



## Education in the Knowledge Society

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

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

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

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](mailto:fgarcia@usal.es)

Director Científico / Editor-In-Chief Education in the Knowledge Society Journal

## ARTICLE INFO

## 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 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.

## RESUMEN

## 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.

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.

## 1. Introduction

As the third anniversary of the public release of ChatGPT (30 November 2022) approaches, the university system is confronting the accelerated maturation of generative artificial intelligence (GenAI) (Jovanović & Campbell, 2022): more capable models, deeper integration into authoring and programming tools, and a social and academic use that is no longer exceptional but everyday (Chatterji et al., 2025). This inflexion point invites us to review what has been learned and, above all, to propose operational frameworks that enable a shift from reaction to pedagogical governance.

In 2023, the first editorial in this series took the pulse of academia in real time: the arrival of ChatGPT precipitated a polarised debate between the promise of disruption and the panic over real risks (plagiarism, opacity, bias, displacement of competences) (García-Peñalvo, 2023). That piece argued that diagnosis must go beyond enthusiasm or alarm and called for the conversation to be framed in terms of critical literacy, traceability and human responsibility in the use of GenAI. This interpretative framework, mindful of the novelty yet demanding of evidence and ethics, enabled the clear identification of the phenomenon's nature and its limitations.

In 2024, the second editorial adopted a multiperspective view, analysing GenAI through the roles that coexist in higher education: academic staff, students, academic leaders and development/support teams (García-Peñalvo, 2024a). This actor map highlighted that decisions are neither homogeneous nor symmetrical: objectives, risk exposure, verification capacities and transparency obligations differ. Building on this cartography, it was emphasised that institutional policies should be aligned with classroom practices and with instructional design that makes explicit the agencies (human and tool), the citation of sources and the recording of artificial intelligence (AI) contributions.

This third contribution, produced in the final quarter of 2025, also arrives with a more clearly defined regulatory context and set of international orientations. The European Union's Artificial Intelligence Act 2024/1689 (AI Act) (European Parliament & The Council of the European Union, 2024) introduces a risk-based approach that directly affects obligations of documentation, transparency, and oversight. However, its focus is not educational per se; its spirit permeates university policies and the requirements for procurement and technological deployment. In parallel, UNESCO's Guidance for GenAI in education and research (UNESCO, 2023) calls for a human-centred vision, institutional capacities and staff development, as well as immediate measures to mitigate risks and close preparedness gaps. This regulatory and programmatic convergence offers a firmer basis for moving from improvisation to responsible implementation.

On this basis, an evolution is proposed: to organise the educational use of GenAI into three scenarios, guided by tool autonomy, human agency and risk exposure, which can function as a common language across roles and as a practical guide for deciding which use is reasonable, how to make it verifiable, and with which safeguards. The aim is not to add yet another taxonomy, but to facilitate adoption by linking each scenario to examples of activities, disclosure criteria, traceability mechanisms (from prompt logs to verifiable citations), and metrics that enable the evaluation of learning, quality, and opportunity cost.

Suppose the first paper (García-Peñalvo, 2023) focused on naming the phenomenon and the second (García-Peñalvo, 2024a) on situating it in its roles. In that case, this third paper advances towards an organisation by exploring scenarios that help inform decisions and actions. The objective of the article is therefore twofold: to define these three scenarios (from responsible support to co-creation) and to provide a roadmap that enables academic staff (and educational institutions) to translate principles and regulations into practices that are verifiable, assessable and sustainable over time.

## 2. Critical AI literacy in higher education

### 2.1. What GenAI is—and is not

GenAI refers to AI models capable of producing novel synthetic content in any form and supporting any task through generative modelling (García-Peñalvo & Vázquez-Ingelmo, 2023). To achieve the natural-language performance (noting that generation is now multimodal) that has propelled GenAI to prominence, these models must be huge (billions of parameters) hence the term Large Language Models (LLMs) (Zhao et al., 2025), understood as state-of-the-art AI systems that can process and generate text with coherent communication and generalise across multiple tasks (Naveed et al., 2025). Formally, parameters are the weights of the layers in neural networks; that is, intrinsic elements of the model that are adjusted to optimise performance on the task of predicting the next word or sequence of text based on prior context. The number of parameters can range from millions to trillions, directly influencing an LLM's capacity and complexity.

Prominent examples of LLMs might include GPT-5 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), or more recent open models such as DeepSeek-V3.2 (DeepSeek, 2025). In educational contexts, GenAI commonly refers to these text-generation tools (for example, ChatGPT-style chatbots) used to answer questions, draft essays, or summarise information, among many other uses.

Equally important as knowing what GenAI *is* is understanding what it *is not*. These systems are neither conscious minds nor infallible oracles, but highly advanced statistical algorithms. They work by predicting the next word in a sequence using context, imitating patterns of natural language. They do not truly *understand* the meaning of what they say, nor do they possess genuine knowledge; instead, they generate text with the *appearance* of meaning from correlations in their training data. Therefore, their fluency should not be confused with truthfulness or genuine understanding.

GenAI is not omniscient “magic” or Tolkien's Palantír (Alier-Forment et al., 2026), even if Arthur C. Clarke's Third Law – “any sufficiently advanced technology is indistinguishable from magic” (Clarke, 1973)– may ring true for many users; it remains a technology with specific scope and limitations that must be understood. Well-used, GenAI can be a valuable tool for improving quality and equity in higher education; however, when poorly used, without a critical understanding or ethical guardrails, it can have the opposite effect (García-Peñalvo, Llorens-Largo, et al., 2024; Lee et al., 2024).

Critical GenAI literacy consists precisely in equipping staff and students with that informed lens (Castañeda & Selwyn, 2018; Veldhuis et al., 2025), which means harnessing its potential with caution and knowledge, understanding what GenAI is and what it is not.

### 2.2. How it works: models, data, and training

Modern generative language models are trained using deep-learning techniques (neural networks) with Transformer architectures (Vaswani et al., 2017, 2023). Introduced by Google in 2017, the Transformer enables the processing of long text sequences by attending to relationships among words. Models such as GPT-5 have hundreds of billions of trainable parameters, affording them substantial capacity to model human language. In practice, they first undergo massive unsupervised pre-training (Brown et al., 2020), where they are fed a vast text corpus and learn to predict the next word based on what they have read, thereby capturing grammatical and semantic patterns. This breadth of training data is precisely what enables them to respond to science, history, art, or technology, depending on the context of the question. Filters are also applied during data collection to exclude highly offensive or biased language; however, many biases and imbalances present on the Internet inevitably end up being incorporated into the model (Bender et al., 2021; Weidinger et al., 2021).

After pre-training, fine-tuning typically follows. A common technique is to train the model on high-quality instruction–response pairs, enabling it to learn to follow human prompts (*instruction tuning* (Wei et al., 2022)). Additionally, Reinforcement Learning from Human Feedback (RLHF) is employed (Christiano et al., 2017; Ouyang et al., 2022): human raters score model outputs, which signals a reward model that steers the LLM to align its responses with human preferences and values.

