

Instrumentos de evaluación, corrección y retroalimentación con apoyo de inteligencia artificial

Construir instrumentos de evaluación con apoyo de inteligencia artificial y situar con precisión la frontera entre lo que puede asistirse y lo que debe seguir siendo decisión del docente.

- 1 Construcción de rúbricas y criterios de evaluación
- 2 Elaboración de bancos de ítems y control de calidad de las pruebas
- 3 Corrección asistida: alcance de la evidencia disponible y límites legales
- 4 Retroalimentación al alumnado asistida por inteligencia artificial
- 5 Interpretación de resultados y toma de decisiones de mejora

3.1. Construcción de rúbricas y criterios de evaluación



Rúbricas y criterios: construirlas, no encargarlas.

Nuria da Lengua Castellana y Literatura en 2.º de Educación Secundaria Obligatoria (ESO) y tiene 28 exposiciones orales por delante. Pide «una rúbrica para evaluar exposiciones orales en 2.º de ESO» y en 15 segundos tiene una tabla impecable de cuatro dimensiones y cuatro niveles. La usa el lunes. Tras la tercera exposición, las tres han caído en la misma casilla —«bastante claridad», «bastante organización»— y la nota se la sigue poniendo ella. La rúbrica no era mala: era decorativa.

Una rúbrica es un instrumento de medida y un modelo de lenguaje produce texto plausible. Ahí está el problema. Lo que sigue reparte el trabajo: qué decides tú, qué puedes encargar y cómo compruebas que el resultado evalúa el criterio del currículo.

Lo que la herramienta hace bien y dónde empieza a rellenar

Hace bien todo lo que es forma. Estructura en tabla, mantiene la simetría entre niveles, ofrece cinco redacciones de un descriptor que no convence, la pasa a segunda persona para el alumnado y detecta que dos dimensiones se solapan. No compromete nada.

Empieza a rellenar en cuanto le pides que decida. Sin criterio, escribirá descriptores sobre la idea genérica de exposición oral que arrastra de su entrenamiento; sin umbral, situará el aprobado donde suene bien. Y como su producto por defecto suena a evaluación —«de manera adecuada», «mostrando dominio»—, pasa cualquier lectura rápida y falla en lo que importa: separar a un alumno de otro.

La regla de reparto es esta: tú aportas el criterio, las dimensiones y el umbral de suficiencia; la herramienta aporta redacción, simetría y variantes. Invertirla produce **rúbricas decorativas**.

Del criterio a las dimensiones: decidir qué se mira

El punto de partida está en el criterio de evaluación literal de tu decreto autonómico de currículo, no en el tema ni en la tarea. En Castilla-La Mancha rige el Decreto 82/2022, de 12 de julio, por el que se establece la ordenación y el currículo de Educación Secundaria Obligatoria en la comunidad autónoma de Castilla-La Mancha (DOCM de 14 de julio de 2022), modificado por el Decreto 31/2024, de 25 de junio (DOCM de 2 de julio de 2024), que retoca ese currículo y la asignación de materias optativas. En Andalucía, el Decreto 102/2023, de 9 de mayo (BOJA núm. 90, de 15 de mayo de 2023). En las demás, el decreto de currículo equivalente.

Los criterios no coinciden en redacción ni en numeración entre comunidades, así que se copia el propio: si se lo pides a la herramienta, te lo recitará con una numeración verosímil que no existe. Si quieres que trabaje sobre el decreto, apórtaselo como contexto con los límites de 1.4.

Del criterio salen las dimensiones: las tres a cinco cosas distintas que vas a mirar. No ocho: cada dimensión añadida es una decisión más por alumno, y las largas se rellenan en diagonal. Nuria fija cuatro desde el criterio de producción oral de su decreto: adecuación al destinatario, organización del discurso, corrección y riqueza léxica, y elementos no verbales y de apoyo. «Originalidad», «esfuerzo» y «actitud» entran por inercia. Si una dimensión no procede del criterio, o la quitas o la sacas de la calificación.

Cada dimensión se convierte después en indicadores **observables**: conductas que ocurren delante de ti. «Mira al público» lo es. «Muestra seguridad» obliga a inferir un estado interno, y dos docentes marcarán cosas distintas. La herramienta depura bien esa lista si se lo pides.

Descriptores de nivel: el punto donde la rúbrica se rompe

Pídele descriptores sin más instrucción y obtendrás una escalera de adverbios: «con mucha claridad / con bastante claridad / con poca claridad / sin claridad». Eso no llega a ser una escala: quien decide el grado sigues siendo tú, sin dejar rastro. Si tapas los adverbios y los niveles dicen lo mismo, no discriminan.

Lo corrigen tres técnicas. La primera separa los niveles por un hecho y no por un adverbio: una diferencia de cantidad, de frecuencia o de condición que puedas señalar. En «organización del discurso», los niveles de Nuria quedan así: *anuncia el guion y cierra recuperándolo / el orden se sigue sin esfuerzo / se entiende, pero hay que reconstruirlo al escuchar / encadena ideas sin orden reconocible*.

La segunda empieza por el nivel de suficiencia. Si el modelo arranca por arriba, describe a un alumno imposible y luego va restando. La instrucción, en el formato de plantilla parametrizada de 1.3, queda así:

PLANTILLA

Escribe cuatro descriptores para la dimensión [dimensión] del criterio [criterio literal]. Cada nivel se distingue del anterior por un hecho observable —cantidad, frecuencia o condición—, nunca por un adverbio de grado. No uses adecuado, correcto, bastante, muy ni excelente. Empieza por el nivel 3, umbral en el que doy por superado el criterio.

La tercera afecta al número de niveles y al más bajo. Cuatro niveles, sin casilla central: con cinco, la del medio se vuelve refugio. Y el nivel más bajo describe lo que el alumno sí hace —«lee íntegro el texto de la pantalla»—, no lo que le falta. «No cumple» no sirve para la retroalimentación posterior (3.4).

Ponderar: una decisión pedagógica, no una operación aritmética

La ponderación por defecto es un reparto simétrico, 25 % en cada dimensión, y siempre parece razonable. Casi nunca lo es. El peso lo marca la centralidad de la dimensión en el criterio, no lo cómodo de observarla: así es como lo fácil de contar acaba pesando lo mismo que lo nuclear.

Nuria pondera organización 30 %, adecuación 25 %, corrección y riqueza léxica 25 %, y elementos no verbales y de apoyo 20 %. Falta decidir algo que casi nadie escribe: cuánto vale cada nivel dentro de su peso. Fija que el nivel 4 vale el peso íntegro; el 3, la mitad, por ser el umbral; el 2, la cuarta parte, y el 1, cero. Quien alcanza el nivel 3 en las cuatro dimensiones obtiene justo el 50 % y supera. Sin esa tabla la rúbrica ordena, pero no produce una calificación.

