the European Union, 2024) introduce un enfoque basado en riesgos que incide directamente en las obligaciones de documentación, transparencia y supervisión; aunque su foco no es educativo per se, su espíritu permea las políticas universitarias y los requisitos de compra y despliegue tecnológico. En paralelo, la Guía de la UNESCO sobre IAGen en educación e investigación (UNESCO, 2023) reclama una visión centrada en las personas, capacidades institucionales y desarrollo docente, así como medidas inmediatas para mitigar los riesgos y cerrar las brechas de preparación. Esta convergencia normativa y programática ofrece una base más sólida para transitar de la improvisación a la implementación responsable.

Sobre esta base, se propone una evolución: organizar el uso educativo de la IAGen en tres escenarios, dirigidos por la autonomía de la herramienta, la agencia humana y la exposición al riesgo, que puedan funcionar como lenguaje común entre roles y como guía práctica para decidir qué uso es razonable, cómo hacerlo verificable y con qué salvaguardas. La aspiración no es añadir otra taxonomía más, sino facilitar la adopción: enlazar cada escenario con ejemplos de actividades, criterios de divulgación, mecanismos de trazabilidad (desde el registro de *prompts* hasta la citación verificable) y métricas que permitan evaluar el aprendizaje, la calidad y el coste de oportunidad.

Si el primer artículo (García-Peñalvo, 2023) se centró en nombrar el fenómeno y el segundo artículo (García-Peñalvo, 2024a) se orientó a situarlo en sus roles, este tercero lleva a una organización en escenarios que sirvan para decidir y actuar. El objetivo del artículo es, por tanto, doble: definir estos tres escenarios (del apoyo responsable a la cocreación) y proveer una hoja de ruta que permita al profesorado (y a las instituciones educativas) convertir principios y regulaciones en prácticas verificables, evaluables y sostenibles en el tiempo.

2. Alfabetización crítica de la IAGen en la educación superior

2.1. Qué es y no es la IAGen

La IAGen se refiere a modelos de IA capaces de producir contenidos sintéticos inéditos, en cualquier forma y para apoyar cualquier tarea, mediante modelización generativa (García-Peñalvo & Vázquez-Ingelmo, 2023). Estos modelos, para conseguir las prestaciones en el manejo del lenguaje natural (aunque su capacidad de generación ya es multimodal en la actualidad) que han llevado a la IAGen a su posicionamiento, deben tener un tamaño muy grande, de miles de millones de parámetros, de ahí su nombre *Large Language Models* (LLM) (Zhao et al., 2025), definiéndose como sistemas de IA de vanguardia que pueden procesar y generar texto con una comunicación coherente y generalizar a múltiples tareas (Naveed et al., 2025). Formalmente, los parámetros son los pesos de las capas de las redes neuronales, es decir, elementos intrínsecos del modelo que se ajustan para optimizar su rendimiento en la tarea de predicción de la siguiente palabra o secuencia de texto, basándose en el contexto previo. El número de estos parámetros puede variar desde millones hasta billones, lo que influye directamente en la capacidad y la complejidad del LLM.

Ejemplos destacados de LLM podrían ser GPT-5 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025) o modelos abiertos más recientes como DeepSeek-V3.2 (DeepSeek, 2025). En contextos educativos, la IAGen suele aludir especialmente a estas herramientas de generación de texto (por ejemplo, *chatbots* tipo ChatGPT), que se emplean para responder preguntas, redactar ensayos o resumir información, entre otras muchas cosas.

Tan importante como conocer qué es la IAGen es entender lo que no es. Estos sistemas no son mentes conscientes ni oráculos infalibles, sino algoritmos estadísticos muy avanzados. Su funcionamiento se basa en predecir la siguiente palabra en una secuencia usando el contexto, imitando patrones del lenguaje natural. No comprenden realmente el significado de lo que dicen ni poseen conocimiento verdadero; simplemente generan texto con apariencia de sentido a partir de correlaciones en sus datos de entrenamiento. Por tanto, no se debe confundir su fluidez con veracidad o entendimiento genuino.

La IAGen no es "magia" omnisciente ni el *Palantír* del universo de Tolkien (Alier-Forment et al., 2026), por más que para un grupo importante de sus usuarios sea válida la denominada Tercera Ley de Clarke "cualquier tecnología lo suficientemente avanzada es indistinguible de la magia" (Clarke, 1973); sigue siendo una tecnología con alcances y limitaciones específicas que deben conocerse. La IAGen bien utilizada puede ser una importante herramienta para mejorar la calidad y la equidad de la educación superior, pero mal utilizada, sin comprensión crítica o sin límites éticos, puede generar el efecto contrario (García-Peñalvo, Llorens-Largo, et al., 2024; Lee et al., 2024).

La alfabetización crítica en IAGen consiste precisamente en dotar a docentes y estudiantes de esa mirada informada (Castañeda & Selwyn, 2018; Veldhuis et al., 2025), es decir, aprovechar su potencial con cautela y conocimiento, sabiendo lo que la IAGen es y lo que no es.

2.2. Cómo funciona: modelos, datos y entrenamiento

Los modelos generativos de lenguaje modernos se entrenan mediante técnicas de aprendizaje profundo (redes neuronales) con arquitecturas tipo *Transformer* (Vaswani et al., 2017, 2023). Introducido por Google en 2017, el *Transformer* permite procesar secuencias largas de texto prestando atención a la relación entre palabras. Modelos como GPT-5 cuentan con cientos de miles de millones de parámetros ajustables, lo que les da una enorme capacidad para modelar el lenguaje humano. En la práctica, primero se someten a un preentrenamiento masivo no supervisado (Brown et al., 2020), esto implica que se les alimenta con un corpus de texto muy grande y aprenden a predecir la siguiente palabra en base a lo leído, capturando patrones gramaticales y semánticos. Esta amplitud en el corpus de entrenamiento es precisamente lo que les permite responder sobre ciencias, historia, arte o tecnología, según el contexto de la pregunta. Se aplican además filtros durante la recopilación de datos para excluir lenguaje altamente soez o sesgado, aunque inevitablemente muchos sesgos y desequilibrios presentes en Internet acaban incorporados en el modelo (Bender et al., 2021; Weidinger et al., 2021).

Tras el preentrenamiento, suele venir un ajuste fino. Una técnica común es entrenar al modelo con ejemplos de instrucciones y respuestas de alta calidad, para que aprenda a seguir indicaciones humanas (lo que se llama *instruction tuning* (Wei et al., 2022)). Adicionalmente, se emplea aprendizaje por refuerzo con retroalimentación humana (*Reinforcement Learning from Human Feedback* – RLHF) (Christiano et al., 2017; Ouyang et al., 2022): evaluadores humanos califican respuestas del modelo, y esa señal entrena un modelo de recompensa que guía al LLM para alinear sus respuestas con las preferencias y valores humanos.

Todo este proceso requiere infraestructura computacional de enormes dimensiones, supercomputadoras con miles de GPU (*Graphics Processing Unit*) trabajando en paralelo durante bastante tiempo (Brown et al., 2020). Por eso, solo organizaciones con grandes recursos (empresas tecnológicas, centros de investigación) han podido desarrollar estos LLM, lo que hace que su interior opere como una caja negra para el usuario común.

Además, esta escala masiva tiene un coste ecológico considerable y a menudo invisible (Google, 2025). El funcionamiento ininterrumpido de miles de procesadores genera una demanda energética enorme, que contribuye a una huella de carbono significativa (Dhar, 2020; Schwartz et al., 2020), especialmente si la electricidad proviene de fuentes no renovables. A esto se suma un enorme consumo de agua dulce (Li et al., 2025; Qiao et al., 2025), utilizada para la refrigeración de los centros de datos que evitan el sobrecalentamiento de los servidores. Cada interacción con un modelo de IA, desde su entrenamiento hasta el uso diario, consume indirectamente estos valiosos recursos (Jegham et al., 2025).

En esencia, un LLM funciona prediciendo texto muy verosímil según las probabilidades derivadas de su entrenamiento. Esto explica tanto su potencia como sus problemas. Por un lado, puede generar respuestas detalladas sobre infinidad de temas, imitando estilos discursivos diversos. Por otro lado, no tiene un mecanismo garantizado de verificación de hechos, más allá de lo que quedó grabado en sus pesos durante el entrenamiento, aunque la mayoría de las versiones de los *chatbots* más extendidos permiten la consulta de Internet para ofrecer unas mejores respuestas con datos factuales actualizados. Esto puede dar lugar a la generación de contenido que, aunque parezca coherente y plausible, es incorrecto o no está fundamentado en datos reales. Esto es lo que se conoce como alucinaciones (Perković et al., 2024; Towhidul Islam Tonmoy et al., 2024). Además, la ausencia de un enlace directo a sus fuentes originales de información (salvo en los casos en los que se utiliza la búsqueda directa en fuentes específicas, ya sean bases de conocimiento o Internet) hace que no pueda citar ni atribuir correctamente las ideas o datos que proporciona. Todos estos factores tecnológicos subrayan la importancia de que los usuarios (docentes y estudiantes en el contexto académico) entiendan cómo operan estos modelos, para poder usarlos de forma informada, conscientes de qué esperar (y qué no) de sus respuestas.

2.3. Riesgos en la práctica universitaria

Integrar la IAGen en la educación superior ofrece ventajas indudables, pero también implica riesgos concretos que una alfabetización crítica debe anticipar. En la universidad, estos riesgos van desde la desinformación involuntaria (por ejemplo, resúmenes que distorsionan resultados científicos o alucinaciones plausibles) hasta dilemas éticos y pedagógicos ligados a integridad, trazabilidad y evaluación. Lejos de justificar su prohibición, constituyen una llamada de atención para integrar estas herramientas con espíritu crítico y cautela informada, reforzando prácticas de verificación, transparencia y diseño instruccional responsable. Todo ello exige identificar las alucinaciones, reconocer sesgos y contrarrestarlos, mantener la autonomía intelectual del estudiantado y salvaguardar principios académicos (veracidad, citación y originalidad) mediante orientaciones claras y desarrollo competencial.

Algunos de los principales riesgos identificados en la práctica académica son (García-Peñalvo, 2024a; García-Peñalvo, Llorens-Largo, et al., 2024):

- **Alucinaciones y veracidad**: Los LLM pueden alucinar contenido, es decir, generar información incorrecta o inexistente presentada con total aparente seguridad (Ji et al., 2023). Por ejemplo, es común que inventen referencias bibliográficas o citas académicas que suenan verosímiles, pero no existen en la realidad. Estudios comparativos han encontrado que incluso modelos avanzados como GPT-4 producen una proporción preocupante de referencias fabricadas (cerca del 28% en pruebas específicas), mientras que otros modelos como Bard han superado el 90% de citas falsas en entornos de revisión sistemática (Chelli et al., 2024). Esta falta de veracidad automatizada supone un peligro en contextos universitarios, que puede llevar al estudiantado a tomar como cierto un dato erróneo bajo la creencia de que la IA nunca se equivoca y basar un trabajo en información falsa, es decir, el sesgo por automatización (*automation bias*) al que los humanos son proclives al confiar ciegamente en los sistemas automatizados pasando por alto su propio juicio (Romeo & Conti, 2025). Sin mecanismos internos de trazabilidad, resulta difícil verificar de dónde proviene cada afirmación o detectar el error sin invertir tiempo en corroborar con fuentes externas (Huang & Chang, 2024; Shao, 2025). La tendencia de la IA a "decir algo" aunque no tenga datos fiables, incluida la fabricación segura de citas, exige que el académico deba ser escéptico por defecto y deba contrastar siempre la información (Gibney, 2025; Peters & Chin-Yee, 2025).