All this requires enormous computational infrastructure, including supercomputers with thousands of GPUs (graphics processing units) that work in parallel for considerable periods (Brown et al., 2020). Hence, only organisations with substantial resources (such as technology companies and research centres) have been able to develop such LLMs, rendering their internals something of a black box for most users.

This massive scale also carries a substantial, and often invisible, environmental cost (Google, 2025). Running thousands of processors continuously requires vast amounts of energy, contributing significantly to a carbon footprint (Dhar, 2020; Schwartz et al., 2020), particularly when electricity is generated from non-renewable sources. There is also considerable fresh-water consumption (Li et al., 2025; Qiao et al., 2025) used to cool data centres and prevent server overheating. Every interaction with an AI model, from training to daily use, indirectly consumes these valuable resources (Jegham et al., 2025).

In essence, an LLM generates highly plausible text according to probabilities derived from its training. This explains both its power and its pitfalls. On the one hand, it can produce detailed responses on myriad topics, imitating diverse discourse styles. On the other hand, it lacks a guaranteed fact-checking mechanism beyond what is embedded in its weights from training, even though most mainstream chatbots can now consult the internet to provide more up-to-date factual information. This can yield content that, although coherent and plausible, is incorrect or not grounded in real data, resulting in hallucinations (Perković et al., 2024; Towhidul Islam Tonmoy et al., 2024). Moreover, the absence of a direct link to sources (except when retrieval is performed over specific sources, such as knowledge bases or the web) means the model cannot correctly cite or attribute ideas or data it provides. All these technical factors underscore the importance of users (teachers and students in academic contexts) understanding how these models operate, so they can use them in an informed way, aware of what to expect (and what not to expect) from their outputs.

### 2.3. Risks in university practice

Integrating GenAI in higher education offers undeniable advantages but also entails concrete risks that critical literacy must anticipate. In universities, these risks range from inadvertent misinformation (for example, summaries that distort scientific results or plausible hallucinations) to ethical and pedagogical dilemmas linked to integrity, traceability, and assessment. Far from justifying prohibition, they are a call to integrate these tools with critical spirit and informed caution, reinforcing verification, transparency, and responsible instructional design. This requires identifying hallucinations, recognising and counteracting biases, maintaining students' intellectual autonomy, and safeguarding academic principles (truthfulness, citation, originality) through clear guidance and competence development.

Some of the main risks identified in academic practice are (García-Peñalvo, 2024a; García-Peñalvo, Llorens-Largo, et al., 2024):

- **Hallucinations and veracity.** LLMs can hallucinate content, i.e., generate incorrect or non-existent information presented with complete apparent confidence (Ji et al., 2023). For instance, they often invent bibliographic references or academic citations that sound plausible but do not exist. Comparative studies have found that even advanced models, such as GPT-4, produce a worrying proportion of fabricated references (around 28% under specific tests). In contrast, other models, such as Bard, have reported false citation rates exceeding 90% in systematic review settings (Chelli et al., 2024). This automated lack of truthfulness is hazardous in university contexts, where students may accept erroneous information, believing “AI does not make mistakes”, and then base work on falsehoods, that is, automation bias, the human propensity to over-trust automated systems and discount one's own judgement (Romeo & Conti, 2025). Without internal mechanisms for traceability, it is difficult to verify the provenance of each assertion or detect errors without investing time in corroboration with external sources (Huang & Chang, 2024; Shao, 2025). The tendency of AI to “say something” even without reliable data (including confidently fabricated citations) demands that academics be sceptical by default and always cross-check information (Gibney, 2025; Peters & Chin-Yee, 2025).
- **Bias and equity.** Generative models inherit the biases present in their training data. They may reproduce historical or social prejudices in their outputs. For example, associating gender or racial stereotypes with particular professions, or overlooking perspectives from underrepresented regions in the academic literature (An et al., 2025; Torres et al., 2025). Used incautiously, they can reinforce inequalities or provide partial information. In addition, because many are trained predominantly in English, some GenAI systems perform worse in other languages, particularly with minority ones, thereby creating linguistic gaps (Roxas, 2024; Xu et al., 2025). Teaching staff should be aware of these biases and mitigate them by filtering outputs, adjusting prompts, or complementing with diverse sources. Inclusion is key; without vigilance, AI may perpetuate only the dominant voices in its corpus (Afreen et al., 2025; UNESCO, 2023).

- **Dependence and reduced learning.** There is a risk of excessive dependence on these tools by students and staff, who may delegate core cognitive tasks (such as searching, cross-checking, synthesis, and argumentation) to AI, thereby eroding the practice of these skills. Recent evidence suggests that overconfidence and cognitive offloading, which involves using external tools to reduce mental effort (Risko & Gilbert, 2016), are associated with lower critical thinking and cognitive engagement, mainly when students accept answers without verification (Gerlich, 2025; Zhai et al., 2024), potentially leading to intellectual laziness. This manifests in a tendency to accept the first AI output as sufficient (automation bias), to cite it as an authority, and to use it to draft work with minimal personal contribution, thereby raising risks for academic integrity and shallow learning (Bittle & El-Gayar, 2025). Neurocognitively, indications of reduced functional connectivity have been observed when academic writing is performed with continual AI assistance versus without tools, reinforcing the need for formative rather than substitutive use (Bai et al., 2023). AI-literacy frameworks and reviews on educational misinformation emphasise the importance of designing activities that require cross-checking, justification, and documentation of AI use to prevent uncritical dependence and its evaluative consequences (Fulsher et al., 2025). Academically, this yields a dual challenge: students may face frustration and penalties when defective AI responses are submitted unverified; staff may drift into delegating core tasks of design and feedback, missing opportunities to foster metacognition and expert judgement. Critical literacy should therefore delimit when AI use is appropriate (for example, brainstorming or planning support) and when students must do the work themselves to achieve programme competences; it should also establish minimum safeguards of transparency (declare use), traceability (log prompts and sources), and verification (check against scholarly literature) (Fulsher et al., 2025).
- **Lack of transparency and citation.** As noted, current models do not natively reveal their sources, hindering attribution and academic accountability; hence, the literature proposes integrating citation/justification mechanisms into LLMs as a safeguard (Huang & Chang, 2024). Unlike a search engine that lists documents, an LLM fuses what it has learned and generates a single response with no verifiable references. This opacity is at odds with the university's values of evidence-based research and correct citation. In practice, there is a growing risk: students may include AI-generated information without citing it, or, even worse, submit fabricated bibliographies generated by the model. Recent works document high rates of false citations and efforts to mitigate them (Gibney, 2025; Glynn, 2025). There are also intellectual-property implications. Models can memorise and reproduce training text, including copyrighted work, especially under specific attacks or prompts, which raises legal risks and necessitates safeguards and controls (Liu et al., 2024; Mueller et al., 2024). The absence of automatic attribution complicates reliability assessment. In this context, institutions must define when and how to cite AI (for example, as support/tutoring or as a source requiring explicit mention), and imply disclosure of use, traceability (including prompt and source logs), and external verification as good academic practice (UNESCO, 2023).