El caso límite se prueba antes de aplicar la rúbrica a nadie. Un alumno con diapositivas impecables y buen contacto visual está en el nivel 4 de elementos no verbales y de apoyo. Se ajusta al destinatario y se expresa con corrección suficiente: nivel 3 en esas dos dimensiones. Y encadena ideas sin orden reconocible: nivel 1 en organización. La cuenta: 20 de 20, más 12,5 de 25, más 12,5 de 25, más 0 de 30. Total, 45 %. No supera, y es lo coherente con un criterio de producción oral.

Conviene estirar el caso. Ese mismo alumno, en el nivel más alto de las otras tres dimensiones, sumaría 70 % pese al discurso sin orden: un 30 % no basta para bloquear la dimensión nuclear. Caben dos salidas, y conviene elegir una por escrito: subir el peso de organización o impedir por regla que se supere con nivel 1 ahí.

Un aviso de frontera. Que la rúbrica lleve porcentajes no significa que la calificación pueda salir sola de una herramienta. El anexo III del Reglamento (UE) 2024/1689 (Reglamento de IA) sitúa entre los usos de alto riesgo en educación los sistemas destinados a «evaluar los resultados del

aprendizaje, también cuando se utilicen para orientar el proceso de aprendizaje». El reparto de obligaciones entre proveedor y centro y el calendario de aplicación los fija 3.3. Aquí construyes el instrumento; quien lo aplica sigues siendo tú.

Comprobar que la rúbrica evalúa el criterio y no otra cosa

Queda comprobar que el instrumento mide lo que dice medir, antes de que la rúbrica toque un trabajo real.

El primero es de trazabilidad: en una tabla de doble entrada, cada dimensión apunta a una parte del criterio y cada parte aparece en alguna dimensión. Lo que sobra o falta se corrige ahora.

El segundo busca el **constructo desplazado**, el error que ninguna lectura rápida detecta. Es la rúbrica de exposición oral que mide el diseño de las diapositivas, la de resolución de problemas que acaba midiendo ortografía, la de proyecto de Formación Profesional (FP) que premia la memoria sobre el procedimiento. La pregunta de control: quien hace lo que el criterio pide, ¿podría suspender esta rúbrica?

El tercero es la **aplicación en seco**. Sitúa con la rúbrica tres trabajos de cursos anteriores que colocabas en niveles distintos. Si dos caen en la misma casilla, reescribe esa dimensión y no bajes la exigencia.

El cuarto es la doble mirada. Que un compañero del departamento corrija dos de esos trabajos con tu rúbrica: una discrepancia sistemática en la misma dimensión señala un descriptor ambiguo. Cómo se acuerda esto en equipo está en 4.5.

Una rúbrica es un instrumento de calificación: le corresponde el nivel 3 de verificación de 1.5, el más alto de la escala. Deja anotado con qué herramienta se construyó y cómo se validó, con el procedimiento de 4.3.

FIGURA 3.1

Del criterio de evaluación a la rúbrica utilizable



Las cinco decisiones que el docente toma antes de que la herramienta escriba nada —criterio literal, dimensiones, indicadores observables, descriptores desde el umbral y ponderación— y la comprobación de validez que cierra el proceso. Elaboración propia.

FIGURA 3.1

Aplicación práctica

1. Coge una rúbrica tuya, copia en su cabecera el criterio literal de tu decreto —es lo que te permite responder ante una reclamación— y tápale los adverbios de grado en una dimensión. Si los cuatro niveles dicen lo mismo, reescríbelos separándolos por un hecho observable, desde el umbral.
2. Escribe la tabla de conversión de nivel a porcentaje y prueba un caso límite: el alumno que falla en la dimensión nuclear y cumple en las demás. Si supera, sube ese peso o añade una regla de mínimos.
3. Antes de aplicar una rúbrica nueva, sitúa con ella tres trabajos de cursos anteriores que ya sabías distinguir. Es la prueba más barata de que discrimina.

3.2. Elaboración de bancos de ítems y control de calidad de las pruebas



Bancos de ítems y pruebas: generar y controlar la calidad.

Álvaro da Geografía e Historia en cuarto curso de Educación Secundaria Obligatoria (ESO) y pide 30 preguntas de opción múltiple sobre la Revolución Industrial. Las recibe en un minuto, bien redactadas y con las cuatro opciones alineadas. Las lee por encima, le parecen razonables y monta el examen. En clase descubre dos cosas. La primera, que 26 alumnos aciertan la pregunta cuatro sin haber estudiado, porque la respuesta correcta era la única que ocupaba tres líneas. La segunda, que 28 de las 30 preguntas caen sobre la primera mitad del tema: máquinas, fábricas y crecimiento urbano. Solo dos rozan la nueva sociedad de clases y el movimiento obrero, que es lo que había trabajado durante dos semanas.

Ninguno de los dos problemas se ve leyendo por encima. Este epígrafe recorre el procedimiento para que no lleguen al aula: qué se decide antes de generar nada, qué se pide para cada tipo de ítem y cómo se depura el banco en una pasada ordenada.

Antes de generar: la tabla de especificaciones

Un banco de ítems se controla por cobertura, y la cobertura se decide antes de escribir la primera instrucción. Cabe en una tabla de dos entradas de media hoja: en las filas, los criterios de evaluación que la prueba debe cubrir; en las columnas, el nivel cognitivo; en las celdas, cuántos ítems quieres de cada cosa.

Para el nivel cognitivo basta una escala de cuatro peldaños, del tipo de la taxonomía de Bloom revisada (Anderson y Krathwohl, 2001): recordar, comprender, aplicar y analizar. No hace falta más finura, y sí hace falta un reparto explícito. El reparto por defecto de cualquier herramienta tiende siempre al mismo sitio: preguntas de recordar, las más fáciles de generar y las que mejor suenan al leerlas. Si no pides seis ítems de aplicar, no los tendrás.

Álvaro rehace su encargo con cuatro filas —transformaciones técnicas, cambios demográficos y urbanos, nueva sociedad de clases y movimiento obrero— y reparte los 30 ítems en siete, siete, nueve y siete. De ellos, 12 se sitúan en los niveles de aplicar y analizar. La pregunta que reconstruiremos más adelante, la del ítem cuatro, ocupa ahora una celda concreta: nueva sociedad de clases por comprender.