- **Sesgos y equidad**: Los modelos generativos heredan los sesgos presentes en sus datos de entrenamiento. Esto significa que pueden reproducir prejuicios históricos o sociales en sus respuestas. Por ejemplo, podrían asociar estereotipos de género o raza a ciertas profesiones, o ignorar perspectivas de regiones subrepresentadas en la literatura académica (An et al., 2025; Torres et al., 2025). Si se usan sin cuidado, podrían reforzar desigualdades o dar información parcial. Además, al estar mayormente entrenados en idiomas como el inglés, algunas IAGen rinden peor en otros idiomas, especialmente minoritarios, lo que plantea brechas lingüísticas (Roxas, 2024; Xu et al., 2025). El profesorado debe ser consciente de estos sesgos para mitigarlos, ya sea filtrando las salidas de la IA, ajustando los *prompts* o complementando con fuentes diversas. La inclusión es un punto clave; sin vigilancia, la IA podría perpetuar solo las voces dominantes de su corpus (Afreen et al., 2025; UNESCO, 2023).
- **Dependencia y disminución del aprendizaje**: Existe el riesgo de una dependencia excesiva del estudiantado y del profesorado hacia estas herramientas, delegando en la IA tareas cognitivas fundamentales (búsqueda, contraste, síntesis y argumentación), lo que puede erosionar la práctica de dichas habilidades. La evidencia reciente sugiere que la sobreconfianza y el usar recursos externos, como herramientas físicas o digitales, para reducir el esfuerzo mental necesario para realizar una tarea (*cognitive offloading*) (Risko & Gilbert, 2016), están asociados con menor pensamiento crítico y menor implicación cognitiva, especialmente cuando el estudiantado acepta respuestas sin verificarlas (Gerlich, 2025; Zhai et al., 2024), induciendo una pereza intelectual. Esta dinámica se manifiesta en la tendencia a aceptar la primera salida de la IA como suficiente (*automation bias*), a citarla como si fuese una autoridad y a usarla para redactar trabajos con mínima aportación personal, lo que incrementa riesgos de integridad académica y de aprendizaje superficial (Bittle & El-Gayar, 2025). En términos neurocognitivos, incluso se han observado indicios de menor conectividad funcional cuando la escritura académica se realiza con ayuda constante de IA, en comparación con condiciones sin herramientas, lo que refuerza la necesidad de un uso formativo y no sustitutivo (Bai et al., 2023). Los marcos de alfabetización en IA y las revisiones sobre desinformación educativa insisten en diseñar actividades que obliguen a contrastar, justificar y documentar el uso de IA para evitar dependencia acrítica y sus consecuencias evaluativas (Fulsher et al., 2025). Académicamente, esto se traduce en un doble desafío: por un lado, el alumnado puede experimentar frustración y sanciones cuando una respuesta defectuosa de la IA se introduce sin verificación; por otro, el profesorado puede deslizarse hacia la delegación de tareas nucleares de diseño y retroalimentación, perdiendo oportunidades para promover metacognición y juicio experto. La alfabetización crítica debe, por tanto, delimitar cuándo es apropiado usar IA (por ejemplo, para una tormenta de ideas o como apoyo a la planificación) y cuándo es esencial que el estudiantado realice el trabajo por sí mismo para alcanzar las competencias del plan de estudios; además, debe establecer políticas de transparencia (declarar el uso), trazabilidad (registro de *prompts* y fuentes) y verificación (contraste con literatura académica) como salvaguardas mínimas (Fulsher et al., 2025).
- **Falta de transparencia y citación**: Como se ha señalado, los modelos actuales no revelan de forma nativa sus fuentes, lo que dificulta la atribución y la rendición de cuentas académica; de ahí que la literatura proponga integrar mecanismos de citación/justificación en los LLM como salvaguarda (Huang & Chang, 2024). A diferencia de un buscador que lista documentos, un LLM fusiona lo aprendido y genera una respuesta unitaria sin referencias verificables, una opacidad que choca con los valores universitarios de rastrear la evidencia y citar correctamente. En la práctica, se observa un riesgo creciente: el estudiantado puede incluir información generada por la IA sin citarla o, peor aún, presentar bibliografías fabricadas por el modelo; varios trabajos recientes documentan tasas elevadas de citas falsas y esfuerzos por mitigarlas (Gibney, 2025; Glynn, 2025). Además, existen implicaciones de propiedad intelectual. Los modelos pueden memorizar y reproducir texto de entrenamiento, incluida obra con *copyright*, especialmente bajo ciertos ataques o solicitudes, lo que plantea riesgos legales y la necesidad de defensas y controles (Liu et al., 2024; Mueller et al., 2024). La ausencia de atribución automática complica evaluar la confiabilidad de lo ofrecido. En este contexto, resulta imprescindible que las instituciones definan cuándo y cómo citar a la IA (por ejemplo, como apoyo/tutoría o como fuente que requiere mención explícita), y que exijan divulgación del uso, trazabilidad (registros de *prompts* y fuentes) y verificación externa como buena práctica docente (UNESCO, 2023).

2.4. Principios operativos de alfabetización crítica

La alfabetización crítica en IAGen no se reduce a saber usar herramientas, sino a usar con juicio dentro de valores y prácticas académicas. Como cierre del marco conceptual, se proponen cuatro principios operativos que guiarán el resto del artículo:

1. **Verificación antes de la adopción**: toda salida generada se contrasta con fuentes académicas verificables; se evita incorporar afirmaciones o citas sin trazabilidad (Huang & Chang, 2024; UNESCO, 2023).
2. **Agencia humana explícita**: la IA asiste, no sustituye la deliberación, el análisis y la autoría; se diseñan tareas que exigen juicio humano y reflexión metacognitiva (Peters & Chin-Yee, 2025).
3. **Equidad e inclusión por diseño**: se anticipan sesgos y brechas lingüísticas; se ajustan los *prompts* y los materiales para incorporar voces subrepresentadas y lenguas del contexto (An et al., 2025; Roxas, 2024).
4. **Transparencia y rendición de cuentas**: toda intervención con IA deja rastro (justificación breve, registro de *prompts*/fuentes, versión final revisada) y cumple con obligaciones de documentación y supervisión (European Parliament & The Council of the European Union, 2024; UNESCO, 2023).

Estos principios se operativizan en la Sección 3, donde se concretan competencias, evidencias y herramientas de aula para desplegar la IAGen con garantías, además de conectarse con los tres escenarios de uso responsable que articulan el núcleo del artículo.

3. Competencias de una alfabetización crítica en IAGen

Partiendo del marco expuesto en 2.1-2.3, esta sección traduce esos fundamentos en prácticas docentes verificables (Artopoulos & Lliteras, 2024). El objetivo es sostener calidad, equidad y trazabilidad del conocimiento mediante cuatro elementos: agencia humana explícita, verificación sistemática, inclusión por diseño y transparencia (European Parliament & The Council of the European Union, 2024; Huang & Chang, 2024; UNESCO, 2023). Las competencias se articulan en cinco ámbitos interconectados: 1) comprensión funcional; 2) evaluación crítica y verificación; 3) interacción eficaz; 4) ética, privacidad y cumplimiento; y 5) equidad e inclusión.

Comprensión funcional. No es necesario dominar la ingeniería de la IA; basta poder explicar en lenguaje llano el funcionamiento y los límites ya resumidos en 2.1–2.3 y reconocer cuándo una salida exige verificación externa (Kassorla et al., 2024; UNESCO, 2023). Esta base desactiva la "magia" y el efecto de "caja negra" y permite integrar la IAGen con criterio en los flujos de trabajo docentes (selección de herramientas, diseño de actividades, criterios de verificación). Como práctica inicial, resulta útil disponer de una ficha del modelo que se vaya a emplear (fecha de corte, idiomas mayoritarios o minoritarios, modos con/sin búsqueda, restricciones de uso), accesible al estudiantado y al equipo docente (Burneo-Arteaga et al., 2025; Yang et al., 2025).

Evaluación crítica y verificación. La verificación pasa de ser una excepción para convertirse en hábito. La literatura reciente muestra que los LLM pueden distorsionar o sobregeneralizar resultados científicos incluso cuando se les pide precisión (Peters & Chin-Yee, 2025) y que la fabricación de citas no es anecdótica (Gibney, 2025). En el contexto universitario, esto transmite errores con apariencia de erudición si no se contrarresta (Zhai et al., 2024). La respuesta pedagógica no es prohibir (García-Peñalvo, Llorens-Largo, et al., 2024), sino diseñar la verificación de forma que toda salida generada sea cotejada con fuentes académicas verificables antes de aceptarse en una tarea (Huang & Chang, 2024) y se documente la revisión de forma breve. Los marcos internacionales recomiendan protocolos de transparencia, verificación y formación docente específica para no sacrificar validez, actualidad y equidad (Frau-Meigs, 2024; UNESCO, 2023). Para sistematizar el proceso, el centro puede adoptar la rúbrica transversal que se muestra en la Tabla 1 y pedir una lista de verificación por tipo de tarea (presentada en la Tabla 2).

Interacción eficaz (*prompts* y post-edición). La interacción con IAGen se apoya en *prompts* bien planteados y en una post-edición responsable. Un buen *prompt* no es una fórmula secreta, sino criterio educativo: aporta contexto (curso, nivel, objetivo), restricciones (extensión, formato, precisión) y expectativas sobre fuentes ("si mencionas literatura, sugiere títulos con DOI; se verificarán después") (Boonstra, 2025; Kotha et al., 2025). Incorporar la ingeniería de *prompts* en el marco curricular universitario ha demostrado beneficios claros (Knoth et al., 2024; Lee & Palmer, 2025). Aun así, la clave está en la post-edición: detectar alucinaciones, completar referencias reales, afinar el razonamiento y firmar el resultado con autoría humana (Bedington et al., 2024; Nguyen et al., 2024). Este patrón alinea uso y aprendizaje: la IAGen ayuda a comenzar; el conocimiento se consolida al revisar. Las evidencias de aprendizaje pueden incluir la versión anotada con correcciones, la lista de fuentes verificadas y una breve reflexión sobre decisiones de edición.

Tabla 1. Propuesta de rúbrica transversal para uso de IAGen en tareas académicas (0-2 por ítem; total 0-10).