#### 2.4. Operational principles for critical literacy

Critical literacy in GenAI is not merely about knowing how to use tools but about using them with judgment within the context of academic values and practices. To conclude the conceptual framework, four operational principles are proposed to guide the rest of the paper:

1. **Verification before adoption:** every generated output is cross-checked against verifiable academic sources; avoid incorporating claims or citations without traceability (Huang & Chang, 2024; UNESCO, 2023).
2. **Explicit human agency:** AI assists rather than replaces deliberation, analysis and authorship; tasks are designed to require human judgement and metacognitive reflection (Peters & Chin-Yee, 2025).
3. **Equity and inclusion by design:** anticipate biases and linguistic gaps; adjust prompts and materials to incorporate underrepresented voices and the languages of the context (An et al., 2025; Roxas, 2024).
4. **Transparency and accountability:** every AI intervention leaves an auditable trail (brief disclosure, record of prompts/sources, revised final version) and complies with documentation and oversight obligations (European Parliament & The Council of the European Union, 2024; UNESCO, 2023).

These principles are operationalised in Section 3, which specifies competencies, evidence and classroom tools to deploy GenAI with safeguards, and connects them to the three responsible-use scenarios that form the core of the paper.

### 3. Competences for critical literacy in GenAI

Building on the framework set out in Sections 2.1-2.3, this section translates those foundations into verifiable teaching practices (Artopoulos & Lliteras, 2024). The aim is to sustain the quality, equity, and traceability of knowledge through four key elements: explicit human agency, systematic verification, inclusion by design, and transparency (European Parliament & The Council of the European Union, 2024; Huang & Chang, 2024; UNESCO, 2023). The competences are organised into five interconnected domains: 1) functional understanding; 2) critical evaluation and verification; 3) effective interaction; 4) ethics, privacy and compliance; and 5) equity and inclusion.

**Functional understanding.** Mastery of AI engineering is not required; it is sufficient to explain in plain language how GenAI works and its limits, already summarised in 2.1-2.3, and to recognise when an output demands external verification (Kassorla et al., 2024; UNESCO, 2023). This foundation demystifies the “magic” and the “black box” effect and enables the informed integration of GenAI into teaching workflows (tool selection, activity design, verification criteria). As a starting practice, it is helpful to have a model factsheet for the tool to be used (knowledge cut-off, majority or minority languages, modes with/without search, usage restrictions), accessible to students and staff (Burneo-Arteaga et al., 2025; Yang et al., 2025).

**Critical evaluation and verification.** Verification shifts from an exception to a habit. Recent literature suggests that LLMs can distort or overgeneralize scientific results, even when instructed to be precise (Peters & Chin-Yee, 2025), and that the fabrication of citations is not an isolated incident (Gibney, 2025). In the university context, if not countered, this transmits errors with an appearance of erudition (Zhai et al., 2024). The pedagogical response is not prohibition (García-Peñalvo, Llorens-Largo, et al., 2024), but designing verification so that every generated output is cross-checked against verifiable academic sources before being accepted in an assignment (Huang & Chang, 2024, with a brief record of the review. International frameworks, therefore, recommend protocols for transparency and verification, as well as targeted staff development, to ensure that validity, currency, and equity are not compromised (Frau-Meigs, 2024; UNESCO, 2023). To systematise the process, institutions can adopt the cross-cutting rubric shown in Table 1 and require task-specific checklists (presented in Table 2).

**Effective interaction (prompts and post-editing).** Interaction with GenAI rests on well-framed prompts and responsible post-editing. A good prompt is not a secret formula but an educational criterion: it provides context (module, level, objective), constraints (length, format, precision) and expectations about sources (“if you mention literature, propose titles with DOIs; these will be verified”) (Boonstra, 2025; Kotha et al., 2025). Embedding prompt engineering within university curricula has shown clear benefits (Knoth et al., 2024; Lee & Palmer, 2025). Even so, the key lies in post-editing: detecting hallucinations, adding real references, sharpening the reasoning and signing off the result with human authorship (Bedington et al., 2024; Nguyen et al., 2024). This pattern aligns use with learning: GenAI helps you start; knowledge is consolidated through review. Evidence of learning may include an annotated version with corrections, a list of verified sources, and a brief reflection on the editing decisions made.

**Ethics, privacy and compliance.** This strand is not ancillary. Regulation (EU) 2024/1689 consolidates a risk-based approach which, although not education-specific, affects university procurement, deployment and documentation (European Parliament & The Council of the European Union, 2024). In the classroom, it entails knowing the limits of assistance (academic integrity), not entering personal or special-category data into services without guarantees, and prioritising institutional deployments or private modes where appropriate (Dúo-Terrón, 2024; Hayes et al., 2025; Nam & Bai, 2023). It also requires attention to intellectual property and the legitimate use of materials (Liu et al., 2024). Every GenAI intervention should leave a trail: a brief disclosure of use, an internal record of prompts and decisions, and a revised final version. This trail enables auditing, learning and improvement (García-Peñalvo et al., 2025).

**Equity and inclusion.** This is the strand most easily overlooked yet decisive for academic quality in the medium term. LLMs inherit biases and can reproduce gender or racial stereotypes (An et al., 2025). Linguistic gaps also persist, such as performance declines in lower-resource languages and non-Anglophone corpora, with direct effects on the quality of generated output (Roxas, 2024; Xu et al., 2025). Inclusion is not delegated; it is designed. In practice, this means framing prompts with an equity lens (“include women authors and under-represented regions; avoid bias”); alternating models/modes when the language of instruction is not English; diversifying sources to avoid perpetuating the dominant canon; and explicitly rewarding such diversification in assessment. The task-type checklist in Table 2 operationalises these controls.

Critical literacy seems time-consuming if improvised; it is not if standardised. Two instruments help ease the burden and turn expectations into a shared routine. With these two elements, any lecturer can assess how GenAI has been used, not only what has been submitted:

- The cross-cutting rubric in Table 1 (five items: veracity/currency; traceability via disclosure and record-keeping; correction of hallucinations; equity/inclusion; quality of GenAI interaction), scored 0-2.
- The checklists in Table 2 (paper summary, essay, quantitative problem, code, image) specify what to check and how to evidence it.

Table 1. Cross-cutting rubric for the use of GenAI in academic tasks (0-2 per item; total 0-10)

| Item                                                             | 0 (Insufficient)                                                    | 1 (Adequate)                                                                     | 2 (Excellent)                                                                                                               |
|------------------------------------------------------------------|---------------------------------------------------------------------|----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| <b>1. Veracity and currency of sources</b>                       | No verification; unverified/ out-of-date data/citations included.   | Partial verification; obvious errors corrected, but gaps or weak sources remain. | All key claims triangulated with current DOIs/ISBNs/academic URLs; figures and citations cross-checked.                     |
| <b>2. Traceability (disclosure + record-keeping)</b>             | No AI use declared; no prompts/versions retained.                   | Generic disclosure without details; incomplete record.                           | Clear disclosure (tool, stage, limits) and auditable record (prompts, versions, decisions).                                 |
| <b>3. Correction of hallucinations/errors</b>                    | Hallucinations/fabricated citations detected but not corrected.     | Main hallucinations corrected; minor issues remain.                              | All hallucinations identified and corrected; the correction process is documented.                                          |
| <b>4. Equity and inclusion (voices/languages)</b>                | Anglocentric or low-diversity sources/examples; biases unaddressed. | Some diversifications of sources; partial attention to bias/language.            | Diverse sources/examples (gender/region/paradigm); explicit attention to the language of instruction.                       |
| <b>5. Quality of interaction with AI (prompt + post-editing)</b> | Vague prompt; output adopted without review.                        | Prompt with basic criteria; limited post-editing.                                | <i>Prompt with context, criteria and constraints; substantive post-editing (structure, accuracy, style, real citation).</i> |

*Suggested interpretation:*

0-3 = Redo with supervision; 4-6 = Acceptable with improvements; 7-8 = Good; 9-10 = Excellent.