Con la tabla delante, la instrucción deja de ser «30 preguntas sobre la Revolución Industrial» y pasa a ser una petición por celda, en tandas de cinco o seis. Es el uso de plantilla parametrizada de 1.3, con el criterio literal del decreto de tu comunidad pegado dentro. Si quieres que la herramienta trabaje sobre el temario que has impartido, aporta tus apuntes o la programación como contexto, con los límites que marca 1.4. Las tandas cortas tienen además una razón técnica: los lotes largos degradan. A partir del ítem 15, el modelo repite estructuras y recicla el mismo distractor.

Qué pedir según el tipo de ítem

En la opción múltiple, especifica el número de opciones, una sola respuesta inequívocamente correcta y distractores de longitud y estructura gramatical semejantes a las de la correcta. La petición que más cambia el resultado es otra: **que cada distractor represente un error real y frecuente del alumnado**, no una opción absurda. Prohíbe expresamente «todas las anteriores», «ninguna de las anteriores» y los enunciados en negativo.

Sobre cuántas opciones usar, Rodríguez (2005) reunió 80 años de investigación en pruebas de rendimiento. Su conclusión: tres opciones bien construidas rinden lo que cuatro, y así caben más ítems en el mismo tiempo de examen. Tómallo como permiso para no rellenar, no como norma:

un distractor que nadie elige no está midiendo nada.

En la respuesta corta, el punto delicado no está en la pregunta, sino en la corrección. Pide siempre, junto al ítem, la respuesta esperada y la lista de respuestas alternativas aceptables: sinónimos, formulaciones equivalentes y grados de precisión admisibles. Sin esa lista, la pregunta corta acaba puntuándose por parecido con lo que tenías en la cabeza.

En las preguntas de desarrollo, lo que menos cuesta es el enunciado. Necesitas que acote la tarea —extensión, número de aspectos, qué debe relacionar el alumno— y que la acompañe de indicadores de corrección. Ese instrumento es una rúbrica y se construye con el procedimiento de 3.1.

Los seis defectos típicos del ítem generado

Aparecen porque el modelo optimiza plausibilidad, y una respuesta correcta plausible tiende a estar mejor argumentada, más matizada y a ser más larga que las falsas. Cada uno se caza con una comprobación distinta.

El primero, y el dominante, es que **la respuesta correcta resulta más larga o más precisa que las demás**. Un alumno entrenado en exámenes tipo test lo detecta sin saber la materia. El segundo son los distractores inverosímiles: opciones que nadie elegiría, que reducen la pregunta de cuatro a dos alternativas reales y regalan medio acierto a quien duda. El tercero son las pistas involuntarias. Delata, por ejemplo, la concordancia gramatical —«un proceso que...» con una sola opción masculina singular—, o la repetición literal de una palabra del enunciado en la opción correcta.

El cuarto es la convergencia de opciones. Tres distractores dicen variantes de lo mismo y el cuarto va por libre; el que va por libre suele ser el bueno y se elige por eliminación. El quinto son los ítems encadenados, el 12 que no se puede responder si has fallado el 11: penaliza dos veces el mismo error y contamina la lectura de resultados que se trabaja en 3.5. El sexto es el contenido inventado con apariencia de dato: cifras, fechas y nombres que el modelo añade para dar concreción. Es el fallo descrito en 1.5 aplicado aquí, y en un examen ese error no se corrige hablando, se reclama.

Depurar: una pasada por columnas, no ítem por ítem

Revisar preguntas de una en una rinde poco, porque los defectos son sistemáticos. Se depura por columnas: la misma comprobación en los 30 ítems seguidos.

La primera pasada es la de la clave sola. Tapa los enunciados y lee solo la lista de respuestas correctas. Si la mayoría son la b y la c, reordena; si son visiblemente las más largas, reescribe.

La segunda pasada mira los distractores. Para cada opción falsa, pregúntate qué alumno concreto la elegiría y por qué. La que se quede sin respuesta, se sustituye. Aquí la herramienta ayuda de verdad, con una instrucción de este tipo:

PLANTILLA

Reescribe los distractores de este ítem de forma que cada uno corresponda a un error conceptual frecuente en 4.º de ESO. Iguala su longitud a la de la respuesta correcta y explícame qué error representa cada uno.

La tercera pasada es la resolución en seco: responde tú la prueba entera sin mirar la clave. Es la única comprobación que detecta el ítem con dos respuestas defendibles y el que no tiene ninguna. La cuarta vuelve a la tabla de especificaciones y coloca cada ítem en su celda. Lo que sobra en una fila y falta en otra se corrige ahora, no después del examen.

Así queda el ítem cuatro de Álvaro. Antes: «¿Cuál fue una consecuencia social de la Revolución Industrial? a) La aparición del proletariado industrial, una nueva clase social formada por trabajadores asalariados que vendían su fuerza de trabajo en las fábricas a cambio de un salario. b) La desaparición de las ciudades. c) La prohibición del comercio internacional. d) El regreso masivo de la población al campo». Acumula tres defectos: correcta larguísima, dos distractores absurdos y una opción d que nadie elige.

Después: «a) Un aumento del trabajo asalariado en fábricas concentradas en las ciudades. b) Un aumento del trabajo artesanal en talleres familiares urbanos. c) Un aumento del trabajo agrícola por la mejora de las cosechas. d) Un aumento del trabajo doméstico rural organizado por comerciantes». Las cuatro opciones tienen longitud pareja y los tres distractores recogen confusiones reales: la pervivencia del taller artesanal, el efecto de la revolución agrícola y el trabajo a domicilio anterior a la fábrica. Para acertar hay que distinguir entre cuatro situaciones plausibles.

Los seis defectos del ítem generado y la pasada que los caza

Seis defectos del ítem generado por IA

Pasada 1. La clave sola

- Defecto 1: la respuesta correcta es la más larga o la más precisa
- Comprobación: tapa los enunciados y lee solo la lista de correctas
- Delata también la concentración de aciertos en una misma letra

Pasada 2. Los distractores

- Defecto 2: distractores inverosímiles que nadie elegiría
- Defecto 4: convergencia de opciones (tres iguales y una por libre)
- Comprobación: di qué alumno concreto elegiría cada opción falsa

Pasada 3. La resolución en seco

- Defecto 3: pistas involuntarias de concordancia o palabra repetida
- Defecto 6: cifras, fechas y nombres inventados con apariencia de dato
- Comprobación: responde tú la prueba entera sin mirar la clave

Pasada 4. Cobertura y montaje

- Defecto 5: ítems encadenados que penalizan dos veces el mismo error
- Comprobación: coloca cada ítem en su celda de la tabla de especificaciones
- Comprobación: ningún ítem contiene la respuesta de otro posterior

Cada defecto característico del ítem generado por IA tiene una comprobación concreta que lo detecta en el depurado por columnas.