Ítem	0 (Insuficiente)	1 (Adecuado)	2 (Excelente)
1. Veracidad y actualidad de fuentes	No hay verificación; se incorporan datos/citas no comprobadas o desactualizadas.	Verificación parcial; se corrigen errores obvios, pero persisten lagunas o fuentes débiles.	Todas las afirmaciones clave trianguladas con DOI/ISBN/URL académicas vigentes; cifras y citas contrastadas.
2. Trazabilidad (declaración de uso + registro)	No se declara el uso de IA ni se conservan *prompts*/versiones.	Declaración de uso genérico sin detalles; registro incompleto.	Declaración de uso claro (herramienta, fase, límites) y registro auditable (*prompts*, versiones, decisiones).
3. Corrección de alucinaciones/errores	Se detectan alucinaciones/citas falsas sin corrección.	Se corrigen alucinaciones principales; quedan faltas menores.	Se identifican y corrigen todas las alucinaciones; se documenta el proceso de corrección.
4. Equidad e inclusión (voces/idiomas)	Fuentes y ejemplos anglocéntricos o poco diversos; sesgos no atendidos.	Alguna diversificación de fuentes; atención parcial a sesgos/idioma.	Fuentes y ejemplos diversos (género/región/escuela); atención explícita a la lengua de docencia.
5. Calidad de la interacción con IA (*prompt* + post-edición)	*Prompt* vago; salida adoptada sin revisión.	*Prompt* con criterios básicos; post-edición limitada.	*Prompt* con contexto, criterios y restricciones; post-edición sustantiva (estructura, precisión, estilo, citación real).

Interpretación sugerida:

0-3 = Rehacer con supervisión; 4-6 = Aceptable con mejoras; 7-8 = Bueno; 9-10 = Excelente.

Tabla 2. Listas de verificación por tipo de tarea.

Resumen de artículo científico
◻ Se identifica el artículo original (PDF) y el DOI. ◻ Se verifica objetivo, método, muestra/datos y resultados frente al PDF. ◻ Se comprueban cifras, tablas y gráficos (valores idénticos; sin "promedios creativos"). ◻ Se detectan y marcan exageraciones o sobregeneralizaciones; se corrigen. ◻ Las citas inventadas por la IA se reemplazan por referencias reales o se eliminan. ◻ Se registra en un anexo: *prompt* → salida → correcciones → fuentes.
Ensayo argumentativo
◻ Se declaran la tesis y los criterios (alcance, límites, definición de términos). ◻ Se verifican hechos y citas académicas con DOI/ISBN/URL. ◻ Se integran voces no dominantes (autoras/regiones/escuelas) relevantes. ◻ Se marcan y revisan posibles sesgos en ejemplos/analogías. ◻ Se ha reescrito con su propio estilo; se indica qué aportó la IA (reconocimiento). ◻ Se registran en un anexo las decisiones (por qué se aceptaron/rechazaron partes de la salida).
Problema cuantitativo / análisis de datos
◻ Se replica paso a paso el procedimiento (cálculo o código) y se anexa evidencia. ◻ Se validan unidades, redondeos y propagación de errores. ◻ Si hay código: se han ejecutado pruebas mínimas y se han añadido comentarios. ◻ Se compara con la fuente oficial o estándar (manual, base oficial). ◻ Se registran discrepancias y cómo se resolvieron. ◻ Se registra en un anexo: *prompt* → salida → correcciones → fuentes.
Código / ingeniería
◻ Se especifican los requisitos (entrada/salida, complejidad, seguridad, licencias). ◻ Se ha hecho revisión estática y dinámica (tests unitarios). ◻ Se citan fragmentos de código externos si procede (licencia). ◻ Se analizan riesgos (dependencias, datos, vulnerabilidades). ◻ Se registra en un anexo: *prompt* → versiones → justificación de cambios → resultados de tests.
Imagen / material visual
◻ Se indica procedencia (propia, generada, banco) y licencia; si aplica, metadatos C2PA. ◻ Se verifica fidelidad (mapas/gráficos: escalas, leyendas, proyecciones, unidades). ◻ Se evitan estereotipos visuales (representaciones sesgadas). ◻ Se anotan transformaciones (edición, *upscaling*, *inpainting*). ◻ Se registran fuentes y permisos. ◻ Se registra en un anexo: *prompt* → versiones → justificación de cambios.

Ética, privacidad y cumplimiento. Este eje no es accesorio. El Reglamento (UE) 2024/1689 consolida un enfoque basado en riesgos que, aunque no es específico para educación, impacta en criterios de adquisición, despliegue y documentación universitaria (European Parliament & The Council of the European Union, 2024). En el aula, implica conocer los límites de la asistencia (integridad académica), no introducir datos personales o especialmente protegidos en servicios sin garantías, y priorizar despliegues institucionales o modos privados cuando proceda (Dúo-Terrón, 2024; Hayes et al., 2025; Nam & Bai, 2023). También exige atender a la propiedad intelectual y uso legítimo de materiales (Liu et al., 2024). Toda intervención con IAGen debe dejar rastro: declaración de uso (*disclosure*) breve, registro interno de *prompts* y decisiones, y versión final revisada. Este rastro posibilita auditar, aprender y mejorar (García-Peñalvo et al., 2025).

Equidad e inclusión. Es el ámbito que con más facilidad se olvida y, sin embargo, condiciona la calidad académica a medio plazo. Los LLM heredan sesgos y pueden reproducir estereotipos de género o raza (An et al., 2025). Además, persisten brechas lingüísticas, como por ejemplo, el rendimiento decrece en lenguas con menos recursos y en corpus no anglófonos, con efectos directos en la calidad de lo generado (Roxas, 2024; Xu et al., 2025). La inclusión no se delega; se diseña. En la práctica, esto significa formular *prompts* con "lentes de equidad" ("incluye autoras y regiones subrepresentadas; evita sesgos"); alternar modelos/modos cuando la lengua de docencia no es el inglés; diversificar fuentes para no perpetuar el canon dominante; y premiar explícitamente esta diversificación en la evaluación. La tabla de verificación por tipo de tarea de la Tabla 2 incorpora estos controles de forma operativa.

La alfabetización crítica parece costosa si se improvisa; no lo es si se estandariza. Dos instrumentos pueden ayudar a aliviar la carga y convertir las expectativas en rutina compartida. Con estos dos elementos, cualquier docente puede valorar cómo se ha usado la IAGen, no solo qué se ha entregado:

- La rúbrica transversal presentada en la Tabla 1 (cinco ítems: veracidad/actualidad; trazabilidad mediante declaración de uso y registro; corrección de alucinaciones; equidad/inclusión; calidad de la interacción con IAGen), con escala 0-2.
- Las listas de verificación presentadas en la Tabla 2 (resumen de artículo, ensayo, problema cuantitativo, código, imagen) especifican qué comprobar y cómo evidenciarlo.

La formación docente cierra el círculo. Dado lo vertiginoso del campo, la alfabetización en IAGen exige una actitud de actualización permanente porque evolucionan los modelos y las salvaguardas y surgen nuevas regulaciones y guías de buenas prácticas con impacto en docencia y evaluación (European Parliament & The Council of the European Union, 2024; Kassorla et al., 2024; UNESCO, 2023). La capacitación continua, las comunidades de práctica y el ajuste metodológico son condiciones para una integración responsable (Abegglen et al., 2024; Jin et al., 2025; Nerantzi et al., 2023; Sozon et al., 2025).

Conviene subrayar que esta alfabetización crítica no compite con los escenarios de uso que articulan el núcleo del artículo; los habilita. En el Escenario 1 (apoyo responsable), basta con verificación básica y declaración de uso breve para asegurar calidad sin sobrecarga. En el Escenario 2 (colaboración guiada), el foco está en la iteración trazable (diarios de aprendizaje, rúbricas explícitas, comparación entre versiones). En el Escenario 3 (cocreación con declaración de uso reforzada), la trazabilidad es total: *prompts*, versiones, fuentes, decisiones y revisión por pares. En todos, la agencia humana es la bisagra que convierte la capacidad técnica en práctica académica con sentido.

La alfabetización crítica en IAGen convierte la promesa y los límites señalados en las secciones 2.1-2.3 en hábitos profesionales: comprender lo suficiente para evitar la ilusión de infalibilidad; verificar antes de adoptar; diseñar la interacción para aprender, no solo para producir; respetar privacidad y normativa; y asegurar que todas las voces (lenguas, regiones, perspectivas) entran en el aula. Es fundamental promover la alfabetización en IAGen, fomentando competencias digitales responsables en la comunidad educativa (Vivas Urias & Ruiz Rosillo, 2025). Al dominar estas competencias, el profesorado podrá convertir la llegada de la IAGen en oportunidad pedagógica, enriqueciendo la enseñanza con innovación y huyendo de mitos y amenazas (García-Peñalvo, 2024b). Con los instrumentos presentados en la Tabla 1 y en la Tabla 2, estas prácticas se vuelven transmitibles, observables y evaluables. Así, la universidad se prepara para un futuro en el que la IAGen se integra en la creación y difusión del conocimiento ampliando pensamiento crítico, creatividad y rigor, nunca limitándolos.

4. Normas y garantías

Esta sección enlaza los hábitos profesionales de la Sección 3 con cuatro marcos que hoy constituyen una referencia en la aplicación de IA de forma segura en el contexto académico: las orientaciones globales de la UNESCO

(2023), el Reglamento (UE) 2024/1689 (AI Act) (European Parliament & The Council of the European Union, 2024), el marco SAFE (*Safety, Accountability, Fairness, Efficacy*) (EDSAFE AI, 2021) y, como columna vertebral, el *Safe AI in Education Manifesto* (Alier et al., 2024). Todos ellos implican la toma de decisiones, ya sea a nivel de aula o de centro, con un foco en los principios que subyacen en las secciones anteriores, esto es, la importancia de la agencia humana, el aseguramiento de la privacidad y la seguridad, el compromiso ético y la transparencia.

4.1. El Manifiesto como hilo conductor

El Manifiesto *Safe AI in Education* (Alier et al., 2024) plantea principios de supervisión humana y derecho de apelación, confidencialidad, alineación con la estrategia y la didáctica, precisión/explicabilidad, transparencia en interfaz y comportamiento y entrenamiento ético y transparencia de datos (García-Peñalvo, Alier, et al., 2024). El espíritu es claro, la IA complementa al profesorado, nunca lo sustituye; el contenido generado ha de marcarse y ser verificable; y el centro debe poder auditar cómo se usó la herramienta. Esto encaja con los cuatro vectores propuestos (agencia, verificación, inclusión y transparencia) y con los instrumentos de las Tablas 1 y 2.

Este manifiesto enlaza con los escenarios propuestos de la siguiente manera:

- **Escenario 1 (apoyo responsable)**: aplicar la precisión, la explicabilidad y la transparencia mediante declaración de uso breve, marcado visible de lo generado y verificación básica antes de incorporar la salida.
- **Escenario 2 (colaboración guiada)**: reforzar la supervisión, el derecho a la apelación y la alineación didáctica mediante iteraciones con registro, revisión del docente y posibilidad de que el alumnado discuta/rectifique inferencias.
- **Escenario 3 (cocreación con declaración de uso reforzado)**: criterios explícitos sobre datos/fuentes/sesgos; registro de auditoría completo y revisión por pares.

