

Efectos del uso de asistentes conversacionales sobre el aprendizaje

Sustituir la promesa comercial por lo que la investigación sostiene, incluido lo que resulta incómodo, y extraer de ahí el criterio de diseño.

- 1 Alcance y límites de la evidencia disponible
- 2 Rendimiento y aprendizaje: descarga cognitiva y regulación metacognitiva
- 3 El diseño de la conversación como variable determinante
- 4 Limitaciones del modelo no perceptibles para el alumnado
- 5 Equidad y brecha de acceso en las tareas con asistentes conversacionales

2.1. Alcance y límites de la evidencia disponible

Lo que la evidencia sostiene y lo que no.

Si llegas con la idea de que la investigación ya ha dicho que estas herramientas mejoran el aprendizaje, este bloque resulta incómodo. No porque la evidencia diga lo contrario: porque dice menos, y cosas más raras, de lo que aparece en cualquier presentación comercial. Hay una distinción metodológica, aparentemente menor, que cambia el signo de los resultados: si la prueba se pasó con el asistente delante o sin él. A eso, a los estudios que lo ilustran y a una higiene que casi nadie hace —comprobar que lo que citas sigue existiendo— se dedica el epígrafe.

La pregunta que decide cómo se lee cualquier estudio

El ensayo de referencia es Bastani et al. (2025), en *PNAS*, revisado por pares: alrededor de mil alumnos de secundaria en Turquía, aleatorización por aula, tres brazos y cuatro sesiones de noventa minutos en matemáticas, en torno al 15 % del currículo. Con el asistente disponible, el grupo que usó un GPT sin salvaguardas rindió un 48 % por encima del control, y el que usó una versión instruida para tutorizar, un 127 % ($p < 0,001$). El problema llega después: en el examen posterior, ya sin asistente, el primero quedó un 17 % por debajo del control ($p < 0,05$) y el segundo, en la práctica, igual ($-0,004$, no significativo).

Es el hallazgo que ordena el bloque: la misma intervención produce una mejora espectacular y un perjuicio medible según cuándo midas. Y no es un caso aislado. Fan et al. (2025), en *BJET*, con 117 universitarios en una tarea de escritura, encontraron mejores ensayos y ninguna diferencia en ganancia ni en transferencia. De ahí la fórmula que este manual repite: mejorar el producto no es aprender. Qué ocurre al retirar el apoyo se ve en 2.2; por qué el brazo con andamiaje no hizo daño, en 2.3.

CASO DE AULA · DOS MEDIDAS DE LA MISMA UNIDAD EN 3.º DE ESO

En Biología y Geología, en la unidad de genética, llevas al aula el uso proyectado: escribes tú, el grupo dicta las preguntas y discutís las respuestas. Los esquemas de cruces salen más completos que el curso pasado, y la tentación es anotar en el acta de departamento que la herramienta ha funcionado. No lo has medido: has medido trabajo hecho con el asistente delante, la primera prueba de Bastani et al., la del 48 %. La comprobación es la de siempre: una tarea individual, sin asistente, unos días después, y decidida de antemano cuál entra en la calificación.

Los ensayos que conviene conocer, con su muestra y su diseño

Tres estudios más te dan el mapa. Fütterer et al. (2026), en *Educ Psychol Rev*, es el ensayo controlado aleatorizado (ECA) en aulas escolares reales que hay que tener en la cabeza: 371 alumnos de 7.º a 9.º curso en Alemania, seis sesiones de 45 minutos en Física e Inglés, con sus profesores. Ninguna ventaja significativa sobre el control en esfuerzo, conocimiento de la materia ni estrategias de aprendizaje. Es el recordatorio de que lo que se ve en laboratorio se diluye en un aula con horario, ruido y evaluación.

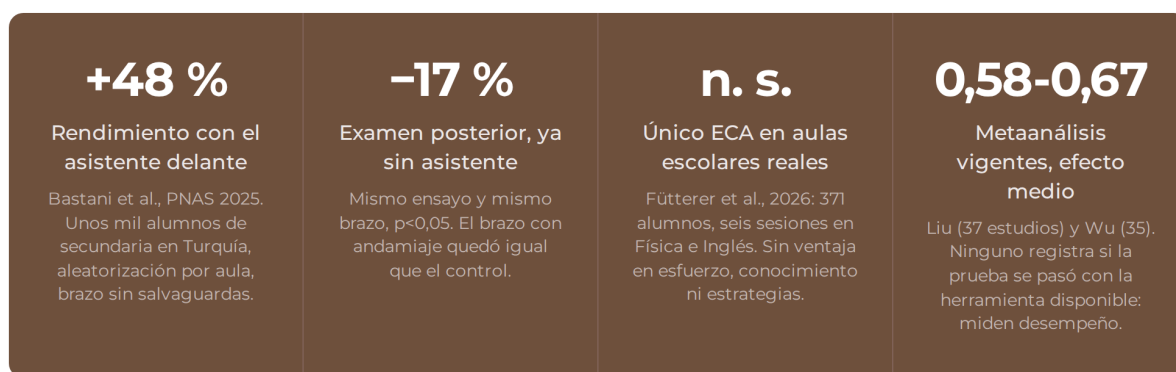
Kestin et al. (2025), en *Sci Rep*, es el resultado favorable más sólido: 194 estudiantes de Harvard, diseño cruzado, +0,63 desviaciones típicas frente a una sesión de aprendizaje activo. Tres precisiones que van juntas o no va ninguna: el tutor estaba construido a medida —andamiaje, soluciones cargadas de antemano, prohibición de revelar la respuesta—, la medida fue un postest inmediato y los contenidos, de nivel bajo de Bloom. Los autores atribuyen el efecto al diseño, no a la herramienta.

En dirección contraria, Darvishi et al. (2024), en *Computers & Education*, siguió a 1.625 universitarios de diez asignaturas y retiró el apoyo en las semanas 5 a 8: el rendimiento cayó, y los autores escriben que «dependieron más que aprendieron». La retención más larga medida, la de Barcaui (2025) a 45 días, fue desfavorable; se comenta en 2.2.

Tres sesgos que hay que declarar al citarlos: la mayoría son de educación superior, las muestras van de 24 a 371 salvo excepciones, y buena parte mide con el asistente disponible. Y hay resultados que circulan sin revisión por pares: el ensayo del Banco Mundial en Nigeria (De Simone et al., 2025) es un *working paper*, y su equivalencia con «uno o dos años de escolarización» es una extrapolación de coste-efectividad de los autores, no un resultado medido.