FIGURA 3.2

Del banco a la prueba: montaje, límites y rastro

Terminado el banco, queda montar el examen, que es tarea aparte. Ordena de menor a mayor dificultad, comprueba que ningún ítem contiene la respuesta de otro, revisa que la extensión quepa en la sesión real y decide tú la puntuación. Un examen es material de nivel 3 de 1.5: verificación total, contraste con el criterio del decreto y, si califica, una segunda mirada del departamento.

Conviene también situar el límite jurídico sin agrandarlo. Generar preguntas no equivale a evaluar aprendizajes. El supuesto de alto riesgo del anexo III del Reglamento (UE) 2024/1689 (Reglamento de IA) alcanza a los sistemas destinados a «evaluar los resultados del aprendizaje» y a «vigilar y detectar comportamientos prohibidos durante los exámenes»; ese régimen, con su calendario de aplicación, se trata en 3.3. Aquí produces un instrumento y lo firmas tú. Si el banco se comparte en el departamento, deja anotado que se generó con apoyo de inteligencia artificial y quién lo depuró, con el procedimiento ligero de 4.3.

Aplicación práctica

1. Antes de pedir un solo ítem, dibuja la tabla de criterios por nivel cognitivo y encarga por celdas, en tandas de cinco o seis.
2. Haz la pasada de la clave sola en el próximo test que tengas hecho: lee solo las respuestas correctas y comprueba si son las más largas o si se concentran en una letra.
3. Convierte en instrucción fija de tu plantilla la petición de que cada distractor represente un error frecuente y de que todas las opciones tengan longitud pareja.

3.3. Corrección asistida: alcance de la evidencia disponible y límites legales



Corrección asistida: hasta dónde llega la evidencia y hasta dónde la ley.

Lorena cierra el segundo trimestre en un aula de cinco años, el último del segundo ciclo de Educación Infantil, con 22 niños y el cuaderno del proyecto del mercado. Tiene que decidir, criterio a criterio, qué ha aprendido cada uno, y se hace la pregunta de cualquier sala de profesores: si la herramienta valora parecido a como valoro yo, ¿por qué no dejarla valorar?

Contestarla obliga a mirar en dos sitios: la investigación sobre la concordancia con la corrección docente, y el Reglamento (UE) 2024/1689 (Reglamento de IA), que sitúa la evaluación de los resultados del aprendizaje en el alto riesgo. De ambos sale la misma frontera: asistir no es delegar.

La evidencia: qué está documentado, y quién lo firma

La concordancia se mide casi siempre con el **kappa ponderado cuadrático** (QWK): 0 es coincidir por azar; 1, coincidir siempre.

El trabajo que más pesa —un preprint— es el de Gaggioli y otros (2025): 67 composiciones universitarias en italiano, cinco modelos y cuatro dimensiones de rúbrica. La concordancia con los correctores humanos fue baja y no significativa en QWK. Peor aún: al repetir tres veces la misma corrección con la misma instrucción, la mediana de la W de Kendall quedó por debajo de 0,30. Un corrector que no coincide consigo mismo no es un corrector.

En español hay ya trabajo revisado por pares. Reina y otros (2026) publican en la *RIED* la validación de una aplicación web de ayuda a la corrección de exámenes de desarrollo, con 91 exámenes reales de un curso tipo MOOC —curso en línea masivo y abierto—. Midieron concordancia con el profesorado y consistencia entre tres versiones de la misma corrección, y concluyen lo que sostiene este epígrafe: la herramienta apoya el juicio del docente, no lo sustituye.

La tercera mira el conjunto. Li y otros (2025) sintetizan 65 estudios de 2022 a 2025 sobre escritura. También es preprint, pero una síntesis con método declarado: los valores van de -0,10 a 0,97, habitualmente entre 0,30 y 0,80, y la disparidad de métricas impide una media agrupada. Retén dos cosas: la concordancia sube cuando la instrucción incorpora la rúbrica y ejemplos resueltos —la plantilla de 1.3—, y 53 de los 65 estudios son en inglés, con solo tres sobre Educación Primaria: en español y en las primeras etapas la evidencia es delgadísima.

Dónde se sostiene y dónde se cae, según la tarea

Las tres dibujan lo mismo: la corrección asistida se parece a la docente en la distribución del grupo y mucho menos en el alumno individual, que es sobre quien se juzga. Y la concordancia no es propiedad de la herramienta, sino de la tarea: ordena las tayas por el margen interpretativo que dejan.

Aguanta bien lo que tiene respuesta cerrada y verificable: opción múltiple, verdadero-falso, cálculo con procedimiento único. Ahí no juzga, compara. Aguanta a medias la respuesta corta y las rúbricas analíticas con descriptores operativos, las de 3.1: cuanto más observable y menos adverbial el descriptor, más se sostiene.

Se cae en todo lo demás: valoración global de un ensayo, originalidad, argumentación, proyecto de Formación Profesional, respuesta correcta por vía no prevista. Con un patrón reconocible: la **tendencia a premiar la coherencia aparente**. El texto bien encadenado y vacío puntúa alto; el desordenado con una idea excelente, bajo. En Infantil no hay siquiera producto textual que leer: la evidencia es conducta observada, y eso deja fuera lo que evalúa.

Qué concordancia está documentada entre corrección asistida y corrección docente



< 0,30 Consistencia del modelo consigo mismo (W de Kendall) al repetir tres veces la misma corrección 67 composiciones universitarias en italiano, cinco modelos y cuatro dimensiones de rúbrica (Gaggioli y otros, 2025, preprint)	Baja y no significativa Concordancia (QWK) con los correctores humanos en ese mismo estudio Un corrector que no coincide consigo mismo tampoco coincide con el docente	91 exámenes Validación en español, en revista revisada por pares: la herramienta apoya el juicio del docente, no lo sustituye Exámenes escritos de desarrollo de un curso tipo MOOC; concordancia con el profesorado y consistencia entre tres versiones de la misma corrección (Reina y otros, 2026, RIED)	0,30-0,80 QWK habitual con el docente en una síntesis de 65 estudios sobre tareas de escritura Li y otros (2025), preprint. Rango total: de -0,10 a 0,97, sin media agrupada posible; 53 de los 65 estudios son en inglés y la evidencia en español y en las primeras etapas es delgadísima
---	---	--	--

Fuente: Gaggioli y otros (2025), Reina y otros (2026) —en revista revisada por pares— y Li y otros (2025); los rangos de esta última son globales, porque la literatura sintetizada no los desagrega tarea a tarea.

