

Evaluación del efecto de la inteligencia artificial en el aprendizaje

INFORME COMPLETO

La premisa · Aprender · Elegir · El tutor · La medida · España · Para qué aprender

Índice general

Antes de empezar	1
1 Ha llegado la máquina que resuelve; la que enseña depende de quien la use	2
1.1. El verano en que llegó el tutor	2
1.2. Qué es la premisa vista desde un pupitre	3
1.3. La mitad que ya se cumple	4
1.4. La mitad que falta	6
1.5. Dos velocidades	7
1.6. Qué se deduce y qué no	8
1.7. Veredicto sobre la premisa en educación	10
2 La máquina hace el ejercicio, y el ejercicio era el aprendizaje	12
2.1. Treinta meses en un condado chino	12
2.2. Con sorteo y con vigilancia, lo mismo	14
2.3. Rendimiento no es aprendizaje	16
2.4. El quinto alumno	17
2.5. Qué añade la premisa	18
2.6. ¿Se ve ya en los grandes exámenes?	19
2.7. Quién pierde más	20
2.8. Dos maneras de equivocarse	20
2.9. Por qué estos veredictos	21
3 El alumno elige la respuesta porque no nota lo que pierde	24
3.1. Dos cursos con el tutor a un clic	24
3.2. El mismo patrón en todas partes	26
3.3. No nota lo que pierde	27
3.4. Paga hoy lo que ganaría mañana	28
3.5. Lo que sí cambia la conducta	29
3.6. La objeción de siempre	30

3.7.	Quién decide	31
3.8.	Qué añade la premisa	32
3.9.	Por qué estos veredictos	32
4	Un tutor mejor que cualquier profesor no sube el techo si no sube el tiempo	34
4.1.	La escuela del tutor	34
4.2.	De dónde sale la promesa	35
4.3.	El techo a escala	36
4.4.	Lo que sube el techo es el tiempo	37
4.5.	Dónde sí rinde la IA	39
4.6.	Cómo leer un anuncio	41
4.7.	Qué añade la premisa	42
4.8.	Por qué estos veredictos	43
5	La nota deja de medir: lo que se evalúa fuera de una sala mide a la máquina	45
5.1.	Dos carriles en Sídney	45
5.2.	El termómetro que marca al revés	46
5.3.	La señal que deja de separar	47
5.4.	Por qué no basta buscar a la máquina en el texto, y por qué la sospecha castiga al bueno	48
5.5.	La vuelta a la sala	50
5.6.	Qué añade la premisa	52
5.7.	El cuello de botella, en el profesor	52
5.8.	Por qué estos veredictos	53
6	España certifica sobre todo con lo que ya no mide	56
6.1.	Málaga, septiembre de 2026	56
6.2.	Dónde se certifica en España	57
6.3.	Lo que ya se ve	59
6.4.	Las respuestas de 2026	60
6.5.	Qué respuesta sostienen los datos	61
6.6.	Dónde va el dinero	63
6.7.	Una debilidad propia	63
6.8.	Lo que no se sabe	64
6.9.	Por qué estos veredictos	64

7	Para qué aprender lo que la máquina sabe	67
7.1.	Cuando la máquina se equivoca	67
7.2.	La ironía de Bainbridge	68
7.3.	Con la premisa, la ironía se agudiza	69
7.4.	Qué hace falta saber para verificar	70
7.5.	La generación siguiente	72
7.6.	Ya no hacer falta	73
7.7.	Con quién está el informe	74
7.8.	Por qué estos veredictos	75
8	Qué hacer: un sitio, una hora y alguien que compruebe	77
8.1.	Dos alumnos con la misma máquina	77
8.2.	Si estudia	78
8.3.	Si tiene hijos	80
8.4.	Si enseña	81
8.5.	Si decide sobre un centro o un presupuesto	82
8.6.	Lo que no sirve	83
	Síntesis	84
	Referencias	97

Antes de empezar

Este informe parte de la misma premisa que los otros dos: la que anuncian los grandes laboratorios de inteligencia artificial, sistemas más capaces que cualquier experto en casi todo, funcionando en millones de copias, en pocos años. No la da por cierta. La toma como prueba de esfuerzo, con las fechas y las dudas que recoge el capítulo 1 del Informe I, y se pregunta qué se seguiría de ella. En la escuela, esa pregunta tiene una particularidad que no tiene en otros terrenos. La premisa dice que la máquina resolverá todo problema cuya respuesta se pueda comprobar, y casi todo lo que la escuela pide es de ese tipo: el ejercicio, el problema y, a efectos prácticos, también la redacción escolar. Por eso, en lo que toca a los daños, el mundo de la premisa no hay que imaginarlo en las aulas, porque ya se puede observar. En España, el 54 % de los alumnos de 15 años usa chatbots cada semana para ayudarse a aprender.¹ Los estudios de 2025 y 2026 que usa este informe miden, por eso, ese mundo en su parte escolar, y no una IA todavía floja.

La pregunta es qué se sigue de ahí para aprender, y qué no. El debate público suele ir en dos direcciones que no se tocan. Unos ven llegar un tutor para cada alumno, mejor que cualquier profesor, y dan por hecho que el alumno lo usará. Otros ven una generación que ya no piensa y dan por hecho que la máquina atonta. Este informe sigue otro camino: va del alumno a la herramienta, de la herramienta a las instituciones que tienen que saber quién ha aprendido, y de ellas a la pregunta de fondo, para qué aprender lo que la máquina ya sabe. En cada paso separa lo que se deduce de la premisa, lo que solo se deduce si se añade algo y lo que no se deduce, y dice con quién está y qué le haría cambiar de juicio.