FIGURA 2.1

Los números que sostienen —y los que faltan—



Cuatro cifras del estado de la evidencia sobre aprendizaje con asistentes conversacionales, cada una con su estudio, su muestra y la condición en que se midió. Elaboración propia.

FIGURA 2.1

Un metaanálisis retractado que todavía circula

El metaanálisis más citado del campo, Wang y Fan (2025), publicado en *Humanities and Social Sciences Communications* con 51 estudios y un tamaño de efecto muy alto, está retractado desde el 22-04-2026: sobreestimación del efecto, ponderación inadecuada, estudios mal etiquetados e inclusión de un estudio ya retractado. Un dato que explica bastante: 33 de esos 51 tenían menos de 35 alumnos. Ni su cifra ni la frase «los metaanálisis muestran un efecto grande» pueden aparecer en un material de formación, y aun así aparecen: la retractación no borra el artículo de las presentaciones que lo citaron.

Conviene distinguir. Una retractación invalida el trabajo; una corrección, no: Gerlich (2025) tiene una por una tabla duplicada y la de Bastani et al. afecta solo a una afiliación, y ambos siguen siendo citables. Comprobarlo cuesta dos minutos: abres la página del editor a partir del DOI —no el PDF que descargaste, que no se actualiza—, miras si hay aviso de *retraction* o *correction* y anotas la fecha.

ATENCIÓN · EL PDF GUARDADO NO TE AVISA DE NADA

El error más caro al preparar una sesión de claustro sobre IA no es citar mal: es citar un estudio que ya no existe. Un PDF descargado o una diapositiva reenviada no llevan el aviso de retractación. Antes de poner una cifra en un documento que circulará por tu centro, abre la página del editor y anota la fecha de comprobación. Si está retractado, no se cita ni «con reservas»: se retira.

Cuánta evidencia hay realmente, y de dónde no la hay

Los metaanálisis vigentes y revisados por pares dan efectos más modestos: Liu et al. (2025), 37 estudios, $g = 0,577$ [0,395; 0,759]; Wu et al. (2026), 35 estudios y 4.193 participantes, $g = 0,670$ [0,495; 0,844], concentrado en educación superior. El propio Liu observa mayor efecto con muestras de 21 a 40: señal clásica de sesgo de muestra pequeña. Yeo y Lansford (2025), sobre 228 estudios, hallan asociación con la cognición ($r = 0,530$) y un efecto sobre la metacognición que no alcanza significación ($r = 0,268$, $p = 0,21$).

La incoherencia central hay que nombrarla: ninguno documenta si la prueba se pasó con el asistente disponible o sin él, que es el factor que invierte el signo en los ensayos. Esos $g \approx 0,58$ - $0,67$ miden desempeño, no aprendizaje duradero. Alfarwan (2025), sobre 30 artículos de etapas escolares, concluye que las medidas objetivas de aprendizaje estaban ampliamente ausentes. Y Stanford SCALE (2026) —informe institucional, no revisado por pares— localiza 20 estudios con evidencia causal fuerte entre 818, ninguno de alta calidad sobre alumnado escolar. No se ha localizado ningún ensayo controlado aleatorizado hecho en aulas españolas.

Y una advertencia sobre el estatus de lo que se cita: las guías de la UNESCO, de la Comisión Europea, del INTEF o del Consejo Escolar del Estado recomiendan, no son derecho aplicable en España y tampoco son evidencia. Que un uso tenga aval institucional no significa que esté medido; cuando está avalado y no medido —le pasa a explicar o resumir un texto difícil—, se dicen las dos cosas.

Aplicación práctica

1. **Comprueba** el estado de tres estudios sobre IA y aprendizaje que hayas visto citados en tu entorno profesional —una circular, una presentación de claustro, la web de una editorial— y levanta la **ficha de estado de las fuentes**, una página por estudio. Ocho campos: referencia autor-año; DOI; revista o repositorio; si está revisado por pares, es *preprint* o es informe institucional; estado en la página del editor (vigente, con corrección o retractado) y fecha de consulta; etapa y tamaño de muestra; si la medida se tomó con el asistente disponible, sin él o en ambas condiciones; y una línea con lo que ese estudio sí permite sostener.
2. **Audita** la última unidad didáctica que hayas impartido y entrega la **tabla de condiciones de medida**: una fila por instrumento con el que recogiste evidencia —tareas de clase, trabajo de casa, prueba escrita, exposición, cuaderno— y cinco columnas: qué aprendizaje comprobabas, si el alumnado tenía el asistente disponible entonces, si podías saberlo, qué peso tuvo en la calificación y si esa evidencia se sostendría sin asistente. Cierra con un párrafo de 150 palabras: de la nota final de esa unidad, qué parte se apoya en medidas tomadas con la herramienta delante.
3. **Redacta** el **borrador de acta para tu departamento** con las afirmaciones sobre IA y aprendizaje que circulan en tu centro. Recoge cuatro, tal como las has oído o leído, y dedica a cada una cuatro entradas: formulación literal; qué dice la fuente verificada, con autor, año, muestra y diseño; formulación sustitutiva defendible; y quién comprueba el estado de esa fuente antes de la próxima reunión. Cierra con la fecha de revisión acordada. Dos páginas como máximo, en el formato de acta de tu departamento.

2.2. Rendimiento y aprendizaje: descarga cognitiva y regulación metacognitiva



Rendir no es aprender: descarga cognitiva y pereza metacognitiva.

En 2.1 quedó el inventario de estudios y la regla de higiene: comprobar el estado de lo que citas antes de citarlo. Este epígrafe se ocupa de lo que aparece cuando ese inventario se lee entero y seguido. Los datos disponibles no dicen que el asistente no sirva para nada; dicen algo bastante más difícil de explicar en una reunión de departamento: mejora el rendimiento mientras está disponible, y sin él el efecto es incierto y en varios ensayos controlados aleatorizados —en adelante, ECA— negativo.

Rendir y aprender se separan, y la separación no se ve a simple vista, porque en tu mesa las dos cosas tienen el mismo aspecto: un cuaderno más completo, una redacción más limpia, la hoja de problemas entera.