FIGURA 3.3

El límite jurídico: por qué esto es alto riesgo

El anexo III, punto 3, del Reglamento de IA recoge cuatro usos de alto riesgo en educación:

- determinar el acceso o la admisión a centros educativos y de formación profesional, o la asignación a ellos;
- **evaluar los resultados del aprendizaje, también cuando se utilicen para orientar el proceso de aprendizaje;**
- evaluar el nivel de educación adecuado que una persona recibirá o al que podrá acceder;
- vigilar y detectar comportamientos prohibidos durante los exámenes.

El segundo describe lo que Lorena se plantea, y no hay excepción por etapa: en Infantil no hay calificación, pero sí un grado de adquisición que orienta la respuesta educativa y llega al informe a las familias. Su aplicación quedó aplazada al 2 de diciembre de 2027 por el Reglamento (UE) 2026/1744 (Ómnibus Digital sobre IA), en vigor desde el 27 de julio de 2026. Es aplazamiento de calendario, no derogación: lo que hoy montes lo tendrás dentro. El artículo 50 y la transparencia son materia de 4.1.

Esa clasificación admite un matiz, el del **artículo 6.3**: un sistema del anexo III puede no considerarse de alto riesgo cuando no plantee un riesgo importante de perjuicio ni influya sustancialmente en la toma de decisiones. Concreta cuatro supuestos: tarea de procedimiento limitada; mejora de una actividad humana ya realizada; detección de patrones sin sustituir ni influir en la evaluación humana ya hecha; y tarea preparatoria. Y será siempre de alto riesgo el sistema que elabore perfiles de personas físicas.

Conviene precisar de qué es propiedad ese matiz, porque no es un refugio en el que se entre y del que se salga reescribiendo la petición del día. El artículo 6.3 se predica del sistema y de su finalidad prevista, no de cada consulta, y el artículo 6.4 obliga a quien lo invoca a documentar esa

valoración antes de introducir el sistema en el mercado y a registrarlo igualmente. La decisión es previa, escrita e imputable a quien provee o despliega.

De ahí, dos consecuencias de aula. Una herramienta pensada para calificar no vuelve al 6.3 porque hoy le pidas que «solo agrupe»: sigue evaluando resultados del aprendizaje. Y configurar un asistente generalista para que produzca el grado de adquisición no te mete en el 6.3, sino en el artículo 25.1.c), que hace proveedor a quien cambia así la finalidad de un sistema de uso general. Dentro del matiz queda lo que no sustituye ni influye en un juicio docente ya emitido — depurar indicadores, agrupar anotaciones, revisar una valoración hecha— y solo mientras esa salida no viaje al informe.

Qué le cae al centro si se cruza la línea

El centro que despliega un sistema del anexo III asume las obligaciones del responsable del despliegue, y la primera es la vigilancia humana. El **artículo 14** vincula al **proveedor**: debe diseñar el sistema, interfaz incluida, para que personas físicas puedan vigilarlo de forma efectiva mientras esté en uso: entender sus límites, interpretar su salida y poder descartarla o detenerlo. El **artículo 26.2** vincula al **centro**: debe encomendar esa vigilancia a personas con la competencia, la formación y la autoridad necesarias. Le toca además informar a las personas afectadas.

Se añaden dos cargas. Los organismos de Derecho público y quienes prestan servicios públicos deben realizar antes del primer uso una evaluación de impacto relativa a los derechos fundamentales (artículo 27) y notificarla a la autoridad de vigilancia del mercado. Y el artículo 86 da a la persona afectada derecho a explicaciones sobre el papel del sistema en la decisión: ahí puede terminar la reclamación de una familia.

Asistir sin delegar: el trimestre de Lorena

Lo que Lorena no hace cabe en una frase: volcar el cuaderno, con los nombres dentro, y pedir un nivel por niño. Serían dos errores: delegaría el juicio que el anexo III sitúa en alto riesgo y sacaría del centro datos de menores identificables, la frontera de 3.4, que aquí aprieta más.

Su procedimiento es otro. Construye primero el instrumento, sin ningún dato de aula: pasa los criterios de evaluación del área «Descubrimiento y exploración del entorno» a indicadores observables, con el método de 3.1. Es tarea preparatoria y no sale un solo dato. Después observa ella, dos semanas, en la asamblea y en los rincones: la observación directa y sistemática que el artículo 12 del Real Decreto 95/2022 fija como técnica principal de la evaluación global, continua y formativa.

La comprobación siguiente calibra el instrumento, no a los niños. Toma 10 anotaciones y las reescribe como conducta observable, sin identificadores —«apila cinco piezas y las cuenta señalando cada una»—, comprobando que ninguna arrastre una circunstancia familiar, un dato de salud o una medida de apoyo. Con esas 10 pide tres cosas que no son valorar: qué indicador documenta cada descripción, cómo se agrupan por dificultad y cuáles están escritas en términos valorativos —«es muy tímido»— en vez de observables. Si señala indicadores que ella no vio, el fallo está en la redacción del indicador, no en su juicio: la valoración docente no se cambia porque la máquina discrepe.

El grado de adquisición lo decide ella, niño a niño. El informe a las familias es material de nivel 3 de verificación (1.5): deja constancia de en qué se apoyó y cómo lo validó, con el registro ligero de 4.3.

Aplicación práctica

1. Antes de asistir ninguna valoración, resuelve tú a mano 10 trabajos o 10 anotaciones: si la herramienta no coincide contigo en esos 10, no la uses en el resto.

2. Escribe, antes de usar la herramienta y no consulta a consulta, para qué se emplea y qué hace el docente después. Si ahí no cabe decir que el juicio ya estaba emitido o que la salida no llega al informe, no estás en el supuesto del artículo 6.3.
3. Lleva a la Comisión de Coordinación Pedagógica (CCP) una regla de una línea antes de diciembre de 2027: ningún grado de adquisición ni ninguna calificación procede de la salida de un sistema de IA sin decisión docente sobre cada alumno.

3.4. Retroalimentación al alumnado asistida por inteligencia artificial



Retroalimentación: donde la IA sí aporta de verdad.

Pregunta en cualquier claustro qué tarea se hace peor de lo que se debería y la respuesta se repite: los comentarios individuales. No por desidia. Devolver a 25 alumnos un párrafo que cada uno pueda usar cuesta más que preparar la unidad. Así que se devuelve una nota y unas correcciones en rojo que el alumno mira tres segundos.

Aquí es donde una herramienta generativa rinde más: no sustituye un juicio que tú ya haces, sino que produce un material que hoy no llega a existir. La calificación la pones igual; el comentario de 10 líneas no lo escribes. Pero lo que devuelve por defecto no sirve, y lo que necesitaría para mejorarlo no puede salir del centro.