4.2. UNESCO: visión centrada en lo humano y capacidad institucional

La guía propuesta por UNESCO (2023) pide acciones inmediatas (ética, privacidad, seguridad, equidad, transparencia y alfabetización), políticas a medio plazo y desarrollo docente para sostener una visión centrada en lo humano.

Esta guía enlaza con los escenarios propuestos de la siguiente manera:

- **Escenario 1 (apoyo responsable)**: declaración de uso, verificación y no utilizar datos personales.
- **Escenario 2 (colaboración guiada)**: consolidación de la capacidad docente mediante el uso de rúbricas.
- **Escenario 3 (cocreación con declaración de uso reforzado)**: validación ética y pedagógica previa a usos de alta exposición (por ejemplo, coevaluación asistida).

4.3. AI Act: riesgo proporcional y obligaciones de transparencia

El AI Act (European Parliament & The Council of the European Union, 2024) no es una norma educativa, pero ordena la adopción por niveles de riesgo, diferenciando riesgo mínimo (sin cargas), riesgo limitado (con obligaciones de transparencia, informar sobre si se interactúa con una IA y marcar el contenido sintético) y alto riesgo (gestión de riesgos, gobernanza de datos, documentación técnica, registro y evaluaciones de impacto cuando proceda). Para el contexto educativo, la clave es la proporcionalidad, es decir, en usos de aprendizaje no evaluativos se habla de transparencia reforzada; cuando la IA impacta en decisiones académicas (por ejemplo, calificación), conviene emular controles de "alto riesgo", con revisión humana, trazabilidad completa, análisis de sesgos y documentación.

Esta norma enlaza con los escenarios propuestos de la siguiente manera:

- **Escenario 1 (apoyo responsable)**: tratar como riesgo limitado, lo que implica una declaración de uso consistente, el marcado del contenido generado por IA y la prohibición de usar la salida como evidencia única de calificación.
- **Escenario 2 (colaboración guiada)**: si la IA se usa para dar retroalimentación o definir rutas de aprendizaje, se debe preparar documentación de propósito y riesgos a nivel de curso, así como registros de interacción.
- **Escenario 3 (cocreación con declaración de uso reforzado)**: se deben emplear controles de "alto riesgo": revisión humana antes de efectos académicos, auditoría de sesgo y documentación técnica accesible para acreditación.

4.4. SAFE: del principio a la práctica

El marco SAFE de EDSAFE AI (2021) funciona como puente entre la política pública y el aula. Resume lo que toda implantación debe garantizar: Seguridad (S), Responsabilidad (*Accountability*) (A), Justicia (*Fairness*) (F) y Eficacia (E). Se despliega en recursos prácticos: políticas de uso, bibliotecas de recursos, plantillas de uso aceptable y materiales de desarrollo profesional. Para un departamento, centro o institución, SAFE es la lista de comprobación operativa que complementa el Manifiesto (principios) y el AI Act (cumplimiento).

Este marco enlaza con los escenarios propuestos de la siguiente manera:

- **Escenario 1 (apoyo responsable)**: S = no datos personales; A = declaración de uso y registro; F = comprobaciones rápidas de sesgo; E = dejar evidencia del aprendizaje añadido.
- **Escenario 2 (colaboración guiada)**: más A y E, lo que implica versionado trazable, rúbricas y comparación con/sin IA.
- **Escenario 3 (cocreación con declaración de uso reforzado)**: SAFE completo con auditorías periódicas de sesgo/error y pruebas de eficacia del diseño instruccional asistido por IA.

4.5. Hoja de ruta para el uso seguro de la IAGen en educación

Los cuatro referentes convergen en un mensaje operativo: usar la IAGen sí, pero con agencia humana, riesgo proporcional y transparencia verificable. El Manifiesto aporta el idioma práctico de aula (supervisión, privacidad, precisión, explicabilidad y transparencia); la UNESCO, la visión centrada en las personas y la capacidad institucional; el AI Act, la arquitectura de riesgo y las obligaciones de marcado y documentación; y SAFE, el puente de recursos entre política y clase. Con esta brújula, los tres escenarios del artículo se convierten en decisiones coherentes de diseño, evaluación y compra, con salvaguardas graduadas y métricas de aprendizaje que evitan tanto la improvisación como la parálisis.

Se propone un sistema de capas, con una secuencia mínima (ver Tabla 3), que enlaza los escenarios con los cuatro marcos.

Tabla 3. Capas para una hoja de ruta para el uso seguro de la IAGen en educación.

Capa 1: Transparencia base (para todo uso)	**Capa 2: Gestión proporcional del riesgo (cuando la IA media en aprendizaje/evaluación)**	**Capa 3: Controles reforzados (cuando la IA afecta decisiones sensibles)**
• Declaración de uso (herramienta, fase, límites). • Marcado visible de contenido generado y registro de *prompts*/versiones. • Verificación previa de hechos/citas (prohibidas bibliografías fabricadas). *(Manifiesto: transparencia y precisión; UNESCO: visión centrada en las personas; AI Act Art. 50: transparencia de contenidos; SAFE: responsabilidad).*	• Matriz de riesgo por actividad (impacto en la nota, datos tratados, sesgos previsibles, mitigaciones). • Cribado estructurado y documentado (ligero) en el contexto educativo (mini-EIPD: quién/para qué/qué datos/cómo se supervisa y apela), para asegurarse de que no hace falta una EIPD completa. • Evidencia de eficacia didáctica (qué mejora justifica el uso). *(AI Act: enfoque por riesgo; Manifiesto: supervisión y precisión; SAFE: eficacia y responsabilidad).*	• Revisión humana obligatoria y derecho de apelación del estudiantado. • Auditoría de sesgos y pruebas de rendimiento multilingüe; comunicar limitaciones. • Documentación técnica y trazabilidad completas para evaluación y acreditación. *(Manifiesto: supervisión, privacidad y transparencia de datos; AI Act: obligaciones reforzadas; SAFE: S/F/A).*

Nota sobre la EIPD: En el contexto de la UE, la "mini-EIPD" es una evaluación superficial para decidir si se requiere una evaluación completa del impacto de la protección de datos (EIPD) en virtud del Artículo 35 del RGPD (European Parliament & Council of the European Union, 2016), y para documentar las medidas de mitigación (por ejemplo, supervisión y apelación humanas, minimización de datos, controles de acceso, marcado, verificación).

5. Tres escenarios de uso de IA en educación

A continuación, se definen los tres escenarios de uso, con sus actividades tipo, riesgos, salvaguardas, evidencias mínimas y métricas, conectando explícitamente con el Manifiesto *Safe AI in Education* (Alier et al., 2024), la guía para el uso de la IAGen en la educación y la investigación de la UNESCO (2023), el AI Act (Artículo 50) (European Parliament & The Council of the European Union, 2024) y el marco SAFE (EDSAFE AI, 2021).

5.1. Escenario 1: Apoyo responsable (bajo riesgo, baja autonomía)

Este primer escenario describe un uso instrumental y no evaluativo de la IA. El profesorado usa herramientas de IA como apoyo a su trabajo. La herramienta ayuda a pensar, ordenar y redactar, pero no decide nada sustantivo por sí misma. Es el terreno del diseño instruccional y de la creación de contenidos: generar esquemas de estudio, proponer preguntas de autoevaluación, clarificar conceptos difíciles con explicaciones alternativas o crear contenidos docentes. La clave pedagógica está en que la IA es un apoyo y no un atajo. Sirve para abrir posibilidades, no para cerrar el razonamiento. El *Safe AI in Education Manif*esto organiza bien esta filosofía, ya que establece que la IA complementa al profesorado, nunca lo sustituye; su intervención debe quedar transparente y auditada; y las salidas que se adopten han de ser verificables, no un acto de fe.

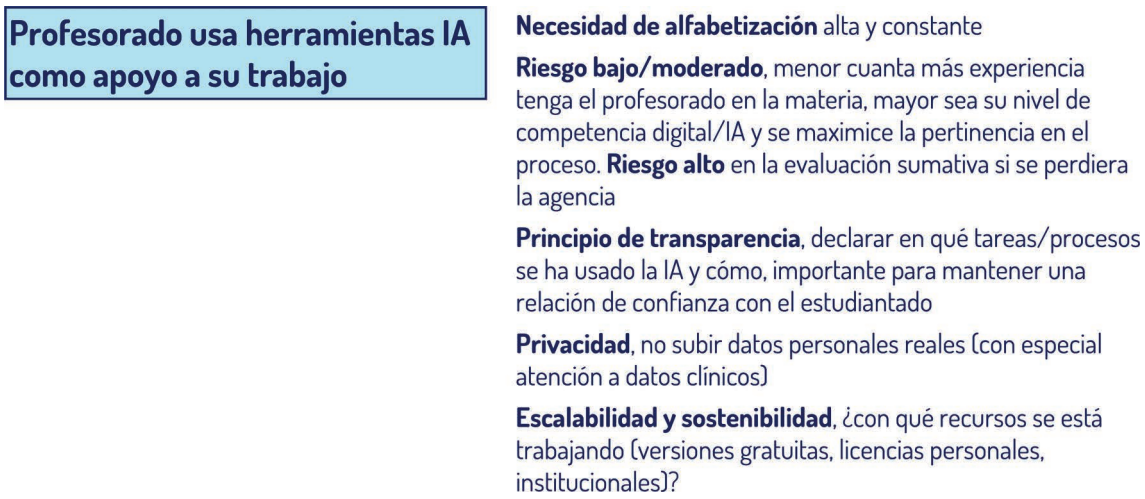
Los riesgos, aunque contenidos, existen. El más habitual es la apariencia de verdad sin veracidad. La IA sintetiza con fluidez, pero puede introducir errores o exagerar conclusiones. También existen los sesgos cognitivos humanos, por ejemplo, la sobreconfianza en una respuesta bien redactada, y descuidos de privacidad cuando se pegan fragmentos con datos personales en servicios externos. Nada de esto exige prohibición, pero sí buenos hábitos. Aquí el marco de la UNESCO resulta pragmático: acciones inmediatas de marcado, alfabetización y verificación, políticas que preserven la agencia humana y un desarrollo docente capaz de explicar límites y buenas prácticas.

Las garantías, por tanto, son proporcionales y sencillas. Primero, transparencia: informar de forma breve de que se ha usado IA, marcar de manera visible el contenido generado y registrar de forma simple el flujo "*prompt* → salida → correcciones". El AI Act lo refuerza con obligaciones de transparencia para sistemas de riesgo limitado: el usuario debe saber cuándo interactúa con IA y cuándo un contenido sintético ha sido generado por IA; ese marcado y aviso son coherentes con el Artículo 50. Segundo, verificación previa: toda afirmación factual y toda referencia deben cotejarse con DOI/ISBN/URL académicos antes de incorporarse. Tercero, minimización de datos: no introducir información personal ni sensible en servicios sin garantías y preferir despliegues institucionales cuando existan. Por último, encaje didáctico, esto es, cada uso debe tener un propósito claro de aprendizaje (mejorar claridad, ampliar ejemplos, reexplicar), no un objetivo de sustitución del trabajo intelectual.