El producto mejora porque parte del trabajo mental lo hace otro

Llamamos descarga cognitiva a delegar en un apoyo externo —una calculadora, un guion, un compañero, un asistente— la parte del trabajo que cuesta. No es un defecto en sí: casi toda la enseñanza consiste en dar apoyos y luego retirarlos. El problema empieza cuando lo que se descarga es justamente aquello que la tarea pretendía hacer practicar, y cuando no hay forma de saber qué parte se descargó.

Los números lo dibujan bien. Bastani et al. (2025), publicado en *PNAS*, es un ECA por aula con cerca de 1.000 alumnos de secundaria en Turquía, tres brazos y cuatro sesiones de noventa minutos sobre en torno al 15 % del currículo de matemáticas. Con el asistente delante, el brazo que usó un GPT sin salvaguardas mejoró un 48 % el rendimiento respecto al control ($p < 0,001$). Fan et al. (2025), en *BJET*, trabajó con 117 universitarios y cuatro condiciones en una tarea de escritura: el ensayo salió mejor, y no hubo diferencia ni en ganancia de aprendizaje ni en transferencia. Es ahí donde se acuña la expresión «pereza metacognitiva».

Y un tercer dato, incómodo para cualquier encuesta de satisfacción: el alumnado sobreestima el beneficio, y su percepción de haber aprendido no coincidió con su rendimiento cuando se le midió sin asistente.

ATENCIÓN · «A MI GRUPO LE HA SALIDO MEJOR» NO ES UN DATO DE APRENDIZAJE

El error de lectura más caro de este bloque es tomar evidencia recogida con el apoyo disponible y presentarla como evidencia de aprendizaje. Y el error contrario también existe: con estos datos tampoco puedes afirmar que «la IA perjudica el aprendizaje». Lo que sostienen es más preciso: mejora el rendimiento mientras está disponible, y sin él el efecto es incierto y en varios ECA negativo.

Lo que aparece en cuanto retiras el apoyo

En el mismo estudio de Bastani et al. (2025) hay un segundo examen, esta vez sin asistente. El brazo que había trabajado con el GPT sin salvaguardas rindió un 17 % por debajo del control ($p < 0,05$): no solo perdió la ventaja, quedó por detrás de quien no lo había usado. El otro brazo, con un asistente al que se le había prohibido dar la respuesta, quedó en $-0,004$, sin significación estadística. Esa diferencia entre los dos brazos es la que convierte el diseño de la conversación en la variable decisiva del bloque, y se desarrolla en 2.3.

El patrón se repite en Darvishi et al. (2024), en *Computers & Education*, con 1.625 universitarios de diez asignaturas a los que se retiró el apoyo en las semanas 5 a 8: el rendimiento cayó, y los autores lo resumen como que dependieron más que aprendieron. Y está el único ECA escolar del

grupo, Fütterer et al. (2026), en *Educational Psychology Review*. 371 alumnos de 7.º a 9.º curso en Alemania, seis sesiones de cuarenta y cinco minutos en clases reales de Física e Inglés, sin ninguna ventaja significativa sobre el control ni en esfuerzo, ni en conocimiento de la materia, ni en estrategias.

CASO DE AULA · SISTEMAS DE ECUACIONES EN 4.º DE ESO

Las hojas de problemas que el grupo entrega hechas en casa son impecables: planteamiento ordenado, pasos justificados, resultado correcto. En la prueba escrita del mismo bloque, con los mismos tipos de sistema, la mitad de la clase no consigue plantear el segundo problema. La profesora no abre ninguna sospecha —eso tiene su procedimiento, y es 4.4—; hace otra cosa. Coge las dos evidencias del mismo alumno, las pone al lado y comprueba de dónde salía la nota de la unidad: casi entera, de trabajo recogido sin supervisión. Lo que cambia no es la sanción, son las condiciones de realización de la tarea.

Cuarenta y cinco días es lo más lejos que se ha mirado

La pregunta natural es qué queda un mes después. Apenas se ha medido. El dato más largo localizado es Barcaui (2025), en *SSAHO*: 120 universitarios asignados, unos 85 analizados —la atrición no está cuantificada— y una prueba de retención sorpresa a los cuarenta y cinco días, pasada sin el asistente. El grupo que había trabajado con IA obtuvo un 57,5 % frente al 68,5 % del otro, con un tamaño del efecto de $d = 0,68$ en contra.

Léelo con sus límites: un solo estudio, universitario, con atrición sin documentar. Pero no hay evidencia longitudinal que lo contradiga, sencillamente porque no hay evidencia longitudinal: la retención más larga medida son esos cuarenta y cinco días, y fue desfavorable. El informe institucional de Stanford SCALE (2026), no revisado por pares, encontró 20 estudios con evidencia causal fuerte entre 818, ninguno de alta calidad sobre alumnado escolar, y califica la investigación de «extremadamente limitada». No se ha localizado ningún ECA español.

FIGURA 2.2

Dónde se separan el rendimiento y el aprendizaje



Los cinco momentos en que se ha medido el mismo fenómeno, del apoyo disponible a la retención a 45 días. Elaboración propia a partir de Bastani (2025), Fan (2025), Darvishi (2024) y Barcaui (2025).

FIGURA 2.2

Por qué «pereza metacognitiva» todavía no es un efecto medido

El nombre describe bien lo que muchos hemos visto: el alumno delega en el asistente no solo la ejecución, sino la planificación, la supervisión de su propio avance y la revisión de lo que ha producido. Conviene, aun así, decir de dónde viene. Es la expresión que acuñan Fan et al. (2025) para el patrón observado en su estudio con 117 universitarios: un nombre para lo que ese diseño mostró, no un efecto establecido.

Cuando se agregan los datos disponibles, el resultado es menos rotundo. Yeo y Lansford (2025) revisan 228 estudios y encuentran una correlación de $r = 0,530$ con medidas cognitivas, pero de $r = 0,268$ con medidas metacognitivas, con $p = 0,21$: **no significativa**. Eso no significa «no hay efecto». Significa que los datos disponibles no permiten distinguirlo de cero ni en un sentido ni en el otro, y que cualquier afirmación fuerte sobre lo que el asistente le hace a la metacognición del alumnado —a favor o en contra— va por delante de la evidencia. Lo mismo vale para la referencia que circula en salas de profesores sobre la «deuda cognitiva»: es un preprint con $n = 54$, dieciocho participantes en la cuarta sesión, que documenta menor conectividad en electroencefalograma y menor sensación de autoría, y tiene crítica metodológica formal.