Por qué aquí sí, y qué significa exactamente «aquí»

Conviene separar dos cosas que se confunden. Una es **corregir**: decidir si el trabajo alcanza el criterio y con qué nivel. Tiene evidencia discutida y régimen jurídico propio, y se trata en 3.3. Otra es **redactar la devolución** de un juicio ya tomado: escritura repetitiva y de baja irreversibilidad, porque un comentario desacertado se corrige hablando el lunes. Ese es el perfil donde una herramienta de lenguaje aporta sin comprometer nada.

La investigación coincide en un punto que ordena todo lo demás: el efecto de la retroalimentación depende de cuánta información contenga, no de su extensión ni de su amabilidad. El metaanálisis de Wisniewski, Zierer y Hattie (2020), sobre 435 estudios, sitúa el comentario de alta información en un tamaño de efecto de 0,99 y el que se limita a reforzar o a penalizar en 0,24. Hattie y Timperley (2007) ya habían fijado las preguntas que debe responder: adónde va el alumno, cómo va y qué hace después.

Comparar comentarios generados con comentarios humanos ya tiene síntesis publicada: Kaliisa, Misiejuk, López-Pernas y Saqr (2025) reúnen 41 estudios con 4.813 estudiantes y no hallan diferencias significativas de rendimiento. El borrador parte, pues, de un suelo aceptable, y el resultado lo deciden la información que aportas y tu revisión. El riesgo real no es el fraude, sino industrializar el elogio vacío.

Qué información del alumnado no puede salir del centro

El límite duro cabe en una línea: trabajas sobre el producto, no sobre la persona.

No salen del centro, con ninguna herramienta que este no haya validado, el nombre y los apellidos, el correo corporativo, el número de expediente ni las fotografías. Tampoco los informes psicopedagógicos, los dictámenes de escolarización ni ningún diagnóstico o dato de salud: son categoría especial del artículo 9 del Reglamento (UE) 2016/679 (RGPD). Quedan fuera igualmente las circunstancias familiares, las medidas de protección de menores, las becas y los expedientes disciplinarios. Y el texto del propio alumno cuando los contiene: una redacción sobre «un fin de semana en mi casa» puede ser un dato familiar.

Queda un aviso: un grupo de clase es una población pequeñísima y quitar el nombre no anonimiza. «Alumna repetidora de 2.º de Educación Secundaria Obligatoria (ESO) con dislexia» identifica a una persona para cualquiera que conozca el centro. El RGPD llama a eso **seudonimización**, en su artículo 4.5, y el dato seudonimizado sigue siendo dato personal. Anonimizar de verdad exigiría que reidentificar fuera imposible, y en un grupo de 25 casi nunca lo es.

De ahí no se sigue que quitar los identificadores sea inútil. Es lo primero que exige el protocolo de 1.5 antes de que un material salga, y reduce la exposición de forma decisiva; lo que no hace es dejar el tratamiento fuera del RGPD ni sustituir a la decisión sobre qué herramienta se usa.

Sí puedes aportar la producción sin identificadores, el criterio de evaluación literal, la rúbrica con los niveles que ya has decidido y la descripción neutra de lo observado. Con eso se genera el borrador. La personalización —el nombre, el tuteo, la entrega anterior, el plazo— la haces después, en tu documento. Ese paso de 30 segundos es el que más pesa: los identificadores directos no salen y el vínculo con la persona se queda dentro del centro.

Quién decide esto no eres tú en solitario: el responsable del tratamiento es el centro. Con una plataforma contratada con encargo de tratamiento del artículo 28 del RGPD el margen es mayor. La ficha de idoneidad de 1.2 termina ahí, en uno de sus tres veredictos sobre un uso concreto: «apta con datos de alumnado», «apta solo para trabajo docente sin datos personales» o «no apta». Sin ese veredicto por escrito, rige el segundo.

Del borrador genérico al comentario que el alumno puede usar

El borrador por defecto falla siempre igual. Elogia sin evidencia, no se ancla a ningún criterio, señala seis cosas a la vez y promete una evolución que nadie sostiene, con un registro que un alumno de 5.º de Primaria no descodifica. Todo eso se corrige en la instrucción, no reescribiendo el resultado.

Un comentario utilizable se sostiene sobre cuatro piezas que conviene pedir por su nombre. Qué está logrado, con la evidencia delante: no «buena estructura», sino qué parte del trabajo demuestra que el criterio se alcanza. Qué falta, referido a la tarea y nunca a la persona: «el párrafo no cuantifica el hecho» informa, «eres poco riguroso» no. Una sola cosa que hacer después, con verbo de acción, porque la tercera no se hace. Y cuándo, porque lo que no puede reutilizarse se lee como una sentencia.

La instrucción, como plantilla parametrizada según 1.3, se parece a esta:

PLANTILLA

Actúas como docente de [materia o módulo] en [nivel]. Te doy el criterio de evaluación literal, la rúbrica y el trabajo de un alumno sin identificar.

Escribe un comentario de 90 a 120 palabras, en segunda persona y con registro adecuado a su edad: qué está logrado, citando la evidencia del trabajo; qué impide alcanzar el siguiente nivel de la rúbrica; una única acción en imperativo, y el plazo. Sin elogios genéricos, sin mencionar esfuerzo ni actitud y sin prometer resultados.

Iván imparte Comunicación Empresarial y Atención al Cliente en un ciclo de grado medio de Formación Profesional (FP). Sus 24 alumnos entregan una carta de reclamación y su rúbrica, construida con el procedimiento de 3.1, mide estructura, registro, eficacia y corrección lingüística. El primer borrador dice: «Has realizado un buen trabajo. Se aprecia esfuerzo y dedicación. La estructura es adecuada, aunque podrías mejorar la redacción. ¡Sigue así!». Ninguna información y nada que hacer el martes.

Con esa instrucción y la rúbrica, el borrador cambia y él lo remata así: «La carta tiene la estructura completa —encabezamiento, exposición y cierre— con las fórmulas de la unidad 4. Lo que impide que sea una reclamación eficaz es que el hecho no está fechado ni cuantificado: escribes que “el pedido llegó incompleto”, pero no dices cuándo ni qué faltaba, y así el proveedor no puede tramitarla. Reescribe el segundo párrafo con la fecha de entrega, el número de albarán y las unidades que faltaron. Tienes hasta el viernes y lo vuelvo a corregir».

La máquina aportó la redacción. Iván, el resto: sabe que ese alumno ya entregó una reclamación sin fechar en octubre, decide corregir un solo párrafo y compromete una segunda revisión.

FIGURA 3.4

Del borrador genérico al comentario que el alumno puede usar



Los seis pasos de la retroalimentación asistida. El juicio evaluador y todo dato identificativo se quedan dentro del centro; a la herramienta solo salen el producto sin identificadores, el criterio y la rúbrica. Elaboración propia.