¿Cómo se evidencia que el uso ha sido responsable? En las tareas es suficiente con una declaración de uso breve, el marcado de los pasajes generados y una breve nota de post-edición que indique qué se aceptó y qué se corrigió. A partir de ahí, los indicadores son operativos: tasa de afirmaciones o citas verificadas frente a las que tuvieron que corregirse; y, en tareas de estilo, mejora observable de claridad y coherencia entre el borrador inicial y la versión final. En privacidad, el umbral es de incidentes cero. De nuevo, el Manifiesto aporta el idioma concreto de aula y el marco SAFE el puente entre política y práctica (seguridad y responsabilidad), mientras que UNESCO recuerda que la transparencia y la alfabetización son la base de una adopción centrada en la persona.

Este escenario convierte a la IA en un asistente visible del profesorado que facilita la preparación y comprensión, sin desplazar la autoría ni la verificación, como se resume en la Figura 1. Es el punto de entrada natural para infundir hábitos de marcado, contraste y uso consciente. Sirve, además, como cultura común antes de avanzar a formas de colaboración más intensas. El equilibrio es simple: bajo riesgo, alta transparencia y criterio académico en primer plano.

Figura 1. Características del escenario de apoyo responsable.



5.2. Escenario 2: Colaboración guiada (riesgo moderado, autonomía compartida)

La clave de este escenario es que el profesorado incorpora herramientas de IA en las actividades con sus estudiantes. Aquí la IA coparticipa en el proceso creativo o analítico, pero siempre bajo iteración trazable y con post-edición humana significativa. Un ensayo puede nacer de un diálogo con la herramienta para explorar contraargumentos, pero el estudiantado debe reescribir con sus propias palabras, sustituir referencias sugeridas por citas reales y explicitar cómo cambió su razonamiento entre versiones. En la programación, la herramienta acelera esbozos y pruebas, pero la persona mantiene la propiedad de diseño, añade pruebas unitarias y comenta decisiones. En el análisis de datos, la IA puede proponer rutas exploratorias; el estudiantado ejecuta, valida supuestos, corrige errores y documenta la trazabilidad. El valor educativo reside en hacer visible la toma de decisiones y la mejora entre iteraciones, con la revisión docente como salvaguarda real, en línea con el Manifiesto (supervisión y derecho de apelación) y con los protocolos de transparencia y verificación defendidos por la UNESCO.

El riesgo se encuadra entre bajo y moderado, menor cuanta más experiencia tenga el profesorado en la materia, mayor sea su nivel de competencia digital/IA y se maximice la pertinencia en el proceso. Este riesgo se desplaza del dato puntual erróneo al arrastre de errores entre versiones si falta post-edición crítica. Puede generarse una dependencia al delegar la síntesis o la argumentación. Por otro lado, en flujos con búsquedas externas o basados en *grounding* (Kenthapadi et al., 2024) o sistemas RAG (*Retrieval-Augmented Generation*) (Zhao et al., 2024), pueden aparecer alucinaciones fundamentadas en fuentes erróneas. La solución no es frenar la iteración, sino aplicarla con juicio crítico de forma sistemática. Primero, se puede aplicar una trazabilidad reforzada que incluya un *log* de *prompts* y un versionado claro (v1, v2, v3) con *diffs* y comentarios que expliquen qué se aceptó y qué se descartó. En segundo lugar, una revisión humana significativa antes de publicar una retroalimentación automatizada o cerrar una versión que cuente para evaluación. Tercero, controles de sesgo y lengua, conviene comprobar que el modelo mantiene calidad y no arrastra prejuicios (SAFE-F). Cuarto, si se tratan datos de estudiantes, realizar un cribado ligero de protección de datos (mini-EIPD) y, si procede, escalar a una evaluación completa. Todo ello, sin olvidar los avisos y marcado que exige la transparencia, el estudiantado debe saber qué parte de la retroalimentación o del borrador provino de IA y con qué límites.

Las evidencias de aprendizaje en este escenario se derivan del proceso. Evidencias que incluyan el historial de versiones, la lista de fuentes verificadas con DOI/ISBN/URL, y una nota de post-edición donde se explica qué se cambió y por qué. La rúbrica recogida en la Tabla 1 (veracidad y actualidad, trazabilidad, corrección de alucinaciones, equidad e inclusión y calidad de interacción con IA) permite evaluar de forma homogénea al estudiantado y se puede aplicar en asignaturas diferentes. En métricas conviene observar la mejora por iteración (por ejemplo, la calidad argumental o la robustez del código), la tasa de corrección de alucinaciones, la eficiencia medida en tiempo sin sacrificar validez, la diversidad en la bibliografía final y el cumplimiento del registro de *prompts* y versiones. El resultado esperable no es un producto final mejor sino un pensamiento más justificado y un proceso más trazable, coherente con el Manifiesto, la visión centrada en la persona de la UNESCO, el principio de proporcionalidad del AI Act y la lógica de responsabilidad/eficacia de SAFE.

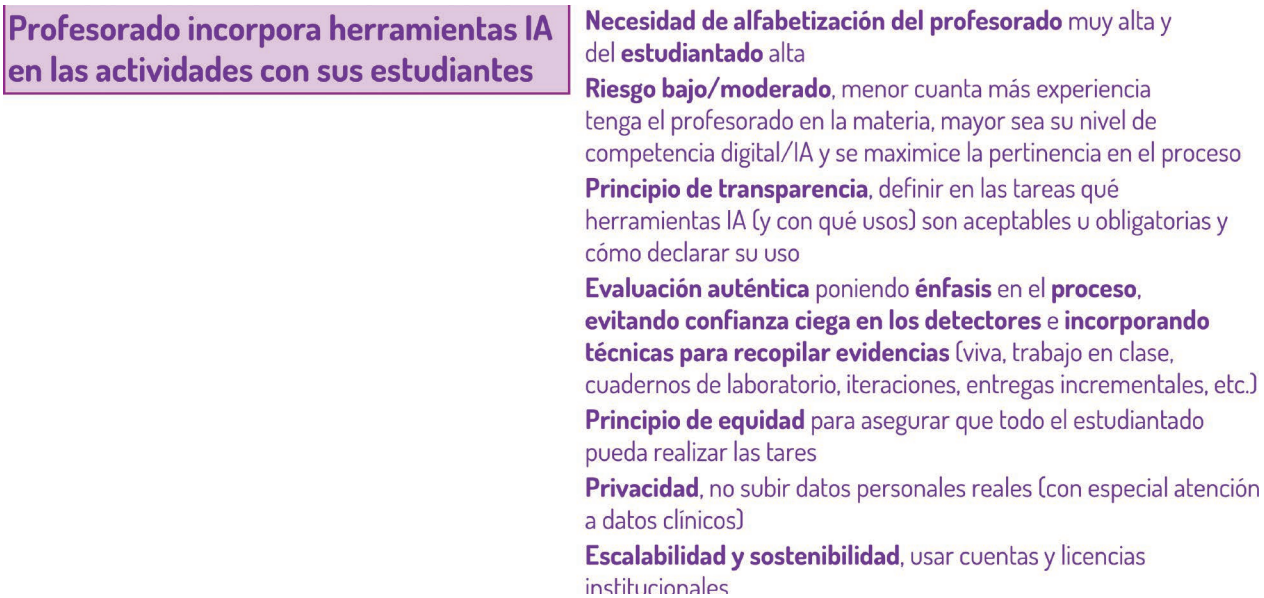
La colaboración guiada convierte a la IA en una pareja de práctica que acelera, provoca, sugiere; la persona discierne, corrige y firma. Se gana velocidad sin perder método y se enseña a pensar con la herramienta, pero manteniendo el control intelectual y ético del proceso. Es el escenario natural para madurar competencias de *prompting* con criterio, post-edición y verificación sistemática, que queda resumido en la Figura 2.

5.3. Escenario 3: Cocreación con declaración de uso reforzada (alto impacto, alta trazabilidad)

Representa el escenario en el que el estudiantado va a usar las herramientas de IA para su propio proceso de aprendizaje de forma autónoma. Se relaciona con diferentes tipos de productos académicos, que van desde entregas en una asignatura a resultados que pueden llegar a tener un alto impacto académico o público porque pueden llegar a depositarse en repositorios institucionales, repositorios de código abierto, exposiciones abiertas, etc. que circularán más allá del aula.

No es que la IA haga más en términos de autonomía, sino que el riesgo de las consecuencias es mayor (menor cuanta más madurez tenga el estudiantado para mantener la agencia de su aprendizaje y mayor sea su competencia digital y en IA); por ello, las salvaguardas deben aproximarse a las de sistemas de alto riesgo en la lógica del AI Act, con revisión humana obligatoria, trazabilidad completa, auditorías de sesgo y documentación técnica accesible para revisión y acreditación. Todo ello sin perder el foco didáctico, la cocreación con IA sigue siendo un proceso de aprendizaje donde se explicitan autorías, se valida la veracidad y se comunica con honestidad qué parte del resultado fue asistido por herramientas de IA.

Figura 2. Características de colaboración guiada.



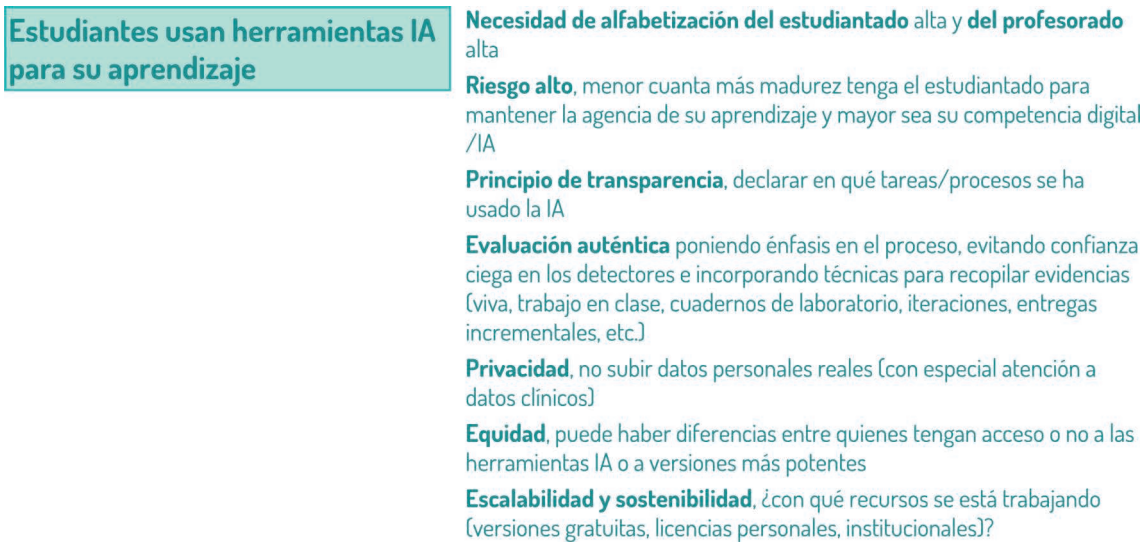
La integridad y la autoría se vuelven cuestiones centrales. Es imprescindible separar lo que es contribución humana de lo que es asistencia automatizada, así como evitar las referencias bibliográficas inventadas y razonamientos que solo parecen sólidos. La equidad también sube de nivel. Si el producto es público, los sesgos de género, raza o región y las brechas lingüísticas al trabajar en idiomas minoritarios, dejan de ser un detalle para convertirse en un requisito de calidad. A esto se suman riesgos de propiedad intelectual y de privacidad cuando se manipulan datos reales o materiales con licencias incompatibles y riesgos técnicos de inyección de *prompts* o de fuentes manipuladas cuando se usa *grounding* o RAG. El Manifiesto y el marco SAFE ofrecen las pautas mediante supervisión humana efectiva y derecho de apelación, privacidad, precisión y explicabilidad, transparencia de interfaz y datos; seguridad, responsabilidad, equidad y eficacia.