Table 2. Task-type checklists

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Scientific paper summary</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <input type="checkbox"/> Identify the original paper (PDF) and DOI.<br><input type="checkbox"/> Verify objective, method, sample/data and results against the PDF.<br><input type="checkbox"/> Check numbers, tables and figures (identical values; no “creative averages”).<br><input type="checkbox"/> Flag and correct exaggerations or over-generalisations.<br><input type="checkbox"/> Replace fabricated citations with real references or remove them.<br><input type="checkbox"/> Record in an appendix: prompt → output → corrections → sources.                     |
| <i>Argumentative essay</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <input type="checkbox"/> State thesis and criteria (scope, limits, definitions).<br><input type="checkbox"/> Verify facts and citations with DOI/ISBN/academic URLs.<br><input type="checkbox"/> Integrate relevant non-dominant voices (authors/regions/schools).<br><input type="checkbox"/> Flag and revise potential biases in examples/analogies.<br><input type="checkbox"/> Rewrite in your own style; indicate what the AI contributed (disclosure).<br><input type="checkbox"/> Record in an appendix the decisions (why parts of the output were accepted/rejected). |
| <i>Quantitative problem / data analysis</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| <input type="checkbox"/> Replicate the procedure step by step (calculation or code) and attach evidence.<br><input type="checkbox"/> Validate units, rounding and error propagation.<br><input type="checkbox"/> If there is code: run minimum tests and add comments.<br><input type="checkbox"/> Compare with an official or standard source (manual, official database).<br><input type="checkbox"/> Record discrepancies and how they were resolved.<br><input type="checkbox"/> Record in an appendix: prompt → output → corrections → sources.                           |
| <i>Code / engineering</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <input type="checkbox"/> Specify requirements (I/O, complexity, security, licences).<br><input type="checkbox"/> Perform static (linters) and dynamic (unit tests) review.<br><input type="checkbox"/> Cite external snippets where applicable (licence).<br><input type="checkbox"/> Analyse risks (dependencies, data, vulnerabilities).<br><input type="checkbox"/> Record in an appendix: prompt → versions → rationale for changes → test results.                                                                                                                        |

(continued)

Table 2. Task-type checklists (*continued*)

| <i>Image / visual material</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="checkbox"/> State provenance (own, generated, stock) and licence; include C2PA metadata where applicable.<br><input type="checkbox"/> Verify fidelity (maps/graphs: scales, legends, projections, units).<br><input type="checkbox"/> Avoid stereotyped visual representations.<br><input type="checkbox"/> Note transformations (editing, upscaling, inpainting).<br><input type="checkbox"/> Record sources and permissions.<br><input type="checkbox"/> Record in an appendix: prompt → versions → rationale for changes. |

Staff development closes the loop. Given the pace of change, GenAI literacy requires a stance of continuous updating, as models and safeguards evolve and new regulations and guidance emerge with implications for teaching and assessment (European Parliament & The Council of the European Union, 2024; Kassorla et al., 2024; UNESCO, 2023). Ongoing training, communities of practice, and methodological adjustments are necessary conditions for responsible integration (Abegglen et al., 2024; Jin et al., 2025; Nerantzi et al., 2023; Sozon et al., 2025).

It is worth underlining that this critical literacy does not compete with the usage scenarios at the core of the paper; it enables them. In Scenario 1 (responsible support), basic verification and a brief disclosure are sufficient to ensure quality without overburdening the process. In Scenario 2 (guided collaboration), the focus is on traceable iteration (learning journals, explicit rubrics, version comparison). In Scenario 3 (co-creation with reinforced disclosure), traceability is complete, encompassing prompts, versions, sources, decisions, and peer review. In all cases, human agency is the hinge that turns technical capability into meaningful academic practice.

Critical literacy in GenAI turns the promises and limits outlined in 2.1-2.3 into professional habits: understanding enough to avoid the illusion of infallibility; verifying before adopting; designing interaction to learn, not merely to produce; respecting privacy and regulation; and ensuring that all voices (languages, regions, perspectives) enter the classroom. Promoting GenAI literacy and fostering responsible digital competences across the educational community is essential (Vivas Urias & Ruiz Rosillo, 2025). By mastering these competences, academic staff can convert the arrival of GenAI into a pedagogical opportunity, enriching teaching with innovation and avoiding myths and threats (García-Peñalvo, 2024b). With the instruments presented in Tables 1 and 2, these practices become teachable, observable, and assessable. In this way, the university prepares for a future in which GenAI is integrated into the creation and dissemination of knowledge, expanding critical thinking, creativity, and rigour, never constraining them.

#### 4. Standards and safeguards

This section links the professional habits from Section 3 to four frameworks that today serve as reference points for safe AI use in academic contexts: UNESCO’s global guidance (2023), Regulation (EU) 2024/1689 (the AI Act) (European Parliament & The Council of the European Union, 2024), the SAFE framework (Safety, Accountability, Fairness, Efficacy) (EDSAFE AI, 2021), and, serving as the backbone, the Safe AI in Education Manifesto (Alier et al., 2024). All four imply decision-making at both classroom and institutional levels, with a focus on the principles underlying the previous sections, namely explicit human agency, privacy and security assurance, ethical commitment, and transparency.

##### 4.1. The Manifesto as the guiding thread

The Safe AI in Education Manifesto (Alier et al., 2024) outlines principles for human oversight and the right to appeal, confidentiality, alignment with strategy and pedagogy, accuracy/explainability, interface and behaviour transparency, and ethical training and data transparency (García-Peñalvo, Alier, et al., 2024). The spirit is clear: AI complements educators and never replaces them; AI-generated content must be marked and verifiable; and the institution must be able to audit how the tool was used. This aligns with the four proposed vectors (agency, verification, inclusion, and transparency) and the instruments outlined in Tables 1 and 2.

This manifesto connects to the proposed scenarios as follows:

- **Scenario 1 (responsible support):** apply accuracy, explainability, and transparency through a brief disclosure of use, visible marking of AI-generated content, and basic verification before incorporating outputs.
- **Scenario 2 (guided collaboration):** reinforce supervision, the right to appeal, and didactic alignment through iterative versions with logging, instructor review, and opportunities for students to discuss/rectify inferences.

- **Scenario 3 (co-creation with reinforced disclosure):** explicit criteria on data/sources/biases; complete audit trail and peer review.

#### 4.2. UNESCO: Human-centred vision and institutional capacity

UNESCO's guidance (2023) calls for immediate actions (ethics, privacy, security, equity, transparency, and literacy), medium-term policies, and staff development to sustain a human-centred vision.