FIGURA 3.4

Lo que aporta siempre el docente, y el límite jurídico

Del comentario final respondes tú, y hay partes que no se delegan. **El juicio sobre si el criterio se alcanza** se toma antes de generar el borrador y no se contrasta con él. También la veracidad: el modelo escribirá con naturalidad que el alumno «ha mejorado respecto a la entrega anterior» aunque no haya visto ninguna, y eso, en un informe que llega a una familia, es afirmación tuya. Y el **compromiso**: plazos, recuperaciones y tono.

El anexo III, punto 3, del Reglamento (UE) 2024/1689 (Reglamento de IA) incluye entre los usos de alto riesgo los sistemas destinados a «evaluar los resultados del aprendizaje, también cuando se utilicen para orientar el proceso de aprendizaje». Un sistema que puntúa el trabajo y de ahí deriva la orientación entra en el supuesto; una herramienta que redacta la devolución de una valoración ya hecha por ti, no.

Si el centro desplegara un sistema del primer tipo, el reparto de la vigilancia humana entre proveedor y centro —artículos 14 y 26.2— y su calendario están en 3.3.

Un comentario que se queda en el aula es material de nivel 2 de la escala de 1.5; el que va a un boletín o a un informe sale del centro, es nivel 3 y exige comprobación previa de datos personales.

Que sobreviva a un grupo de 30

Un procedimiento que solo funciona con tres alumnos no llega a procedimiento. Genera por tandas de cinco o seis trabajos, no de 30: los lotes largos repiten la misma frase con distinta ropa, y quien compara comentarios lo detecta antes que tú. Reserva el comentario extenso para las tareas que se reescriben; en las demás, tres líneas ancladas al criterio rinden más. Y deja constancia del apoyo de inteligencia artificial y de quién revisó, con el procedimiento de 4.3.

Aplicación práctica

1. Escribe una vez la plantilla con las cuatro piezas —logro con evidencia, obstáculo referido a la tarea, una sola acción, plazo— y guárdala en tu biblioteca de 1.3.
2. Antes de pegar un trabajo en una herramienta, quítale nombre, grupo y todo rasgo identificativo, y comprueba que el texto del alumno no traiga datos familiares ni de salud.
3. Revisa el borrador buscando una frase concreta: cualquier afirmación sobre la evolución del alumno que tú no hayas observado. Es el error que más daño hace fuera del aula.

3.5. Interpretación de resultados y toma de decisiones de mejora



Leer los resultados y decidir la mejora.

Terminas de corregir, pasas las notas y el resultado cabe en una frase: «han ido regular». En la sesión de evaluación se comenta que la unidad ha costado, en el acta se escribe que se reforzará y en marzo se imparte igual. Ese es el análisis que no cambia nada: descriptivo, tardío y sin destinatario.

Este epígrafe trata del otro análisis, el que empieza en una tabla de resultados y termina en dos o tres decisiones sobre la próxima unidad. Tampoco es tarea opcional: la evaluación de los procesos de enseñanza figura entre las funciones del profesorado del artículo 91.1.b) de la Ley Orgánica 2/2006, de 3 de mayo, de Educación. Lo que aporta la herramienta es el cálculo y la ordenación que hasta ahora obligaban a resolver esa función de memoria.

Antes de abrir nada, fija el producto. Un análisis útil termina en una lista corta de decisiones, y cada decisión tiene cuatro partes: qué se hace, cuándo, con quién y cómo sabrás si ha funcionado. Dos o tres, no ocho, por la razón ya vista en las rúbricas de 3.1: lo que no cabe en el calendario real no se ejecuta. Y los destinos posibles son cuatro: reenseñar en clase, incorporarlo a la unidad siguiente como contenido previo, corregir el instrumento porque el fallo era suyo, o aceptar el resultado. El cuarto es legítimo; lo que no vale es tomarlo por defecto cuando el análisis se detiene en la media.

Preparar la tabla: cada ítem atado a su criterio

La calidad del análisis depende casi por entero de cómo entra el dato. Necesitas una matriz simple, en hoja de cálculo o en texto separado por comas: una fila por alumno, una columna por ítem y, en cada celda, el acierto, el error o la puntuación parcial. Los nombres no entran: se sustituyen por códigos —A01, A02— y la correspondencia se queda en tu ordenador. Qué información del alumnado no sale del centro se trata en 3.4; la mecánica de aportar un archivo propio, en 1.4.

A esa matriz le añades una segunda tabla, pequeña, que convierte números en información pedagógica: **qué criterio de evaluación mide cada ítem**. Si generaste la prueba con la tabla de especificaciones de 3.2, ya la tienes; si no, hazla ahora, aunque sea a mano. Sin ella la herramienta te dirá qué preguntas han fallado, y ese dato no llega a la programación. Con ella te dirá qué aprendizajes no se han alcanzado, que es el idioma de los criterios y de las actas.

Una advertencia operativa. El cálculo aritmético sobre tablas es donde un modelo de lenguaje se equivoca en silencio si trabaja solo con texto. Si tu herramienta ejecuta código gana fiabilidad, pero comprueba igual a mano un dato que ya conozcas —la media, los aprobados— y sigue solo si coincide.

Los cinco indicadores que hay que mirar

Con la matriz cargada, pide cinco indicadores y no 20: son los que separan un problema de aprendizaje de uno de instrumento y de programación.

La **dificultad** de cada ítem es la proporción de alumnado que lo acierta. Por convención psicométrica, no por norma, un valor inferior a 0,30 señala un ítem muy difícil; por encima de 0,90 lo acierta casi todo el mundo: informa poco sobre diferencias, aunque confirma que el contenido está adquirido. La discriminación compara el rendimiento en ese ítem del tercio con mejor puntuación total y del tercio con peor puntuación. Si el superior acierta menos que el inferior, el ítem mide al revés: detrás suele haber un enunciado ambiguo, una clave errónea o dos opciones defendibles.

El comportamiento de los distractores solo existe en opción múltiple y es el más desaprovechado. Uno elegido por media clase es un diagnóstico regalado: nombra el error conceptual dominante del grupo. Uno que no elige nadie no medía nada y conviene sustituirlo, con los criterios de depuración de 3.2.

El **porcentaje de logro por criterio** agrega los ítems según la segunda tabla. Es el único de los cinco que se lleva tal cual a una sesión de evaluación, porque habla el idioma del currículo y no el de la prueba. El quinto es la no respuesta y su posición: una concentración de blancos en los últimos ítems no habla de aprendizaje, habla de tiempo insuficiente, e invalida cualquier conclusión sobre ellos.