De forma operativa, este escenario requiere declaraciones de uso que acompañen al producto: quién hizo qué, traza completa de *prompts* y versiones, fuentes primarias verificadas con DOI/ISBN/URL, informe de verificación que explique qué se contrastó y cómo, chequeo de equidad e idioma, y un acta de revisión humana (y, si procede, revisión por pares). El contenido generado por IA debe estar marcado de manera visible y el documento final incluir una declaración pública de metodología, cumpliendo el principio de transparencia del Artículo 50, de forma que el receptor no tenga duda de si hubo o no intervención de IA. En paralelo, conviene ejecutar una auditoría de sesgos y una prueba multilingüe cuando la lengua de docencia o de difusión no sea el inglés, comunicando limitaciones y mitigaciones. Si se tratan datos personales, el centro debería documentar un cribado de EIPD o, si el riesgo lo aconseja, una evaluación completa.

Las métricas deben ser más exigentes, por ejemplo, porcentaje de afirmaciones con fuente verificable y reproducibilidad del resultado; incidencias detectadas y corregidas en revisión humana y por pares; indicadores de equidad (diversidad de voces/regiones en referencias y casos); presencia y detectabilidad del marcado de IA; y cumplimiento de privacidad y licencias sin excepciones. La eficacia se demuestra con impacto formativo (dominio del proceso, mejor argumentación, mayor calidad técnica, transferibilidad del producto). En términos de gobernanza, el triángulo Manifiesto–UNESCO–AI Act ofrece la guía: agencia humana y derecho de apelación, visión centrada en la persona y controles proporcionados al riesgo.

Este escenario se resume en la Figura 3, toma como centro la cocreación con declaración de uso reforzada precisamente para que no sea más de lo mismo, pero con IA, sino que se incremente la responsabilidad y se asegure la trazabilidad por el impacto del resultado. Es decir, este escenario exige una cultura de evidencias y auditoría que conecta plenamente con los estándares internacionales y la regulación vigente, que prepara al estudiantado para enfrentarse con garantías a la era de la IAGen en su futura vida profesional.

Figura 3. Características de la cocreación con declaración de uso reforzada.



6. Del mapa a la práctica: fases del ciclo docente

Tras la presentación de los tres escenarios, se enlazan con las cuatro fases principales del ciclo docente: i) planificación y diseño instruccional; ii) creación de materiales (que puede subdividirse en la búsqueda y síntesis de evidencia, la creación de los recursos educativos y actividades relacionadas con la gestión de datos y programación); iii) apoyo al aprendizaje; y iv) evaluación. Estas fases deben estar bajo el paraguas de una capa de ética y transparencia que ofrezca las garantías propias de los marcos presentados en la Sección 4. Esta estructura se refleja en el esquema de la Figura 4, en la que las fases se completan con un catálogo de herramientas que, actualmente, se podrían utilizar en dichas fases.

Figura 4. Fases del ciclo docente.

6.1. Planificación y diseño instruccional

Esta fase es el hábitat natural del Escenario 1 (apoyo responsable) y del Escenario 2 (colaboración guiada). En la práctica, la IAGen puede apoyar en i) la descripción de los resultados de aprendizaje, el diseño de las actividades, la propuesta de evidencias y criterios de evaluación, etc.; ii) el análisis de riesgos y requisitos, qué datos se manejarán, qué sesgos son previsibles, qué idioma(s) se usarán; iii) la selección de herramientas y de su encaje institucional; entre otras. Desde la perspectiva de garantías, UNESCO pide acciones inmediatas (alfabetización, privacidad, equidad, transparencia) y capacidad institucional (políticas y desarrollo docente); SAFE traduce esto en listas de comprobación para seguridad, responsabilidad, justicia y eficacia; y el Manifiesto recuerda que toda integración debe explicitar supervisión humana, privacidad y transparencia.

6.2. Creación de materiales

Esta fase es propia del Escenario 1 (materiales que el docente revisa y valida) y del Escenario 2 (cuando se trabaja con estudiantes en codiseño). La IAGen puede ayudar a encontrar y sintetizar evidencia académica y preparar actividades específicas relacionadas con la programación y el análisis de datos. Ahora bien, su mayor potencial reside en que la IAGen se puede convertir en un asistente para la creación de contenidos de diversa índole, enriqueciendo los materiales con elementos complementarios (ejemplos, contraejemplos, ejercicios, etc.), visuales, interactivos, personalizados (por ejemplo, explicaciones alternativas para diferentes niveles), etc.

En la creación de contenido hay dos reglas operativas básicas. Primero, la declaración de uso cuando hay parte de los materiales que se han generado con IA, siendo recomendable que se tenga acceso a la traza "*prompt* →salida→correcciones". Esta transparencia es coherente con las obligaciones del Artículo 50 del AI Act al tratar contenidos sintéticos (texto, imagen, audio o vídeo). Segundo, verificación de los hechos, las cifras y las citas, así como atención a las posibles brechas de idioma y voces subrepresentadas para evitar sesgos curriculares. Al seguir el Manifiesto (precisión, explicabilidad y transparencia) y SAFE (S/A/F/E), el contenido generado por IA se convierte en un insumo revisable, no en un sustituto de la autoría docente.

6.3. Apoyo al aprendizaje

El Escenario 2 (colaboración guiada) se desarrolla de forma natural en esta fase, por más que en un contexto propio del Escenario 1 (apoyo responsable) se pueda generar contenido relacionado con procesos de tutoría, guía, etc.

El proceso se convierte en el centro para el control de la evidencia en esta fase. UNESCO enfatiza protocolos de transparencia y verificación y desarrollo docente; SAFE refuerza la responsabilidad; y el Manifiesto da las pautas para la interacción docente, la supervisión y el derecho de apelación.

Las evidencias del proceso (historial de versiones, lista de fuentes verificadas, nota de post-edición, etc.) se valoran con la rúbrica de la Tabla 1 y las listas de la Tabla 2.

6.4. Evaluación

La evaluación es el punto más sensible y, por ello, donde más claramente se gradúan los tres escenarios. La evaluación formativa (por ejemplo, en la retroalimentación de entregas pautadas o ejercicios) puede aparecer en el Escenario 2, con marcado de la intervención de IA, trazabilidad y revisión humana antes de consolidar la respuesta.

Cuando la evaluación es sumativa con productos de alto impacto o decisiones sustantivas (por ejemplo, trabajos finales públicos, prototipos con terceros, exámenes), se deben aplicar los controles reforzados propios del Escenario 3, con independencia de si la IAGen la usa el profesorado o el estudiantado, incluyendo revisión humana obligatoria, declaración de uso, *prompts* y versiones, etc., así como una declaración pública conforme al Artículo 50 del AI Act para que el receptor no tenga duda de si hubo o no intervención de IA y en qué términos. Los factores decisivos son el impacto y la trazabilidad, no quién utiliza la herramienta.

Si se permite el uso de la IAGen por parte del estudiantado en una asignatura, en la evaluación sumativa puede aplicarse la rúbrica de la Tabla 1 y las listas de verificación de la Tabla 2 para valorar veracidad, actualidad, trazabilidad, corrección de alucinaciones, equidad, aspectos de idioma y calidad de la interacción con la IA. La orientación sectorial (Joint Council for Qualifications, 2025; Office of Qualifications and Examinations Regulation, 2024, 2025; Walker, 2025) converge en reforzar la integridad académica, elevar la autenticidad de las tareas y potenciar defensas orales (Wang et al., 2024) u otras evidencias de agencia y propiedad del proceso, desaconsejando depender de detectores como estrategia principal.

Además, la evidencia muestra que los patrones de mala praxis evolucionan y que detectar IA es sustancialmente distinto de detectar plagio (Sadasivan et al., 2025; Weber-Wulff et al., 2023); los estudios comparativos sobre detectores reportan limitaciones, lo que refuerza la necesidad de diseños que verifiquen agencia y trazabilidad en lugar de apoyarse en herramientas de detección, que, en todo caso, pueden servir como alerta inicial en algunas situaciones, pero nunca reemplazan el criterio humano.

6.5. Cerrando el círculo

En las cuatro fases que representan una abstracción del ciclo docente, la coherencia se logra si cada decisión responde a tres preguntas: 1) propósito pedagógico (qué aporta la IA al aprendizaje); 2) riesgo y salvaguardas (qué exposición o impacto tiene y qué controles actúan); y 3) evidencia y métricas (qué deja rastro, qué se puede auditar, cómo mejora).

La regla práctica es de proporcionalidad, es decir, cuanto mayor sea el impacto (o la sensibilidad de los datos), más cerca de los controles establecidos en el Escenario 3; cuanto más exploratorio y no evaluativo, más cerca se está del régimen ligero propio del Escenario 1. Esta graduación de garantías materializa la arquitectura de riesgo del AI Act (transparencia para usos limitados; controles reforzados cuando hay decisiones de alto impacto) y evita tanto la prohibición indiscriminada como la adopción acrítica.

Este mapa no añade burocracia innecesaria, sino que convierte en rutina visible lo que define la cultura académica (citar, verificar y dar cuenta) y lo alinea con el Manifiesto (supervisión, privacidad y transparencia), las directrices de la UNESCO (visión centrada en las personas y capacidad institucional), las obligaciones del AI Act (marcado y aviso del contenido sintético) y el puente operativo de SAFE (S/A/F/E con recursos y plantillas).

Cerrando el círculo, el valor de este mapeo no está en la taxonomía, sino en transmitir los hábitos de buen uso. El Escenario 1 presenta la transparencia y la verificación, el Escenario 2 entrena el criterio y la iteración, mientras que el Escenario 3 exige evidencia y auditoría. Disponer de fases y escenarios acoplados permite a cada docente saber qué hacer, qué pedir, qué revisar y cómo justificar la adopción de IAGen con estándares internacionales y con la normativa vigente, mientras mantiene intacta la autoría humana, la equidad y la calidad del aprendizaje.