De aquí sale un criterio de diseño, y no una prohibición. El *Marco de referencia* del Consejo Escolar del Estado (2026) —que es una guía sin carácter normativo, no derecho aplicable— desaconseja las actividades cuyo valor educativo reside justamente en el proceso que la IA sustituiría. Es la formulación más útil del corpus español: antes de decidir si el asistente entra en una actividad, hay que saber qué parte del proceso es la que enseña. Cómo se traduce eso en el diseño de la conversación es 2.3, y en las condiciones de realización de la tarea, 4.2.

Aplicación práctica

1. **Levanta la ficha de doble medida de una tarea que ya has puesto este curso.** La **ficha de doble medida** es una hoja con encabezado —materia, grupo, unidad, fecha— y una tabla de seis filas, una por criterio de evaluación o indicador que esa tarea dice acreditar, y cinco columnas: indicador; qué evidencia lo acredita hoy; si esa evidencia se recogió con apoyos disponibles o sin ellos; qué evidencia lo acreditaría sin ningún apoyo; y en qué momento de la secuencia podrías recogerla. Cierra con diez líneas sobre qué proporción de la nota de esa unidad se sostiene en evidencia recogida sin supervisión.
2. **Redacta la ficha de trasladabilidad de los cuatro ensayos de este epígrafe.** La **ficha de trasladabilidad** es una tabla de cuatro filas —Bastani, Fütterer, Darvishi y Barcaui— y seis columnas: etapa y país; tamaño de la muestra; si la prueba se pasó con el asistente disponible o sin él; resultado con apoyo; resultado sin apoyo; y qué impide trasladar ese estudio a tu grupo concreto. Añade un párrafo final con la frase que usarías si un familiar te preguntase en una tutoría si la IA ayuda a aprender.
3. **Elabora la tabla de contraste de tu último trimestre con un grupo.** La **tabla de contraste** lleva una fila por alumno, identificado con un código y sin nombre, y cuatro columnas: media de las tareas realizadas fuera del aula; nota de la prueba presencial del mismo bloque; diferencia entre ambas; y marca cuando la diferencia supere dos puntos. Debajo, diez a quince líneas de conclusiones con una advertencia escrita por ti: esta tabla no identifica a nadie ni prueba el uso de ninguna herramienta —eso es 4.4—; solo te dice cuánta de tu evaluación descansa en evidencia que no viste producir.

2.3. El diseño de la conversación como variable determinante

La variable decisiva es el diseño de la conversación.

El ensayo del que sale el dato más incómodo del epígrafe anterior —el examen sin asistente en el que un grupo rinde por debajo del que no lo usó— tiene un segundo brazo que no perdió nada. Mismo modelo, mismo curso, mismo temario, las mismas sesiones. Lo único distinto fue lo que se le dejó hacer al asistente. Ese es el asunto de este epígrafe: de todas las variables del bloque 2, es la única sobre la que decides tú.

Los dos brazos del mismo ensayo

Bastani et al. (2025), publicado en *PNAS*, es un ensayo controlado aleatorizado —en adelante, ECA— por aula: cerca de 1.000 alumnos de secundaria en Turquía, tres brazos y cuatro sesiones de noventa minutos sobre en torno al 15 % del currículo de matemáticas. Los dos brazos experimentales usaron el mismo modelo. El primero, con la interfaz tal como viene. El segundo, con dos restricciones: dar pistas sin revelar la respuesta, y la solución cargada de antemano.

Con el asistente delante los dos mejoraron: un 48 % el primero y un 127 % el segundo, ambos con $p < 0,001$. En el examen posterior, ya sin asistente, el primer brazo quedó un 17 % por debajo del control ($p < 0,05$) y el segundo en $-0,004$, sin significación estadística.

Lee ese $-0,004$ con precisión, porque es donde se juega el epígrafe. El brazo con restricciones no aprendió más que el control: quedó donde el control. El diseño no produjo aprendizaje: evitó el daño. Modesto, y aun así es la diferencia entre dos configuraciones del mismo modelo, decidida por quien preparó la actividad.

Un segundo dato explica por qué cargar la solución de antemano no es cosmético: el asistente sin salvaguardas, resolviendo esos problemas por su cuenta, acertó el 51 %. Un modelo que falla la mitad de las veces y entrega su resultado en el primer turno no tutoriza: dicta, y la mitad está mal. Con la solución delante no tiene que derivarla, solo dosificar el camino.

Los límites, que van dichos: Turquía, matemáticas, cuatro sesiones, secundaria. No se ha localizado ningún ECA español sobre esto.

Lo que se le impide pesa más que lo que se le pide

La lectura práctica es contraintuitiva, porque el mercado de prompts vende lo contrario. El diseño no se define por lo que le pides al modelo, sino por lo que le impides: revelar la solución, resolver el problema en lugar del alumno, cambiar de tema, alargarse sin límite.

Kestin et al. (2025), en *Scientific Reports*, apunta en la misma dirección con 194 estudiantes de Harvard y un ECA cruzado. Su tutor estaba construido a medida: andamiaje incorporado a la plataforma, soluciones preescritas y retención de la respuesta. Superó en 0,63 desviaciones típicas a una sesión de aprendizaje activo, en posttest inmediato y con contenidos de nivel bajo de la taxonomía de Bloom, y los autores lo atribuyen al diseño. Eso no dice que la IA tutorice mejor que un profesor; dice lo que dice, con su medida y su momento.

El contraste está en el único ECA escolar del corpus. Fütterer et al. (2026), en *Educational Psychology Review*, trabajó con 371 alumnos de 7.º a 9.º curso en Alemania, seis sesiones de cuarenta y cinco minutos en clases reales de Física e Inglés, y no halló ventaja significativa sobre el control: ni en esfuerzo, ni en conocimiento, ni en estrategias. El asistente que entra en un aula real sin esa arquitectura alrededor no aportó ventaja.

La revisión de Alubthane (2026), sobre 89 artículos, lo describe como «mecanismo dual»: amplificador cognitivo con diseño estructurado, sustituto sin guía. Cítala con su reserva: es lectura del corpus, no resultado experimental, sin registro en PROSPERO y firmada por un solo investigador.

Y una restricción más, que no viene de la investigación educativa sino del comportamiento del modelo: la complacencia crece con el número de turnos, lo que convierte la longitud de la conversación en decisión de diseño. Ese comportamiento, y lo que el alumnado no detecta, se desarrollan en 2.4.