This guidance connects to the scenarios as follows:

- **Scenario 1 (responsible support):** disclosure of use, verification, and no use of personal data.
- **Scenario 2 (guided collaboration):** Strengthen faculty capacity through Rubrics.
- **Scenario 3 (co-creation with reinforced disclosure):** ethical and pedagogical validation before high-exposure uses (for example, AI-assisted peer assessment).

#### 4.3. AI Act: Proportional risk and transparency obligations

The AI Act (European Parliament & The Council of the European Union, 2024) is not an education statute. Still, it structures adoption by risk levels, distinguishing minimal risk (no obligations), limited risk (with transparency obligations, for example, informing users they are interacting with AI and marking synthetic content), and high risk (risk management, data governance, technical documentation, registration, and impact assessments where applicable). For education, the key is proportionality: in non-evaluative learning uses, reinforce transparency; when AI affects academic decisions (for example, grading), emulate "high-risk" controls, with human review, full traceability, bias analysis, and documentation.

This regulation connects to the scenarios as follows:

- **Scenario 1 (responsible support):** treat as limited risk, implying consistent disclosure of use, marking AI-generated content, and prohibiting reliance on AI output as the sole basis for grading.
- **Scenario 2 (guided collaboration):** if AI is used for feedback or learning pathways, prepare purpose-and-risk documentation at the course level and keep interaction logs.
- **Scenario 3 (co-creation with reinforced disclosure):** employ "high-risk"-style controls: human review before academic effects, bias audit, and technical documentation accessible for accreditation.

#### 4.4. SAFE: From principle to practice

The EDSAFE AI SAFE framework (2021) bridges public policy and the classroom. It summarises what any implementation must guarantee: Safety (S), Accountability (A), Fairness (F), and Efficacy (E), and provides practical resources: use policies, resource libraries, acceptable-use templates, and professional-development materials. For departments and institutions, SAFE is the operational checklist that complements the Manifesto (principles) and the AI Act (compliance).

This framework connects to the scenarios as follows:

- **Scenario 1 (responsible support):** S = no personal data; A = disclosure and logging; F = quick bias checks; E = leave evidence of added learning.
- **Scenario 2 (guided collaboration):** more A and E, implying traceable versioning, rubrics, and with/without-AI comparisons.
- **Scenario 3 (co-creation with reinforced disclosure):** full SAFE, with periodic bias/error audits and efficacy testing of AI-assisted instructional design.

#### 4.5. Roadmap for safe GenAI use in education

All four references converge on a practical message: yes, to GenAI, but with human agency, proportional risk, and verifiable transparency. The Manifesto provides the classroom idiom (supervision, privacy, accuracy, explainability, transparency); UNESCO contributes a human-centred vision and institutional capacity; the AI Act brings the risk architecture and obligations on marking and documentation; and SAFE offers the resource bridge between policy and practice. With this compass, the paper's three scenarios become coherent choices in

design, assessment, and procurement, with graduated safeguards and learning metrics that avoid both improvisation and paralysis.

A layered system is proposed, with a minimum sequence (see Table 3), linking the scenarios to the four frameworks.

Table 3. Layers for a roadmap to safe GenAI use in education

| Layer 1: Baseline transparency (for all uses)                                                                                                                                                                                                                                                                                                                                                                           | Layer 2: Proportional risk management (when AI mediates learning/assessment)                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Layer 3: Reinforced controls (when AI affects sensitive decisions)                                                                                                                                                                                                                                                                                                                                            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>• Disclosure of use (tool, stage, limits).</li><li>• Visible marking of AI-generated content and logging of prompts/versions.</li><li>• Prior verification of facts/citations (fabricated bibliographies prohibited).</li></ul> <p><i>(Manifesto: transparency &amp; accuracy; UNESCO: human-centred vision; AI Act Art. 50: content transparency; SAFE: accountability).</i></p> | <ul style="list-style-type: none"><li>• Risk matrix per activity (impact on grade, data processed, foreseeable biases, mitigations).</li><li>• Structured, documented screening in the educational context (mini-DPIA: who/what for/what data/how supervised and appealed to ensure a full DPIA is not required).</li><li>• Evidence of instructional efficacy (what learning improvement justifies use).</li></ul> <p><i>(AI Act: risk-based approach; Manifesto: supervision and accuracy; SAFE: efficacy and accountability).</i></p> | <ul style="list-style-type: none"><li>• Mandatory human review and students' right to appeal.</li><li>• Bias audits and multilingual performance checks; communicate limitations.</li><li>• Complete technical documentation and traceability for evaluation and accreditation.</li></ul> <p><i>(Manifesto: supervision, privacy and data transparency; AI Act: reinforced obligations; SAFE: S/F/A).</i></p> |

Note on DPIA: In the EU context, the “mini-DPIA” is a light-touch screening to decide whether a full Data Protection Impact Assessment (DPIA) is required under GDPR Article 35 (European Parliament & Council of the European Union, 2016), and to document mitigations (for example, human oversight and appeal, data minimization, access controls, marking, verification).

5. Three scenarios for using AI in education

What follows defines three usage scenarios, including their typical activities, risks, safeguards, minimum evidence and metrics, explicitly connected to the Safe AI in Education Manifesto (Alier et al., 2024), UNESCO’s guidance for GenAI in education and research (2023), the AI Act (Article 50) (European Parliament & The Council of the European Union, 2024), and the SAFE framework (EDSAFE AI, 2021).

5.1. Scenario 1: Responsible support (low risk, low autonomy)

This first scenario describes an instrumental and non-summative use of AI. Teachers use AI tools to support their work. The tool aids in thinking, structuring, and refining prose, yet it does not make substantive decisions on its own. This is the space for instructional design and content preparation, where tasks include generating study outlines, proposing self-assessment questions, clarifying complex concepts with alternative explanations, and drafting teaching materials that are then reviewed and validated by staff. The pedagogical key is that AI is assistance, not a shortcut; it opens possibilities rather than closing the reasoning. The Safe AI in Education Manifesto effectively captures this philosophy: AI complements educators rather than replacing them; its intervention should be transparent and auditable; and adopted outputs must be verifiable, not taken on trust.

Risks, while contained, do exist. The most common is the appearance of truth without veracity: AI synthesises fluently but may introduce errors or overstate conclusions. Human cognitive biases also play a role (for example, overconfidence in a well-written response), and privacy can be compromised if personal data is shared with external services. None of this requires prohibition; it does require good habits. Here, the UNESCO guidance is pragmatic: it emphasises immediate actions around marking, literacy, and verification, as well as policies that preserve human agency, and staff development that enables them to explain limits and good practices.

Safeguards are therefore proportionate and straightforward. First, transparency: briefly disclose that AI has been used, mark AI-generated content visibly, and keep a simple record of the flow “prompt → output → corrections”. The AI Act reinforces this with transparency obligations for limited-risk systems, requiring users to be aware when they are interacting with AI and when AI has generated synthetic content. Such marking and disclosure are consistent with Article 50. Secondly, prior verification: every factual claim and every reference should be cross-checked against academic DOI/ISBN/URL sources before inclusion. Thirdly, data minimisation: avoid entering personal or sensitive information into services without assurances and prefer institutional deployments where available. Lastly, didactic fit: each use should serve a clear learning purpose (clarity, examples, re-explanation), rather than aiming to substitute for intellectual work.