7. Discusión y conclusiones

Este artículo ha propuesto una forma pragmática y proporcionada de integrar la IAGen en la educación superior mediante tres escenarios graduados por autonomía, agencia y riesgo, ligados a los hábitos profesionales (Sección 3) y marcos normativo-éticos (Sección 4), y llevados a la práctica en el ciclo docente (Sección 6). El hilo conductor es claro. La IA como complemento, siempre supervisada por el juicio académico, no como sustituto; transparencia y trazabilidad como premisas; equidad e inclusión por diseño; y proporcionalidad del riesgo para decidir salvaguardas. Esta propuesta dialoga estrechamente con el *Safe AI in Education Manifesto* (supervisión humana, privacidad, precisión, explicabilidad y transparencia), las orientaciones de la UNESCO (visión centrada en las personas, acciones inmediatas y capacidad institucional), el AI Act (transparencia y marcado de contenido sintético; controles reforzados según exposición) y el marco SAFE como puente operativo.

El valor de los tres escenarios radica en convertir principios amplios en decisiones docentes verificables. En el Escenario 1 (apoyo responsable), la regla es de bajo riesgo y alta transparencia: declaración de uso breve, marcado de pasajes generados, verificación factual y cero datos personales en servicios sin garantías. En el Escenario 2 (colaboración guiada), la clave es la iteración trazable más la post-edición significativa: versiones con cambios registrados (*diffs*), lista de fuentes verificadas y nota de decisiones. En el Escenario 3 (cocreación con declaración de uso reforzada), el foco pasa a evidencias robustas y auditoría (*prompts* y versiones, chequeo de sesgos e idioma, revisión humana y, si procede, revisión por pares), coherentes con el Artículo 50 del AI Act y

con los principios de SAFE. Este gradiente de garantías materializa el enfoque por riesgo del AI Act sin trasladar burocracia innecesaria al aula.

En términos de roles, los escenarios se ajustan mayoritariamente al uso docente de la IAGen como asistente (Escenario 1), a su incorporación en actividades con estudiantes (Escenario 2) y al uso autónomo por parte del estudiantado (Escenario 3). No obstante, en evaluación sumativa de alto impacto pueden exigirse controles propios del Escenario 3 aunque sea el profesorado quien use la IA, porque lo determinante es el riesgo y el impacto del resultado, no quién use la herramienta. Esto refuerza la proporcionalidad del AI Act y el énfasis de la UNESCO en salvaguardas institucionales.

Precisamente en el terreno evaluativo, las orientaciones sectoriales en el ámbito británico señalan este camino. El *Jisc* recomienda reforzar autenticidad, agencia y propiedad del proceso en el rediseño de la evaluación; la *Office of Qualifications and Examinations Regulation* (Ofqual) enmarca el uso de IA en el sector de cualificaciones subrayando calidad y equidad; y *Joint Council for Qualifications* (JCQ) actualiza reglas para centros sobre divulgación del uso, integridad y mala praxis, desaconsejando depender de detectores como estrategia principal. Estas guías refuerzan la premisa de este trabajo de que la evaluación debe comprobar agencia y trazabilidad más que centrarse en la detección del uso de la IA, así como el valor de las defensas orales u otras evidencias performativas como instrumentos coherentes con este fin.

En términos de viabilidad institucional, la propuesta se apoya en dos palancas que facilitan la adopción. Por un lado, una rúbrica transversal (veracidad y actualidad; trazabilidad; corrección de alucinaciones; equidad e idioma; calidad de la interacción) que permite una evaluación homogénea entre asignaturas. Por otro lado, listas de verificación por tipo de tarea (resumen, ensayo, problema cuantitativo, código, imagen) que convierten la transparencia y la verificación en rutinas. Este mínimo de garantías (declaración de uso, marcado, registro básico y verificación) es plenamente consistente con el Manifiesto y con la UNESCO y escalable hacia controles reforzados en función del impacto (AI Act).

El ecosistema de modelos y políticas evoluciona con rapidez; la operativización concreta (por ejemplo, qué campo exacto incluye una declaración de uso o cómo almacenar *logs* con privacidad) puede variar entre instituciones y jurisdicciones. Además, la medición de impacto (aprendizaje, equidad, carga de trabajo) requerirá estudios cuasiexperimentales y diseños longitudinales. Por último, las brechas lingüísticas y de infraestructura exigen una inversión sostenida y una evaluación multilingüe de herramientas para evitar trasladar desigualdades al aula. Las hojas de ruta institucionales deberán revisarse periódicamente, en línea con el énfasis de la UNESCO en la capacidad institucional y en el desarrollo docente, lo que magnifica el valor de la gobernanza de la IA en las universidades (Molina-Carmona & García-Peñalvo, 2025).

La llegada de la IAGen no obliga a elegir entre entusiasmo y prohibición; obliga a enseñar y aprender con garantías. Los tres escenarios propuestos permiten a las universidades y al profesorado decidir con proporción, evaluar con evidencia y rendir cuentas con transparencia, manteniendo en todo momento la agencia humana y la equidad como criterios rectores. En la práctica, marcar, verificar y documentar no son trámites, sino las formas contemporáneas de cuidar la cultura académica. Si el Manifiesto aporta el idioma de aula, la UNESCO la visión centrada en las personas, el AI Act la arquitectura de riesgo y SAFE el puente operativo, entonces la universidad ya dispone del mapa para convertir la IA en oportunidad pedagógica sin ceder en rigor, justicia y responsabilidad. Esta es, en última instancia, la referencia que debe guiar el proceso de enseñanza-aprendizaje en la era de la IA.

Referencias / References

Abegglen, S., Nerantzi, C., Martínez-Arboleda, A., Karatsiori, M., Atenas, J., & Rowell, C. (Eds.). (2024). *Towards AI Literacy: 101+ Creative and Critical Practices, Perspectives and Purposes*. Zenodo. https://doi.org/10.5281/zenodo.11613520.

Afreen, J., Mohaghegh, M., & Doborjeh, M. (2025). Systematic literature review on bias mitigation in generative AI. *AI and Ethics, 5*(5), 4789–4841. https://doi.org/10.1007/s43681-025-00721-9

Alier, M., García-Peñalvo, F. J., Casañ, M. J., Pereira, J. A., & Llorens-Largo, F. (2024). *Safe AI in Education Manifesto. Version 0.4.0*. https://manifesto.safeaieducation.org

Alier-Forment, M., Casañ-Guerrero, M. J., Pereira, J., García-Peñalvo, F. J., & Llorens-Largo, F. (2026). Inteligencia artificial generativa y autonomía educativa: metáforas históricas y principios éticos para la transformación pedagógica. *RIED: revista iberoamericana de educación a distancia, 29*(1). https://doi.org/10.5944/ried.29.1.45536

An, J., Huang, D., Lin, C., & Tai, M. (2025). Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus, 4*(3), Article pgaf089. https://doi.org/10.1093/pnasnexus/pgaf089

Anthropic. (2025, September 29). Introducing Claude Sonnet 4.5. *Anthropic*. https://d66z.short.gy/Gk55eS

Artopoulos, A., & Lliteras, A. (2024). Alfabetización crítica en IA: Recursos educativos para una pedagogía de la descajanegrización. *Trayectorias Universitarias, 10*, Article e168. https://doi.org/10.24215/24690090e168

Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. *Brain-X, 1*(3), Article e30. https://doi.org/10.1002/brx2.30

Bedington, A., Halcomb, E. F., McKee, H. A., Sargent, T., & Smith, A. (2024). Writing with generative AI and human-machine teaming: Insights and recommendations from faculty and students. *Computers and Composition, 71*, Article 102833. https://doi.org/10.1016/j.compcom.2024.102833

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada, March 3 - 10, 2021)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bittle, K., & El-Gayar, O. (2025). Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. *Information, 16*(4), Article 296. https://doi.org/10.3390/info16040296

Boonstra, L. (2025). *Prompt Engineering*. Google. https://d66z.short.gy/3ok7tY

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv*, Article arXiv:2005.14165v4 https://doi.org/10.48550/arXiv.2005.14165

Burneo-Arteaga, P., Lira, Y., Murzi, H., Balula, A., & Costa, A. P. (2025). Capability-based training framework for generative AI in higher education. *Frontiers in Education, 10*, Article 1594199. https://doi.org/10.3389/feduc.2025.1594199

Castañeda, L., & Selwyn, N. (2018). More than tools? Making sense of the ongoing digitizations of higher education. *International Journal of Educational Technology in Higher Education, 15*(1), 22. https://doi.org/10.1186/s41239-018-0109-y

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How people use ChatGPT* (34255). (NBER Workking Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w34255

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research, 26*, Article e53164. https://doi.org/10.2196/53164

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA* (Vol. 30). Curran Associates, Inc.

Clarke, A. C. (1973). *Profiles of the Future: An Inquiry into the Limits of the Possible* (2nd ed.). Harper & Row.

DeepSeek. (2025, September 29). Introducing DeepSeek-V3.2-Exp. *DeepSeek API Docs*. https://d66z.short.gy/eXidah

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence, 2*(8), 423–425. https://doi.org/10.1038/s42256-020-0219-9

Dúo-Terrón, P. (2024). Generative artificial intelligence: Educational reflections from an analysis of scientific production. *Journal of Technology and Science Education, 14*(3), 756–769. https://doi.org/10.3926/jotse.2680

EDSAFE AI. (2021). *What is the EDSAFE AI SAFE Framework?* EDSAFE AI. https://d66z.short.gy/RNVmzh

European Parliament, & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Brussels, Belgium: European Commission Retrieved from https://bit.ly/2O2juE9

European Parliament, & The Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence

and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). (Official Journal of the European Union). European Union Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Frau-Meigs, D. (2024). *User empowerment through media and information literacy responses to the evolution of generative artificial intelligence (GAI)* (CI/FMD/MIL/2024/3). UNESO. https://d66z.short.gy/Wg2YCU

Fulsher, A., Pagkratidou, M., & Kendeou, P. (2025). GenAI and misinformation in education: a systematic scoping review of opportunities and challenges. *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02536-y

García-Peñalvo, F. J. (2023). The perception of Artificial Intelligence in educational contexts after the launch of ChatGPT: Disruption or Panic? *Education in the Knowledge Society, 24*, Article e31279. https://doi.org/10.14201/eks.31279

García-Peñalvo, F. J. (2024a). Generative Artificial Intelligence and Education: An Analysis from Multiple Perspectives. *Education in the Knowledge Society, 25*, Article e31942. https://doi.org/10.14201/eks.31942

García-Peñalvo, F. J. (2024b). Mito de la inteligencia. Más allá de una educación de silicio. In C. Suárez-Guerrero, J. E. Raffaghelli, & P. Rivera-Vargas (Eds.), *Mitos EdTech. Desmontando el solucionismo tecnológico en educación* (pp. 79–87). Editorial UOC.