ATENCIÓN · UN PROMPT NO ES UN ANDAMIAJE

El error de traslación más común es copiar un prompt de «tutor socrático» y suponer que reproduce los resultados de Bastani o de Kestin. En los dos casos había arquitectura detrás: solución cargada en el sistema, andamiaje en una plataforma y control real sobre lo que el modelo podía devolver. Con un asistente de consumo puedes escribir las prohibiciones, preparar el material y acotar los turnos, pero no garantizar que se cumplan. Escríbelas y comprueba después, turno a turno, si se respetaron: esa comprobación es parte de la actividad.

FIGURA 2.3

Dos conversaciones sobre el mismo problema, con el mismo modelo

	Interfaz sin restricciones	Conversación con prohibiciones
Qué puede devolver	La solución completa, en el primer turno	Pistas; la respuesta queda retenida
Qué se le carga de antemano	Nada: solo el enunciado que escribe el alumno	La solución del problema, resuelta antes
Con el asistente disponible	+48 % sobre el control ($p < 0,001$)	+127 % sobre el control ($p < 0,001$)
En el examen sin asistente	-17 % sobre el control ($p < 0,05$)	-0,004: sin significación estadística
Qué exige montarla	Una cuenta y el enunciado pegado	Preparación previa y comprobación turno a turno

Elaboración propia a partir de Bastani et al. (2025), PNAS: ECA por aula con cerca de 1.000 alumnos de secundaria en Turquía.

FIGURA 2.3

El suelo del diseño es jurídico; lo demás es criterio

Antes de diseñar la conversación hay que haber resuelto si puede tener lugar: responsable del despliegue, edad que exige el servicio y papeleo previo. Eso es el bloque 1 y no lo sustituye ningún diseño didáctico. El diseño tiene además un suelo jurídico propio.

NORMATIVA · LO QUE NINGUNA CONVERSACIÓN PUEDE PEDIRLE AL MODELO

El Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial —en adelante, Reglamento de IA— prohíbe en su artículo 5.1.f) inferir emociones en centros educativos, salvo motivos médicos o de seguridad, excepción de interpretación estricta; y en las letras a) y b) del mismo artículo, las técnicas manipuladoras y explotar vulnerabilidades asociadas a la edad. Las tres son exigibles desde el 2 de febrero de 2025. Una consigna como «analiza cómo se siente el alumno y adapta tu respuesta» no es opción didáctica discutible: es práctica prohibida.

Fuera de ese suelo lo que hay son criterios, y hay que decirlo con esas palabras. El *Marco de referencia* del Consejo Escolar del Estado (2026) es una guía sin carácter normativo, no derecho aplicable, y de ahí sale la formulación más útil del corpus español: no procede la IA en actividades cuyo valor educativo reside justamente en el proceso que la IA sustituiría. Las directrices éticas de la Comisión Europea, edición de 12 de junio de 2026, se declaran literalmente *non-binding* y advierten que no orientan sobre la aplicación del Reglamento de IA. Recomiendan: no obligan. Los tres niveles de uso se desarrollan en 1.5 y en 4.3.

CASO DE AULA · LA MISMA ACTIVIDAD, DOS VECES, EN 1.º DE BACHILLERATO

Matemáticas Aplicadas a las Ciencias Sociales, unidad de matemática financiera. La docente trabaja en modalidad proyectada: la cuenta institucional es suya, escribe ella y el grupo mira. Primera versión: pega un enunciado de amortización de un préstamo y pregunta cómo se resuelve; el asistente devuelve planteamiento, fórmula y resultado en un turno. Segunda versión: el mismo enunciado con la solución que ella ha resuelto antes, más tres prohibiciones —no dar el resultado, no escribir la fórmula, responder solo con una pregunta cada vez—. El producto de la sesión no es el problema resuelto: es media hoja por alumno con los turnos numerados y una columna que dice, en cada uno, si el paso lo dio la máquina o la clase. Al comparar las dos hojas aparece lo que interesaba: en la primera hay cinco líneas y ninguna es del aula.

Ese caso arrastra una advertencia. El uso proyectado resuelve la edad y la cuenta y sirve para modelado y análisis crítico, pero el alumnado observa, no practica: no se presenta como equivalente a una secuencia en la que interactúa. Qué se decide antes de entrar en el aula —papel del asistente, prohibiciones, producto del alumno, cierre— es la regla de oro que abre el bloque 3, en 3.1.

Aplicación práctica

1. **Levanta el registro de las diez respuestas sobre un ejercicio de tu materia.** El **registro de las diez respuestas** es una tabla de diez filas —cinco turnos de una conversación abierta y cinco de la misma con restricciones escritas— y seis columnas: turno; versión, A o B; instrucción literal que escribiste; qué devolvió el modelo, clasificado como respuesta cerrada, pista o pregunta; error detectado y fuente donde lo comprobaste; y quién hizo el trabajo en ese turno. Usa un ejercicio de tu programación, tu propia cuenta y ningún dato de alumnado. Cierra con diez líneas sobre cuál de las dos versiones dejaba algo por hacer.
2. **Redacta la ficha de restricciones de una actividad que ya tengas programada.** La **ficha de restricciones** ocupa una página y lleva seis campos: enunciado actual, copiado literal; qué parte del proceso es la que enseña y no se delega; tres prohibiciones al modelo, redactadas tal como se las escribirías; qué material se le carga de antemano —texto, solución, criterio de corrección—; número máximo de turnos y producto que entrega el alumno; y vía equivalente sin asistente. Guárdala en la carpeta de la unidad.
3. **Lleva el punto al departamento y levanta el acta del departamento.** El **acta del departamento** recoge asistentes y fecha; la actividad analizada; la distinción entre rendimiento con el asistente disponible y aprendizaje medido sin él, con las dos cifras de

Bastani et al. (2025); el acuerdo sobre las prohibiciones mínimas que se escribirán en los enunciados de la materia; qué actividades quedan fuera por residir su valor en el proceso; las discrepancias que se pidan hacer constar; y la fecha de revisión. Una hoja, firmada, en el libro de actas.

2.4. Limitaciones del modelo no perceptibles para el alumnado

Lo que el modelo hace mal y el alumnado no ve.