How is responsible use evidenced? For student work, a brief disclosure, visible marking of generated passages, and a short post-editing note stating what was accepted and what was corrected are sufficient. Operational indicators then follow: the rate of claims or citations verified versus those corrected; and, in style-driven tasks, observable gains in clarity and coherence between draft and final version. For privacy, the threshold is zero incidents. Again, the Manifesto provides the classroom idiom and the SAFE framework as a bridge between policy and practice (safety and accountability). At the same time, UNESCO emphasises that transparency and literacy are the foundation of a human-centred approach.

This scenario turns AI into a visible assistant for staff, easing preparation and understanding without displacing authorship or verification. It is the natural entry point for instilling habits of marking, cross-checking and mindful use, and it serves as a shared culture before advancing to more intensive forms of collaboration. The balance is straightforward: low risk, high transparency, and academic judgment upfront.

Figure 1. Characteristics of the responsible support scenario.

|                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Teachers use AI tools to support their work</b> | <p><b>Need for high and constant literacy</b></p> <p><b>Low/moderate risk</b>, the lower the more experience the teaching staff has in the subject, the greater their level of digital/AI competence, and the more relevance is maximised in the process. <b>High risk</b> in summative assessment if agency is lost</p> <p><b>Principle of transparency</b>: declare in which tasks/processes AI has been used and how; this is important for maintaining a relationship of trust with the students</p> <p><b>Privacy</b>: do not upload real personal data (with special attention to clinical data)</p> <p><b>Scalability and sustainability</b>: what resources are being used (free versions, personal, or institutional licences)?</p> |
|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 5.2. Scenario 2: Guided collaboration (moderate risk, shared autonomy)

Here, the teacher integrates AI tools into activities with students. AI co-participates in creative or analytical processes, but always through traceable iteration and with meaningful human post-editing. An essay may begin with a dialogue with the tool to explore counter-arguments. However, students should rewrite in their own words, replace suggested references with actual citations, and explicitly explain how their reasoning evolved across versions. In programming, the tool accelerates scaffolding and testing, but the person retains design ownership, adds unit tests, and provides comments on decisions. In data analysis, AI can propose exploratory routes; students execute, validate assumptions, correct errors, and document traceability. The educational value lies in making decision-making and improvement across iterations visible, with lecturer review as a real safeguard, in line with the Manifesto (human oversight and right to appeal) and UNESCO's protocols of transparency and verification.

Risk falls between low and moderate, lower as staff expertise and digital/AI competence increase and as pedagogic relevance is maximised. The focus shifts from the isolated erroneous datum to the carry-over of errors across versions if post-editing is weak. Over-reliance may emerge if synthesis or argumentation is delegated to others. With external search, grounding (Kenthapadi et al., 2024) or RAG (Retrieval-Augmented Generation) (Zhao et al., 2024), well-grounded hallucinations can appear when sources are poorly anchored. The answer is not to slow iteration but to rail-guard it. Use reinforced traceability, a prompt log and explicit versioning (v1, v2, v3) with diffs and comments explaining what was accepted or discarded; ensure meaningful human review before publishing automated feedback or finalising an assessed version; bias and language checks, ensuring the model maintains quality and does not import prejudice (SAFE-Fairness); and where student data are processed, complete a lightweight DPIA screening (mini-DPIA) and escalate to a full DPIA if needed. Throughout, maintain clear marking and disclosure so that students are aware of which parts of feedback or drafts were generated by AI and under what specific conditions.

Evidence in this scenario is inherently process-centred: a version history, a list of verified sources with DOI/ISBN/URL, and a post-editing note explaining what changed and why. The rubric in Table 1 (veracity and currency, traceability, correction of hallucinations, equity and inclusion, quality of interaction with AI) supports

consistent assessment across modules. Sensible metrics include iteration gains (for example, argument quality or code robustness), correction rate of hallucinations, efficiency in time without sacrificing validity, diversity in the final bibliography, and compliance with prompt/version logs. The target is not merely a better final product but better-justified thinking and a more traceable process, coherent with the Manifesto, UNESCO's human-centred vision, the AI Act's principle of proportionality, and SAFE's logic of accountability/efficacy.

Guided collaboration makes AI a practice partner that speeds up, provokes and suggests, while the human discerns, corrects and signs. Speed is gained without sacrificing method, and students are taught to think with the tool while maintaining intellectual and ethical control of the process. This is the natural setting for developing prompting skills with discernment, post-editing, and systematic verification, as summarised in Figure 2.

Figure 2. Guided collaboration features.

**Teachers integrate AI tools into activities with students**

**Need** for very high teacher **literacy** and high **student literacy**

**Low/moderate risk**, the lower the more experience the teaching staff has in the subject, the greater their level of digital/AI competence, and the more relevance is maximised in the process

**Principle of transparency**: define in the tasks which AI tools (and for what uses) are acceptable or mandatory, and how to declare their use

**Authentic assessment**, putting **emphasis** on the **process**, **avoiding blind trust in detectors**, and **incorporating techniques to gather evidence** (viva, in-class work, lab notebooks, iterations, incremental submissions, etc.)

**Principle of equity** to ensure all students can complete the tasks

**Privacy**: do not upload real personal data (with special attention to clinical data)

**Scalability and sustainability**: use institutional accounts and licences

### 5.3. Scenario 3: Co-creation with reinforced disclosure (high impact, high traceability)

This scenario represents a situation in which students will use AI tools autonomously for their own learning process. It is related to various types of academic products, ranging from course submissions to results that can have a high academic or public impact, as they may be deposited in institutional repositories, open-source repositories, or open exhibitions, thereby circulating beyond the classroom.

It is not that AI does more in terms of autonomy, but rather that the risk of the consequences is greater (lower the more maturity the student has in maintaining the agency of their learning and the greater their digital and AI competence); therefore, the safeguards must approach those of high-risk systems under the logic of the AI Act, with mandatory human review, full traceability, bias audits, and technical documentation accessible for review and accreditation. All this without losing the didactic focus; co-creation with AI remains a learning process where authorship is made explicit, veracity is validated, and there is honest communication about which AI tools assisted in the result.

Integrity and authorship become central issues. It is essential to distinguish between human contributions and automated assistance, as well as to avoid fabricated bibliographical references and reasoning that appear sound but are not. Equity also escalates in importance. If the product is public, biases related to gender, race, or region, as well as linguistic gaps when working in minority languages, cease to be a minor detail and become a quality requirement. Added to this are intellectual property and privacy risks when manipulating real data or materials with incompatible licences, and technical risks of prompt injection or manipulated sources when using grounding or RAG. The Manifesto and the SAFE framework provide guidelines for adequate human supervision, the right of appeal, privacy, accuracy, explainability, interface and data transparency, security, accountability, fairness, and efficacy.