García-Peñalvo, F. J., Alier, M., Pereira, J. A., & Casañ, M. J. (2024). Safe, Transparent, and Ethical Artificial Intelligence: Keys to Quality Sustainable Education (SDG4). *IJERI – International Journal of Educational Research and Innovation*(22), 1–21. https://doi.org/10.46661/ijeri.11036

García-Peñalvo, F. J., Casañ-Guerrero, M. J., Alier-Forment, M., & Pereira-Valera, J. A. (2025). The ethics of generative artificial intelligence in education under debate. A perspective from the development of a theoretical-practical case study. *Revista Española de Pedagogía, 83*(291), 281–293. https://doi.org/10.22550/2174-0909.4577

García-Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). The new reality of education in the face of advances in generative artificial intelligence. *RIED: revista iberoamericana de educación a distancia, 27*(1), 9–39. https://doi.org/10.5944/ried.27.1.37716

García-Peñalvo, F. J., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(4), 7–16. https://doi.org/10.9781/ijimai.2023.07.006

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

Gibney, E. (2025). Can researchers stop AI making up citations? *Nature, 645*, 569–570. https://doi.org/10.1038/d41586-025-02853-8

Glynn, A. (2025). Guarding against artificial intelligence-hallucinated citations: the case for full-text reference deposit. *European Science Editing, 51*, Article e153973. https://doi.org/10.3897/ese.2025.e153973

Google. (2025). *Google Environmental Report 2025*. Google. https://d66z.short.gy/uxN9Eu

Hayes, J., Swanberg, M., Chaudhari, H., Yona, I., Shumailov, I., Nasr, M., Choquette-Choo, C. A., Lee, K., & Cooper, A. F. (2025). Measuring memorization in language models via probabilistic extraction. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Albuquerque, New Mexico, April 29 - May 4, 2025)* (pp. 9266–9291). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.469

Huang, J., & Chang, K. (2024). Citation: A Key to Building Responsible and Accountable Large Language Models (Mexico City, Mexico, June 16–21, 2024). In K. Duh, H. Gomez, & S. Bethard (Eds.), *Mexico City, Mexico* (pp. 464–473). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.31

Jegham, N., Abdelatti, M., Elmoubarki, L., & Hendawi, A. (2025). How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference. *arXiv*, Article arXiv:2505.09598v4. https://doi.org/10.48550/arXiv.2505.09598

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Jin, Y., Yan, L., Echeverria, V., Gašević, D., & Martinez-Maldonado, R. (2025). Generative AI in higher education: A global perspective of institutional adoption policies and guidelines. *Computers and Education: Artificial Intelligence, 8*. https://doi.org/10.1016/j.caeai.2024.100348

Joint Council for Qualifications. (2025). *AI use in assessments: Your role in protecting the integrity of qualifications* (Revision two). Joint Council for Qualifications. https://d66z.short.gy/G2eDjK

Jovanović, M., & Campbell, M. (2022). Generative Artificial Intelligence: Trends and Prospects. *Computer, 55*(10), 107–112. https://doi.org/10.1109/MC.2022.3192720

Kassorla, M., Georgieva, M., & Papini, A. (2024). *AI Literacy in Teaching and Learning: A Durable Framework for Higher Education*. Educause. https://d66z.short.gy/bPhL3A

Kenthapadi, K., Sameki, M., & Taly, A. (2024). Grounding and Evaluation for Large Language Models: Practical Challenges and Lessons Learned (Survey). In *KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Barcelona, Spain, August 25 - 29, 2024)* (pp. 6523–6533). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671467

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6*, Article 100225. https://doi.org/10.1016/j.caeai.2024.100225

Kotha, A., Lee, J., & Zakariasson, E. (2025, August 7). GPT-5 prompting guide. *OpenAI Cookbook*. https://d66z.short.gy/CaAOnG

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence, 6*, Article 100221. https://doi.org/10.1016/j.caeai.2024.100221

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: a systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education, 22*(1), Article 7. https://doi.org/10.1186/s41239-025-00503-7

Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI Less 'Thirsty'. *Communications of the ACM, 68*(7), 54–61. https://doi.org/10.1145/3724499

Liu, X., Sun, T., Xu, T., Wu, F., Wang, C., Wang, X., & Gao, J. (2024). SHIELD: Evaluation and Defense Strategies for Copyright Compliance in LLM Text Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Miami, Florida, USA, November 12-16, 2024)* (pp. 1640–1670). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.98

Molina-Carmona, R., & García-Peñalvo, F. J. (2025). Safeguarding Knowledge: Ethical Artificial Intelligence Governance in the University Digital Transformation. In E. Vendrell Vidal, U. R. Cukierman, & M. E. Auer (Eds.), *Advanced Technologies and the University of the Future* (pp. 201–220). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71530-3_14

Mueller, F. B., Görge, R., Bernzen, A. K., Pirk, J. C., & Poretschkin, M. (2024). LLMs and Memorization: On Quality and Specificity of Copyright Compliance. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7*(1), 984–996. https://doi.org/10.1609/aies.v7i1.31697

Nam, B. H., & Bai, Q. (2023). ChatGPT and its ethical implications for STEM research and higher education: a media discourse analysis. *International Journal of STEM Education, 10*(1), Article 66. https://doi.org/10.1186/s40594-023-00452-5

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. *ACM Transactions on Intelligent Systems and Technology, 16*(5), Article 106. https://doi.org/10.1145/3744746

Nerantzi, C., Abegglen, S., Karatsiori, M., & Martínez-Arboleda, A. (Eds.). (2023). *101 creative ideas to use AI in education, A crowdsourced collection*. Zenodo. https://doi.org/10.5281/zenodo.8355454

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education, 49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593

Office of Qualifications and Examinations Regulation. (2024, 24 April). *Ofqual's approach to regulating the use of artificial intelligence in the qualifications sector*. Office of Qualifications and Examinations Regulation. https://d66z.short.gy/WLoJbW

Office of Qualifications and Examinations Regulation. (2025, 1 May). *Ofqual strategy 2025 to 2028* Office of Qualifications and Examinations Regulation. https://d66z.short.gy/8T7iPk

OpenAI. (2025, 7 de agosto). Presentamos GPT-5. *OpenAI*. https://d66z.short.gy/hJeA79

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems (New Orleans, LA, USA, 28 November - 9 December 2022)* (pp. 27730 – 27744). Curran Associates Inc.

Perković, G., Drobnjak, A., & Botički, I. (2024). Hallucinations in LLMs: Understanding and Addressing Challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO) (Opatija, Croatia, 20-24 May 2024)* (pp. 2084–2088). IEEE. https://doi.org/10.1109/MIPRO60963.2024.10569238

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science, 12*, Article 241776. https://doi.org/10.1098/rsos.241776

Qiao, H., Bhardwaj, E., Landau, V. G. D., Bonfils, N., Iqbal, M., Jaworsky, O., Munson, R. O. A., Rubisova, L., Smith, N. M., Thapa, A., & Becker, C. (2025). Are You Thirsty? So is Your AI. In *COMPASS '25: Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (Toronto, Canada, July 22 - 25, 2025)* (pp. 811–816). Association for Computing Machinery. https://doi.org/10.1145/3715335.3736308

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

Romeo, G., & Conti, D. (2025). Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02422-7

Roxas, R. E. (2024). Large Language Models and Natural Language Processing On Minority Languages: A Systematic Review (Tokyo, Japan, 7-9 December 2024). In N. Oco, S. N. Dita, A. M. Borlongan, & J.-B. Kim (Eds.), *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation* (pp. 1–8). Institute for the Stufy of Language and Information (ISLI).

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2025). Can AI-Generated Text be Reliably Detected? *arXiv*, Article arXiv:2303.11156v4. https://doi.org/10.48550/arXiv.2303.11156

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. https://doi.org/10.1145/3381831

Shao, A. (2025). New sources of inaccuracy? A conceptual framework for studying AI hallucinations. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-182

Sozon, M., Parnther, C., Wei Lun, W., & Chowdhury, M. A. (2025). Generative AI in higher education: navigating benefits and challenges in the technological era. *Journal of Applied Research in Higher Education*. https://doi.org/10.1108/JARHE-02-2025-0103

Torres, N., Ulloa, C., Araya, I., Ayala, M., & Jara, S. (2025). A comprehensive analysis of gender, racial, and prompt-induced biases in large language models. *International Journal of Data Science and Analytics, 20*(4), 3797–3834. https://doi.org/10.1007/s41060-024-00696-6

Towhidul Islam Tonmoy, S. M., Mehedi Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv*, Article arXiv:2401.01313v3. https://doi.org/10.48550/arXiv.2401.01313

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO. https://d66z.short.gy/SBxqSb

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (pp. 5998–6008).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. *arXiv*, Article arXiv:1706.03762v7. https://doi.org/10.48550/arXiv.1706.03762

Veldhuis, A., Lo, P. Y., Kenny, S., & Antle, A. N. (2025). Critical Artificial Intelligence literacy: A scoping review and framework synthesis. *International Journal of Child-Computer Interaction, 43*, Article 100708. https://doi.org/10.1016/j.ijcci.2024.100708

Vivas Urias, M. D., & Ruiz Rosillo, M. A. (Eds.). (2025). *Inteligencia artificial generativa. Buenas prácticas docentes en educación superior*. Octaedro.

Walker, S. (2025). *Trends in assessment in higher education: considerations for policy and practice*. Jisc. https://d66z.short.gy/ZMRzML

Wang, C., Fogle, E., & Urban, A. (2024). AI-powered viva exams: advancing academic integrity in online education. In *Proceedings of the 17th annual International Conference of Education, Research and Innovation - ICERI 2024 (Seville, Spain, 11-13 November 2024)* (pp. 5673–5678). IATED. https://doi.org/10.21125/iceri.2024.1379

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), Article 26. https://doi.org/10.1007/s40979-023-00146-z

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned Language Models Are Zero-Shot Learners. *arXiv*, Article arXiv:2109.01652v5. https://doi.org/10.48550/arXiv.2109.01652

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from Language Models. *arXiv*, Article arXiv:2112.04359v1. https://doi.org/10.48550/arXiv.2112.04359

Xu, Y., Hu, L., Zhao, J., Qiu, Z., Xu, K., Ye, Y., & Gu, H. (2025). A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science, 19*(11), Article 1911362. https://doi.org/10.1007/s11704-024-40579-4

Yang, Y., Zhang, Y., Sun, D., He, W., & Wei, Y. (2025). Navigating the landscape of AI literacy education: insights from a decade of research (2014–2024). *Humanities and Social Sciences Communications, 12*(1), Article 374. https://doi.org/10.1057/s41599-025-04583-8

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments, 11*(1), Article 28. https://doi.org/10.1186/s40561-024-00316-7

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. *arXiv*, Article arXiv:2402.19473v6. https://doi.org/10.48550/arXiv.2402.19473

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2025). A Survey of Large Language Models. *arXiv*, Article arXiv:2303.18223v16. https://doi.org/10.48550/arXiv.2303.18223