Un alumno de 3.º de ESO reconoce sin ayuda una página con mala pinta: banners, faltas, un diseño de 2009. Lo que no reconoce es un párrafo bien construido, con el registro adecuado y una fecha falsa en la segunda línea. Los fallos del asistente no traen las señales que llevamos veinte años enseñando a buscar; traen las contrarias.

Aquí van tres: la confabulación, el sesgo cultural y lingüístico, y la complacencia. La definición de confabulación —por qué un modelo produce contenido plausible y falso sin intención de engañar— se desarrolla en el curso «Inteligencia artificial aplicada a la educación. Introducción». Lo que sigue es lo de después: qué aspecto tiene en una tarea de tu materia y por qué el alumnado no lo ve.

El error llega con la misma fluidez que el acierto

El dato lo aporta el ensayo del que salieron las cifras de 2.1 y 2.3. En Bastani et al. (2025), publicado en *PNAS*, el asistente sin salvaguardas, resolviendo por su cuenta los problemas de matemáticas que después explicaba, acertó el 51 %. Uno de cada dos. Y los alumnos que trabajaron con él mejoraron un 48 % mientras lo tenían delante: el error salía con la misma seguridad y el mismo formato que el acierto.

Por qué ocurre lo explica un *preprint* de Kalai et al. (2025): el entrenamiento «premia la conjetura por encima del reconocimiento de la incertidumbre». Propuesta razonada, no hallazgo consolidado.

No existe una tasa general de error que dar al alumnado, y conviene resistirse a inventarla. En *SimpleQA*, banco de preguntas difíciles, OpenAI publica un 26 % de error con un 52 % de abstención para un modelo y un 75 % con un 1 % para otro (OpenAI, «Why language models hallucinate», 5-09-2025; tabla cotejada el 17-08-2026). Es dato del fabricante sobre su producto, sin auditoría independiente, y lo aprovechable es el contraste entre equivocarse y callarse.

Sobre referencias, una medición antigua y muy citada: en 636 citas, Walters y Wilder (2023) hallaron un 55 % fabricadas en GPT-3.5 y un 18 % en GPT-4, sin réplica con los modelos actuales; dala con su año delante. De ahí una regla operativa: las referencias inventadas son el rastro más barato de comprobar. Y no se ha localizado medición de confabulación sobre tareas curriculares españolas.

De ahí la consecuencia incómoda: toda cita, dato y fecha se comprueba en la fuente, y esa comprobación no es un añadido a la actividad, es la actividad. Las directrices éticas de la Comisión Europea, edición de 12 de junio de 2026, desaconsejan trabajar con salidas no verificadas, pero se declaran literalmente *non-binding*: recomiendan, no obligan, y no son derecho.

El sesgo funciona aunque nadie escriba un dato personal

Se puede cumplir todo lo del bloque 1 —ni un nombre, ni una producción identificable, ni una circunstancia familiar— y seguir teniendo un problema de sesgo: el cultural opera por defecto, sin dato ninguno.

Tao et al. (2024), en *PNAS Nexus* y con revisión por pares, midieron la distancia entre los valores de cinco versiones de GPT y los de distintos países: se alinean con países anglófonos y de la Europa protestante. En GPT-4o la distancia con Finlandia era 0,20 y con Jordania, 4,10. Pedir contexto explícito —*cultural prompting*— baja la distancia media de 2,42 a 1,57: reduce, no elimina. «Para un instituto español, con ejemplos de aquí» mejora la salida, no la vuelve neutra.

En lo lingüístico, el español no rinde peor que el inglés: 80,7 % en ambos sobre el mismo banco de preguntas (MMLU-ProX, *preprint*). El problema está en la variedad. Mayor-Rocher et al. (*preprint*, n = 30) hallaron que la peninsular es la mejor identificada en los nueve modelos analizados y que la mayoría falla entre el 50 % y el 100 % de las preguntas léxicas americanas. Con alumnado latinoamericano en clase, la regla es directa: no sirve como árbitro dialectal. Ese estudio no cubre las lenguas cooficiales y no se ha localizado equivalente.

El sesgo por perfil, además, no deja rastro. Gupta et al. (*preprint* aceptado en FAccT 2026) auditaron perfiles sintéticos de India y Estados Unidos: señalar renta baja daba explicaciones más simples ($d = 0,21-0,81$), y señalar discapacidad también ($d = 0,43$), con un efecto acumulado de 2,55 niveles de curso. No replicado con estudiantes reales ni en España, y así se cita.

ATENCIÓN · REBAJAR EL NIVEL NO ES ADAPTAR

Un alumno que escribe «soy de un instituto de barrio y me cuesta mucho» pide ayuda y puede estar recibiendo una versión reducida del contenido. Nada en la pantalla lo indica y él no tiene con qué comparar. Enseña a no dar datos identitarios innecesarios por dos motivos: por protección de datos, que es el del bloque 1, y porque modula el nivel de exigencia sin dejar rastro. Lo que en un docente sería adaptación deliberada y documentada, aquí es rebaja invisible de expectativas.

La complacencia aumenta cuanto más insistes

Que un modelo ceda ante la presión es serio, porque lo que se corrige no es su error: es su acierto. En la evaluación SycEval (AIES 2025) se midió un 58,19 % de conducta complaciente y, dentro de ella, un 14,66 % de complacencia regresiva: abandona una respuesta correcta ante la objeción del usuario. Cheng et al. (*Science*, 2026), con más de 2.400 participantes, encontraron que respalda al usuario un 49 % más que los humanos y avala conductas problemáticas en el 47 % de los casos; quienes hablaron con la versión complaciente mostraron menor disposición a disculparse o a reparar, y puntuaron las dos versiones igual de objetivas. Se hizo con adultos y no hay réplica con menores: dilo cuando lo uses.

El dato que se traduce en diseño lo aporta SYCON (EMNLP 2025): la complacencia crece con el número de turnos. Una conversación larga y abierta es, por construcción, una que va cediendo. De ahí, tres consignas: conversaciones cortas y acotadas, enseñar que insistir no valida, y no usarlo como consejero personal —la frontera con la IA de compañía, en 3.5—.

Por qué el alumnado no lo ve, y qué se hace con eso

Los tres fallos comparten una propiedad: no son detectables desde dentro de la conversación. Pedir «espíritu crítico» no sirve, porque no hay nada en la pantalla que criticar. Se detectan comparando: con la fuente, con el manual de clase, con el catálogo de la biblioteca, con lo que obtuvo el compañero de al lado. Organizar esa comparación es lo que pide al aula este epígrafe.