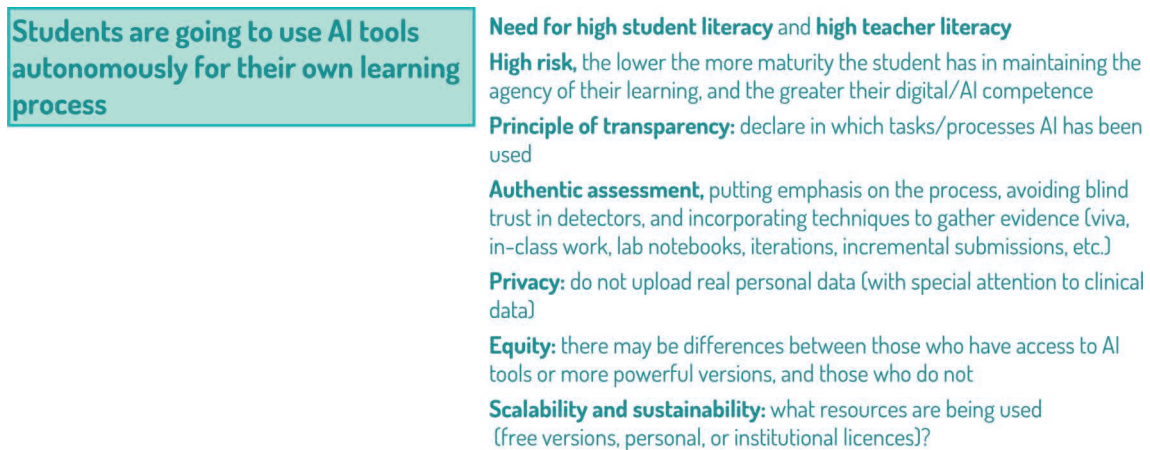
Operationally, this scenario requires declarations of use accompanying the product: who did what, full trace of prompts and versions, primary sources verified with DOI/ISBN/URL, a verification report explaining what was contrasted and how, an equity and language check, and a record of human review (and, where applicable, peer review). AI-generated content must be visibly marked, and the final document must include a public statement of methodology, thereby fulfilling the principle of transparency outlined in Article 50, so that the recipient

does not doubt whether there was AI intervention. In parallel, a bias audit and a multilingual test should be conducted when the language of teaching or dissemination is not English, clearly communicating any limitations and mitigations. If personal data is processed, the institution should document a DPIA screening or, if the risk advises it, a complete assessment.

The metrics must be more demanding, for example, the percentage of claims with a verifiable source and the reproducibility of the result; incidents detected and corrected in human and peer review; equity indicators (diversity of voices/regions in references and cases); presence and detectability of the AI marking; and compliance with privacy and licences without exceptions. Efficacy is demonstrated by formative impact (mastery of the process, better argumentation, higher technical quality, transferability of the product). In terms of governance, the Manifesto–UNESCO–AI Act triangle provides guidance on human agency and the right of appeal, a person-centred vision, and controls proportionate to the risk.

Figure 3 summarises this scenario, highlighting co-creation with a reinforced declaration, precisely so that it is not merely more of the same, but rather that responsibility is increased through the use of AI. Traceability is ensured due to the impact of the result. That is, this scenario requires a culture of evidence and auditing that fully aligns with international standards and current regulations, preparing students to confidently navigate the Generative AI era in their future professional lives.

Figure 3. Characteristics of co-creation with reinforced disclosure.



## 6. From map to practice: Stages of the teaching cycle

Following the presentation of the three scenarios, they are linked to the four main stages of the teaching cycle: i) planning and instructional design; ii) material creation (which can be subdivided into searching for and synthesising evidence, creating educational resources, and activities related to data management and programming); iii) learning support; and iv) assessment. These stages must operate under the umbrella of an ethics and transparency layer that offers the safeguards inherent in the frameworks presented in Section 4. This structure is reflected in the diagram in Figure 4, where the stages are complemented by a catalogue of tools that can currently be used in these phases.

### 6.1. Planning and instructional design

This stage is the natural habitat of Scenario 1 (responsible support) and Scenario 2 (guided collaboration). In practice, GenAI can support i) the description of learning outcomes, the design of activities, the proposal of evidence and assessment criteria, etc.; ii) the analysis of risks and requirements, what data will be handled, what biases are foreseeable, which language(s) will be used; iii) the selection of tools and their institutional fit; among others. From a safeguards perspective, UNESCO calls for immediate actions (literacy, privacy, equity, transparency) and institutional capacity (policies and faculty development); SAFE translates this into checklists for safety, accountability, fairness, and effectiveness; and the Manifesto recalls that all integration must make human oversight, privacy, and transparency explicit.

Figure 4. Stages of the teaching cycle.

## AI Tools for Teaching



Made with Napkin

### 6.2. Material creation

This phase is specific to Scenario 1 (materials that the teacher reviews and validates) and Scenario 2 (when working with students in a co-design setting). GenAI can assist in finding and synthesising academic evidence and preparing specific activities related to programming and data analysis. However, its most significant potential lies in GenAI becoming an assistant for creating diverse content, enriching materials with complementary elements (such as examples, counter-examples, exercises, etc.), as well as visual, interactive, and personalised content (for example, alternative explanations for different levels).

In content creation, there are two basic operating rules. First, the declaration of use when part of the materials has been generated with AI, making access to the 'prompt→output→corrections' trace recommendable. This transparency is consistent with the obligations of Article 50 of the AI Act when dealing with synthetic content (text, image, audio, or video). Second, verifying facts, figures, and citations, as well as addressing potential language gaps and underrepresented voices, is necessary to avoid curricular biases. By following the Manifesto (accuracy, explainability, and transparency) and SAFE (S/A/F/E), AI-generated content becomes a revisable input, not a substitute for teacher authorship.

### 6.3. Learning support

Scenario 2 (guided collaboration) naturally develops in this stage, although content related to tutoring, guidance, and other similar aspects can be generated in a context typical of Scenario 1 (responsible support).

The process becomes the focus for evidence control in this stage. UNESCO emphasises transparency and verification protocols, as well as faculty development. SAFE reinforces accountability, and the Manifesto provides guidelines for teacher interaction, oversight, and the right to appeal.

Process evidence (version history, list of verified sources, post-editing note, etc.) is valued using the rubric in Table 1 and the checklists in Table 2.

### 6.4. Assessment

Assessment is the most sensitive point and, therefore, where the three scenarios are most clearly graduated. Formative assessment (for example, in feedback on scheduled submissions or exercises) may appear in Scenario 2, with AI intervention flagging, traceability, and human review before consolidating the response.

When assessment is summative with high-impact products or substantive decisions (for example, public final projects, prototypes with third parties, exams), the reinforced controls typical of Scenario 3 must be applied, regardless of whether GenAI is used by the faculty or the students, including mandatory human review, declaration of use, prompts and versions, etc., as well as a public statement in accordance with Article 50 of the AI Act so that the recipient does not doubt as to whether or not AI intervention occurred and on what terms. The determining factors are impact and traceability, rather than who operates the tool.

Suppose a student uses GenAI in a module. In that case, the rubric in Table 1 and the checklists in Table 2 can be applied in summative assessment to evaluate veracity, currency, traceability, correction of hallucinations, equity, language aspects, and the quality of interaction with the AI. Sectoral guidance (Joint Council for Qualifications, 2025; Office of Qualifications and Examinations Regulation, 2024, 2025; Walker, 2025) converges on reinforcing academic integrity, elevating the authenticity of tasks, and enhancing oral defences (Wang et al., 2024) or other evidence of agency and ownership of the process, discouraging reliance on detectors as the primary strategy.