FIGURA 3.5

Cinco indicadores, cinco decisiones: cómo se lee la tabla de resultados

	Qué mide	Señal de alerta	Decisión que sugiere
Dificultad del ítem	Proporción de alumnado que acierta el ítem	Por debajo de 0,30 o por encima de 0,90	Reenseñar; si falla el tercio superior, sospecha del ítem
Discriminación	Rendimiento del tercio superior frente al tercio inferior	Nula o negativa: el tercio superior acierta menos	Retirar del cómputo, recalcular y corregir el banco
Distractores	Qué opción falsa se ha llevado los errores	Uno elegido por media clase; otro por nadie	Trabajar el error dominante; sustituir el que nadie elige
Logro por criterio	Ítems agregados según la tabla ítem-criterio	Varios ítems del mismo criterio fallados	Aprendizaje no alcanzado: decisión de enseñanza con fecha
No respuesta y posición	Blancos y su lugar en la prueba	Blancos concentrados en los últimos ítems	Tiempo insuficiente: invalida conclusiones y acorta la prueba

Los cinco indicadores del análisis de una prueba —dificultad, discriminación, distractores, logro por criterio y no respuesta— con lo que mide cada uno, su señal de alerta y la decisión docente que sugiere. Los umbrales de dificultad son convenciones psicométricas, no valores normativos. Elaboración propia.

FIGURA 3.5

Distinguir el fallo del alumnado del fallo del instrumento

Es la lectura que más veces se hace mal. Estas son las combinaciones que se repiten.

Cuando falla casi todo el grupo en un ítem aislado, incluido el tercio superior, sospecha del ítem antes que del grupo. Revísalo con los defectos típicos de 3.2 a la vista: enunciado en negativo, clave discutible, vocabulario no trabajado en clase. Si el ítem está mal, la decisión no es reenseñar: retíralo del cómputo, recalcula y arréglalo en el banco.

Cuando fallan varios ítems del mismo criterio y la dificultad se reparte entre ellos, hay un aprendizaje no alcanzado. La decisión es de enseñanza y hay que situarla en el calendario: reenseñar ahora si el contenido es prerrequisito, o incorporarlo como contenido previo de la

unidad siguiente.

Y cuando falla un grupo identificable y no el resto, comprueba las condiciones. Una prueba a última hora, dos ausencias largas o un cambio de aula bastan para que el resultado no compare lo que crees que compara.

Una petición que funciona bien aquí, con el formato de plantilla parametrizada de 1.3:

PLANTILLA

Tienes la matriz de resultados y la tabla que asocia cada ítem con su criterio de evaluación. Calcula la dificultad y la discriminación de cada ítem, el porcentaje de logro por criterio y la distribución de errores por distractor. Señala los ítems cuya discriminación sea nula o negativa. No propongamos causas ni perfiles de alumnado: devuelve solo indicadores y la lista de ítems que exigen revisión.

La prohibición final no es un adorno: sin ella recibirás explicaciones causales redactadas con toda seguridad y sin apoyo en la tabla.

El caso de Verónica: de la tabla a tres decisiones

Verónica imparte Matemáticas en Educación Primaria. Su grupo, 6.º de Primaria, tiene 24 alumnos y acaba de corregir una prueba de fracciones con 18 ítems repartidos en tres criterios. La media, 5,4, no dice nada por sí sola.

Los ítems 4 y 11 —suma de fracciones con distinto denominador— tienen una dificultad de 0,25 y una discriminación alta: cuestan a todos, pero más a quien menos sabe. Es aprendizaje no alcanzado, y el contenido es prerequisite de la unidad siguiente. El ítem 9 tiene discriminación negativa, porque lo aciertan más quienes peor puntúan; al releerlo, el enunciado admite dos lecturas. Y el criterio de resolución de problemas sale con un 41 % de logro, con el mismo distractor elegido por 14 alumnos, que multiplican donde el problema pide dividir.

Sus tres decisiones quedan escritas en la programación de aula, no en un informe. Repasar la suma con distinto denominador en las dos primeras sesiones de la unidad de decimales, con un cierre de cinco preguntas el jueves siguiente. Retirar el ítem 9 del cómputo, recalcular las notas afectadas y corregir el enunciado en el banco. Y dedicar una sesión a decidir la operación antes de operar, para ver en la prueba siguiente si ese distractor baja de 14 elecciones a menos de cinco.

La herramienta ha hecho el cálculo. Lo demás lo pone Verónica: que esa suma es prerequisite de los decimales, y que 14 alumnos multiplicando revelan un modelo mental equivocado. Eso no está en la matriz.

Dónde se para

Queda el punto de parada. El primer límite es estadístico: con 24 alumnos, una diferencia de medio punto entre dos grupos es ruido, y la herramienta la comentará igual si se lo permites. El segundo es de método. El modelo produce explicaciones causales plausibles —«no dominaban los conocimientos previos»— que son hipótesis tuyas, no hallazgos de la tabla; caben como conjetura, nunca como conclusión.

El tercero es jurídico. El anexo III, punto 3, del Reglamento (UE) 2024/1689 (Reglamento de IA) considera de alto riesgo, entre otros usos, «evaluar los resultados del aprendizaje, también cuando se utilicen para orientar el proceso de aprendizaje». Su aplicación quedó aplazada al 2 de diciembre de 2027 por el Reglamento (UE) 2026/1744 (Ómnibus Digital sobre IA), en vigor desde el 27 de julio de 2026. Es aplazamiento de calendario, no derogación, y el supuesto se desarrolla en 3.3.

Dónde cae tu análisis lo decide el uso. Examinar una prueba ya corregida por ti no clasifica a nadie ni produce una calificación. Pedir que el sistema ordene al alumnado por nivel previsto, anticipe quién suspenderá o asigne itinerarios, sí: eso elabora perfiles de personas.

Deja constancia breve de que el análisis se apoyó en una herramienta y de qué comprobaste tú, con el procedimiento ligero de 4.3. Y guarda la plantilla junto a la matriz vacía: es de lo primero que conviene poner en común en el departamento (4.5).

Aplicación práctica

1. Añade a tu prueba, antes de pasarla, la tabla que asocia cada ítem con su criterio de evaluación. Sin ella el análisis solo te dirá qué preguntas fallaron, no qué aprendizajes faltan.
2. Pide indicadores y prohíbe explícitamente las causas: dificultad, discriminación, distractores, logro por criterio y blancos. Las causas las pones tú, que conoces al grupo.
3. Cierra cada análisis con un máximo de tres decisiones escritas en la programación, cada una con fecha y con la comprobación que te dirá si ha funcionado. Lo que no tenga fecha y comprobación, bórralo.