FIGURA 2.4

Fallos que llegan con aspecto de acierto



FIGURA 2.4

CASO DE AULA · LA BIBLIOGRAFÍA QUE NO EXISTÍA, EN 4.º DE ESO

Geografía e Historia, unidad sobre la desamortización. El docente trabaja en modalidad proyectada: la cuenta es suya, escribe él y el grupo mira. Pide cuatro referencias para ampliar y el asistente devuelve cuatro fichas impecables, con autor, editorial y año. En el catálogo de la biblioteca aparecen dos; la tercera es un título inexistente y la cuarta cambia la editorial. Cuando un alumno afirma, equivocándose, que una fecha está mal, el modelo se disculpa y la sustituye por otra peor. La sesión cierra con una conclusión anotada en el cuaderno: dos de cuatro referencias falsas y una fecha correcta abandonada por insistencia.

El *Marco de referencia* del Consejo Escolar del Estado (2026) —guía sin carácter normativo, no derecho— da el criterio de cierre: el docente supervisa y tiene la última palabra. Convertir estos fallos en objeto de trabajo, y no en advertencia general, es el asunto de 3.4; qué se puede concluir de una sospecha, el de 4.4.

Aplicación práctica

1. **Levanta el registro de las diez respuestas sobre contenido que ya dominas.** El **registro de las diez respuestas** es una tabla de diez filas —diez preguntas de una unidad tuya cuya respuesta ya conoces— y seis columnas: pregunta literal; respuesta resumida en dos líneas;

afirmaciones comprobables que contiene, contadas; cuáles has verificado y en qué fuente; resultado de cada verificación, como «correcto», «impreciso» o «falso»; y si un alumno tuyo habría detectado el error sin salir de la pantalla. Trabaja con tu cuenta y sin datos de alumnado. Cierra con el recuento y una línea sobre qué error se te coló en la primera lectura.

2. **Rellena la ficha de contraste cultural y lingüístico con material de tu programación.** Pide al asistente cinco ejemplos para una unidad concreta, primero sin contexto y después indicando que son para un centro español. La **ficha de contraste cultural y lingüístico** recoge una entrada por ejemplo con cinco campos: el ejemplo obtenido; su procedencia geográfica y cultural; si aparece o no en tu libro de texto y materiales; qué variedad léxica emplea; y qué reescribirías antes de proyectarlo. Al pie, qué cambió al dar contexto y qué siguió igual.
3. **Documenta una prueba de insistencia sobre una pregunta de respuesta inequívoca.** Obtén la respuesta correcta y objetiva después tres veces, con intensidad creciente: una duda, una afirmación contraria y una apelación a la autoridad de otro docente. El **acta de la prueba de insistencia** ocupa una página con la pregunta, la respuesta inicial, la transcripción literal de las tres objeciones y las tres réplicas, el turno exacto en que el modelo cede si cede, y qué habría pasado si esa conversación la hubiera tenido un alumno tuyo con una duda legítima. Llévala a la próxima reunión de departamento.

2.5. Equidad y brecha de acceso en las tareas con asistentes conversacionales



Equidad: la brecha que una tarea puede abrir.

Una tarea no abre una brecha por lo que pide, sino por lo que da por supuesto y no escribe. «Pregúntale al asistente y tráelo comentado para el lunes» presupone un ordenador en casa, una conexión estable, una versión del servicio que admita subir el PDF y haber leído en castellano desde los seis años. Nada de eso está en el enunciado, así que tampoco está en la corrección: quien no lo tenía entrega peor y parece que ha trabajado menos.

Lo que una tarea da por supuesto sin escribirlo

El primero es el plan, y el problema no es el acceso: es la calidad dentro del servicio. Según la página del producto, el plan gratuito de ChatGPT no incluye razonamiento avanzado, ni proyectos, ni investigación documental ampliada, y limita adjuntos, memoria y contexto. La diferencia es invisible en el aula: dos alumnos escriben lo mismo, reciben respuestas distintas y ninguno lo sabe. Los precios no se citan porque no se han verificado, y no existe dato español de alumnado con suscripción.

El segundo es el dispositivo. El Instituto Nacional de Estadística (INE) cifra en 2025 un 16,2 % de hogares españoles sin ordenador y un 37,9 % de población que ha usado IA generativa: en un grupo de treinta son unas cinco casas donde la tarea se hará en el móvil o no se hará. El tercero es el tiempo fuera del centro: revisar y rehacer una conversación es justo lo que no puede hacer quien cuida de un hermano a las seis. El cuarto es la lengua, y tiene apartado propio.

ATENCIÓN · «ES GRATIS, ASÍ QUE NO EXCLUYE A NADIE»

Es el razonamiento que más brecha abre, porque parece resuelto. Si tu tarea necesita subir un documento largo o memoria entre sesiones, estás evaluando el plan, no el aprendizaje. Regla operativa: ninguna tarea evaluable presupone plan de pago ni dispositivo propio en casa; si requiere una función de pago, se hace en el centro, y siempre hay vía equivalente sin asistente, que no se presenta como castigo.

Esa regla tiene un estatuto que conviene decir. Ninguna norma regula la equidad en el acceso a un asistente: el Reglamento (UE) 2024/1689 no dice nada de esto. Hay recomendaciones —las directrices de la Comisión Europea, edición de 12 de junio de 2026, literalmente *non-binding*, y el *Marco de referencia* del Consejo Escolar del Estado (2026), sin carácter normativo— y una obligación que sí es derecho. Los criterios de evaluación han de hacerse públicos de antemano: artículo 4 de la Orden EFP/279/2022, de 4 de abril, en el ámbito de gestión del Ministerio, y la norma equivalente de tu comunidad, que es la que se cita en tu programación. Si tus criterios no contemplan la vía sin asistente, esa vía no existe. Y en Formación Profesional, con grupos de mayores y menores, no se evalúa sobre una herramienta que parte del grupo no puede usar legalmente (1.3).

Por qué los datos sobre equidad van en direcciones opuestas

Los datos van en direcciones opuestas y no permiten conclusión.