Furthermore, evidence shows that patterns of malpractice evolve and that detecting AI is substantially different from detecting plagiarism (Sadasivan et al., 2025; Weber-Wulff et al., 2023); comparative studies on detectors report limitations, reinforcing the need for designs that verify agency and traceability instead of relying on detection tools, which, in any case, can serve as an initial alert in some situations, but never replace human judgement.

### 6.5. Closing the loop

In the four stages that represent an abstraction of the teaching cycle, coherence is achieved if each decision answers three questions: 1) pedagogical purpose (what the AI contributes to learning); 2) risk and safeguards (what exposure or impact it has and what controls are in effect); and 3) evidence and metrics (what leaves a trace, what can be audited, how it improves).

The practical rule is one of proportionality; that is, the greater the impact (or data sensitivity), the closer it is to the controls established in Scenario 3. Conversely, the more exploratory and non-evaluative it is, the closer it is to the lighter regime typical of Scenario 1. This graduation of safeguards materialises the AI Act's risk architecture (transparency for limited uses; reinforced controls when there are high-impact decisions) and avoids both indiscriminate prohibition and uncritical adoption.

This map does not add unnecessary bureaucracy but makes visible routine what defines academic culture (citation, verification, and accountability) and aligns it with the Manifesto (oversight, privacy, and transparency), UNESCO guidelines (human-centred vision and institutional capacity), the obligations of the AI Act (marking and notification of synthetic content), and the operational bridge of SAFE (S/A/F/E with resources and templates).

Closing the loop, the value of this mapping is not in the taxonomy but in conveying the habits of good practice. Scenario 1 introduces transparency and verification, Scenario 2 trains judgment and iteration, while Scenario 3 demands evidence and auditing. Having coupled stages and scenarios allows each teacher to know what to do, what to ask for, what to review, and how to justify the adoption of GenAI with international standards and current regulations, while maintaining intact human authorship, equity, and the quality of learning.

## 7. Discussion and Conclusions

This article has proposed a pragmatic and proportionate way to integrate Generative AI (GenAI) into higher education through three scenarios graduated by autonomy, agency, and risk, linked to professional habits (Section 3) and regulatory-ethical frameworks (Section 4), and put into practice in the teaching and learning cycle (Section 6). The common thread is clear. AI as a supplement, always supervised by academic judgement, not as a substitute; transparency and traceability as fundamental premises; equity and inclusion by design; and proportionality of risk to determine safeguards. This proposal is in close dialogue with the Safe AI in Education Manifesto (human supervision, privacy, accuracy, explainability, and transparency), UNESCO's guidelines (a human-centred vision, immediate actions, and institutional capacity), the AI Act (transparency and labelling of synthetic content; strengthened controls according to exposure), and the SAFE framework as an operational bridge.

The value of the three scenarios lies in turning broad principles into verifiable teaching decisions. In Scenario 1 (responsible support), the rule is characterised by low risk and high transparency: a brief declaration of use, marking of generated passages, factual verification, and the absence of personal data in services without guarantees. In Scenario 2 (guided collaboration), the key is traceable iteration, followed by significant post-editing: versions with tracked changes, a list of verified sources, and a note on the decisions made. In Scenario 3 (co-creation with a strengthened declaration), the focus shifts to robust evidence and auditing (prompts and versions, checks for bias and language, human review, and, where appropriate, peer review), consistent with Article 50 of the AI Act and the principles of SAFE. This gradient of safeguards materialises the AI Act's risk-based approach without transferring unnecessary bureaucracy to the classroom.

In terms of roles, the scenarios are primarily aligned with the use of GenAI by teaching staff as an assistant (Scenario 1), its incorporation into student activities (Scenario 2), and autonomous use by students (Scenario 3). However, in high-stakes summative assessments, controls specific to Scenario 3 may be required, even if the lecturer uses the AI, because the determining factor is the risk and impact of the outcome, not who uses the tool. This reinforces the proportionality of the AI Act and UNESCO's emphasis on institutional safeguards.

Precisely in the field of assessment, sector-specific guidance in the United Kingdom points in this direction. Jisc recommends reinforcing authenticity, agency, and ownership of the process in assessment redesign; the Office of Qualifications and Examinations Regulation (Ofqual) frames the use of AI in the qualifications sector by emphasising quality and fairness; and the Joint Council for Qualifications (JCQ) is updating rules for centres on disclosing use, integrity, and malpractice, advising against relying on detectors as a primary strategy. These guidelines reinforce this paper's premise that assessment should verify agency and traceability rather than focusing on detecting AI use, as well as the value of oral defences or other performative evidence as instruments consistent with this aim.

In terms of institutional feasibility, the proposal is supported by two levers that facilitate adoption. On the one hand, a cross-cutting rubric (veracity and currency, traceability, correction of hallucinations, fairness and language, and quality of interaction) allows for consistent evaluation across different subjects. On the other hand, checklists for each type of task (summary, essay, quantitative problem, code, image) turn transparency and verification into routine practices. This minimum set of guarantees (declaration of use, marking, basic logging, and verification) is fully consistent with the Manifesto and UNESCO. It is scalable towards strengthened controls depending on the impact (as outlined in the AI Act).

The ecosystem of models and policies is evolving rapidly; specific operationalisation (for example, which exact field a declaration of use includes or how to store logs while ensuring privacy) may vary between institutions and jurisdictions. Furthermore, measuring the impact (on learning, equity, and workload) will require quasi-experimental studies and longitudinal designs. Lastly, linguistic and infrastructural gaps necessitate sustained investment and multilingual evaluation of tools to prevent transferring inequalities into the classroom. Institutional roadmaps will need to be reviewed periodically, in line with UNESCO's emphasis on institutional capacity and staff development, which magnifies the value of AI governance in universities (Molina-Carmona & García-Peñalvo, 2025).

The arrival of GenAI does not force a choice between enthusiasm and prohibition; it forces us to teach and learn with safeguards in place. The three proposed scenarios enable universities and academic staff to make proportionate decisions, evaluate them with evidence, and be transparently accountable, while always maintaining human agency and fairness as guiding criteria. In practice, marking, verifying, and documenting are not bureaucratic procedures, but rather the contemporary ways of safeguarding academic culture. Suppose the Manifesto provides the language for the classroom. In that case, UNESCO offers the human-centred vision, the AI Act establishes the risk architecture, and SAFE provides the operational bridge. Universities already have the map to turn AI into a pedagogical opportunity without compromising on rigour, justice, and responsibility. This, ultimately, is the standard that must guide the teaching and learning process in the age of AI.

## 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 Working 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/202juE9>
- 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). UNESCO. <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>
- Knonth, 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](https://doi.org/10.1007/978-3-031-71530-3_14)
- Mueller, F. B., Gorge, 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 Study 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., Parnter, 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> Tabla 1. Propuesta de rúbrica transversal para uso de IAGen en tareas académicas (0-2 por ítem; total 0-10).