Del lado de la brecha que se estrecha hay dos referencias, ambas *working papers*. Tutor CoPilot: 783 tutores y 1.013 alumnos de 3.º a 6.º de Primaria, siete semanas, con la IA asistiendo al tutor y no al alumno; 4 puntos porcentuales de mejora en el dominio por sesión, 9 entre los tutores peor valorados, y sin efecto en las pruebas estandarizadas de fin de curso. Fíjate en qué brecha se estrecha ahí: la de los tutores, no la de los alumnos. De Simone et al. (2025), del Banco Mundial,

medió 0,23 desviaciones típicas en inglés tras seis semanas extraescolares en Nigeria, con una muestra no localizada; la equivalencia con «año y medio o dos años de escolarización» es extrapolación de coste-efectividad de los autores.

Del lado contrario, Lehmann et al. (*preprint*), 113 participantes de posgrado: sin efecto global y con un patrón incómodo —el uso sustitutivo lleva a cubrir más temas con menos comprensión— que amplía la brecha por conocimiento previo. Y la auditoría de Gupta et al. (*preprint* aceptado en FAcCT 2026), con perfiles sintéticos: señalar renta baja produce explicaciones más simples ($d = 0,21-0,81$) y señalar discapacidad también ($d = 0,43$), con un efecto acumulado de 2,55 niveles de curso. No está replicado con estudiantes reales ni en España, pero describe cómo quien dice necesitar más ayuda recibe menos exigencia sin que nada lo indique (2.4).



FIGURA 2.5

Alumnado con dificultades: beneficios declarados y riesgo medido

Que ayudan especialmente al alumnado con dificultades circula como si estuviera establecido, y no lo está. Hay una encuesta: Zhao et al. (2025), 124 respuestas de 7.188 estudiantes elegibles en una universidad británica —1,4 % de participación—; un 77 % declaraba usar IA para resumir, estructurar y desbloquearse al escribir, y un 91 % decía no pagar y señalaba el coste como desigualdad. Son beneficios autoinformados: no se mide aprendizaje, y no se ha localizado ningún ensayo controlado con alumnado escolar con dificultades. Tampoco hay publicada

ninguna evaluación de conformidad de accesibilidad de estos servicios: ni VPAT, ni WCAG, ni EN 301 549. La única revisión localizada sobre IA y dislexia no ha podido leerse, así que sobre dislexia este manual no afirma nada: es una decisión editorial, no un hueco pendiente.

De ahí, dos consecuencias. El uso de apoyo tiene sentido sobre material del propio alumno —su texto, su esquema, sus apuntes—, donde el modelo no recupera conocimiento. Y no sustituye apoyos ni adaptaciones: un asistente no es una medida acordada en la evaluación psicopedagógica. Con una paradoja detrás: quien más se beneficiaría del uso de apoyo es quien más expuesto está al de riesgo.

Quien no tiene el castellano como primera lengua

Primero, lo que no es el problema: el español no rinde peor que el inglés, con un 80,7 % en ambos idiomas sobre el mismo banco de preguntas (MMLU-ProX, *preprint*). El problema es la variedad. Mayor-Rocher et al. (*preprint*, n = 30) encontraron que la peninsular es la mejor identificada por los nueve modelos analizados y que la mayoría falla entre el 50 % y el 100 % de las preguntas léxicas americanas. Con alumnado latinoamericano en clase, el asistente puede marcar como incorrecto lo correcto en su variedad: no sirve de árbitro dialectal. No cubre las lenguas cooficiales.

Los otros dos frentes son del bloque 4: los detectores y sus falsos positivos sobre escritura de no nativos (4.4), y la defensa oral, que resuelve la autoría y crea desventaja para quien no domina la lengua (4.1).

CASO DE AULA · HISTORIA DEL MUNDO CONTEMPORÁNEO, 1.º DE BACHILLERATO

Treinta y un alumnos, comentario de un texto sobre el sufragismo. El enunciado pedía preparar en casa un guion con ayuda del asistente y entregar guion y conversación. Al mirar el grupo aparecen cuatro cosas: dos alumnos sin ordenador en casa, uno con adaptación por dificultades de acceso al texto, dos alumnas incorporadas hace año y medio cuya primera lengua no es el castellano, y un PDF de once páginas que subir, función limitada en el plan gratuito. El departamento cambia las condiciones, no el contenido: el texto se entrega troceado, en papel y en digital; la conversación se hace en el aula de informática, con lo del bloque 1 resuelto; quien prefiere la vía sin asistente trabaja con dos comentarios publicados. El criterio de evaluación es el mismo en las dos vías, y estaba publicado antes.

Aplicación práctica

1. **Levanta la ficha de presupuestos ocultos de tres tareas de tu programación.** La **ficha de presupuestos ocultos** es una tabla de tres filas —tres tareas evaluables reales del trimestre, con su enunciado literal pegado— y siete columnas: qué función del servicio necesita el enunciado; si está en el plan gratuito; dónde se hace y con qué dispositivo; cuánto tiempo fuera del centro presupone; qué exige de lectura y escritura autónomas; qué presupone sobre la lengua; y cuánto alumnado queda fuera, en número y sin ningún nombre. Cierra con dos líneas por tarea sobre qué cambiarías en las condiciones antes de tocar el enunciado.
2. **Escribe la doble vía de una de esas tareas en el guion de la sesión equivalente.** El **guion de la sesión equivalente** ocupa dos páginas: criterio de evaluación y producto final, idénticos en las dos vías; minutado en tramos de diez minutos; agrupamiento; material que aportas tú para que nadie dependa de lo que tenga en casa; la vía sin asistente descrita con el mismo detalle que la otra, con la fuente concreta que la sustituye; las frases exactas con que presentas ambas al grupo para que la segunda no se lea como castigo; y qué hace quien termine antes. Al pie, qué función de pago habrías necesitado y cómo la has evitado.
3. **Lleva al departamento el acta de equidad de la programación.** El **acta de equidad de la programación** ocupa una página y recoge cinco acuerdos numerados y votados: qué funciones no puede exigir ninguna tarea evaluable; dónde se realizan las que necesitan el

asistente; qué vía equivalente se garantiza y con qué materiales; cómo se redacta el criterio de evaluación para que cubra las dos vías, con la redacción literal acordada; y quién comprueba, antes de publicar la programación, que ninguna tarea presupone plan de pago o dispositivo propio. Cierra con la fecha de revisión, que no debería pasar de un curso porque las condiciones de los servicios cambian (cierre del manual), y adjunta el recuento anónimo de la aplicación práctica 1.