Ha llegado la máquina que resuelve; la que enseña depende de quien la use

54 %

adolescentes de Estados Unidos que han usado chatbots para los deberes; el 10 % hace con ellos todo o casi todo su trabajo escolar²

12 %

universitarios británicos que meten texto generado directamente en un trabajo que se califica; eran el 3 % en 2024³

≈18 %

alumnos estonios del programa nacional que usan cada semana la herramienta hecha para no dar la respuesta (cuenta de este informe con los datos del programa)⁴

1.1 El verano en que llegó el tutor

El 29 de julio de 2025, OpenAI añadió a ChatGPT un «modo estudio». En lugar de dar la respuesta, el modelo pregunta qué se quiere aprender, guía paso a paso y se guarda la solución hasta que el alumno llega a ella. Sus instrucciones se escribieron con especialistas en enseñanza de más de cuarenta instituciones, y es gratis para cualquier usuario registrado. Se puede volver al modo normal en cualquier momento.⁵

Leah Belsky, vicepresidenta de educación de la empresa, admitió al presentarlo: «Si alguien quiere burlar el aprendizaje y, por así decir, sacar las respuestas y tomar el camino fácil, es posible» (traducción del informe).⁵ OpenAI vende el producto y es, por tanto, parte interesada. La frase va contra su interés, y por eso pesa.

En marzo de 2026, la empresa publicó un ensayo con sorteo de su propio modo estudio. Más de trescientos universitarios prepararon exá-

menes de neurociencia y de microeconomía en sesiones cronometradas de cuarenta minutos: unos con buscador y vídeos, otros con el tutor. En microeconomía, los del tutor sacaron en torno a un 15 % más. En neurociencia, la diferencia no se distinguía de la del grupo de comparación, y la empresa lo atribuye a problemas técnicos. Son resultados preliminares, sin artículo publicado.⁶ La propia OpenAI señaló lo que su ensayo no podía ver: lo que de verdad importa, escribió, es si esas ganancias duran.⁶

El verano dejó tres hechos. El tutor que no da la respuesta existe; en condiciones controladas da algún resultado; y está a un clic de la máquina que sí la da. Lo que no dijo es cuál de las dos elige el alumno cuando nadie le pone delante la buena, ni quién puede hacer que elija una u otra. Qué parte de lo que promete la inteligencia artificial llega a un alumno no lo decide solo la técnica.

Por eso este informe empieza por lo que hace el alumno, y cada capítulo contesta la pregunta que deja abierta el anterior.

1.2 Qué es la premisa vista desde un pupitre

La premisa de los tres informes es lo que los grandes laboratorios anuncian para los próximos años. Dario Amodei, que dirige Anthropic, lo llama «IA potente» y lo resume como «un país de genios en un centro de datos».^{7,8} El capítulo 1 del Informe I recoge su definición, las fechas que da cada cual y por qué este trabajo la toma como prueba de esfuerzo y no como apuesta. Sobre una mesa de estudio, sus rasgos se leen así.

- Es más lista que un premio Nobel en la mayoría de los campos: resuelve cualquier ejercicio, problema, comentario de texto o redacción que un profesor pueda mandar, desde primaria hasta el final de la carrera.
- Maneja todo lo que maneja una persona que trabaja a distancia: puede explicar hablando, dibujando o escribiendo, y en cualquier idioma.
- Acepta encargos de días o semanas y los hace sola. Un agente así puede seguir entero un curso a distancia, entregar las tareas de un trimestre o hacer un examen en línea en nombre de otro.
- Funciona en millones de copias a la vez, cada una más rápida que una persona: no se cansa, no tiene horario y repite la explicación tantas

veces como haga falta.

El capítulo 3 del Informe I añade un rasgo que Amodei no incluye y que en la escuela pesa mucho: lo puntero tarda unos ocho meses en funcionar, en una versión más pequeña, en un ordenador doméstico.⁹ Resolver un examen de bachillerato no exige lo puntero. Con la premisa, la máquina que resuelve acaba cabiendo en un portátil y, en versiones más pequeñas, en un móvil sin conexión (la extrapolación es de este informe), y ningún proveedor ve lo que se le pide ni puede cortarlo.

De esos rasgos salen dos promesas distintas para la educación. Una es una máquina que hace los ejercicios. La otra es un tutor que enseña: que sabe qué no ha entendido un alumno, qué le conviene practicar después y cómo sostener su esfuerzo durante meses. La primera sirve para entregar; la segunda, para aprender. El capítulo 1 del Informe I parte la premisa en dos mitades: la de los problemas cuya respuesta se puede comprobar, que ya se está cumpliendo, y la del criterio y el trabajo sostenido, que no tiene fecha. Cada promesa cae en una mitad distinta.

1.3 La mitad que ya se cumple

La premisa dice que la máquina resolverá lo que se puede comprobar. Casi todo lo que la escuela pide es de ese tipo. Un ejercicio de matemáticas tiene solución y un problema de física, un resultado: se comprueban solos, como un teorema o un programa. El capítulo 1 del Informe I cuenta que GPT-6 Astra resuelve el 97,6 % de los problemas del nivel más difícil de FrontierMath, escritos por matemáticos profesionales para que duraran años.¹⁰ Un examen de bachillerato queda muy por debajo. Una redacción o un comentario de texto no tienen juez automático, pero el listón escolar queda muy por debajo del experto y ya se ha superado a ciegas. En una prueba de 2024 en la Universidad de Reading, el 94 % de los exámenes para casa escritos por la máquina pasó sin ser detectado y sacó de media más nota que los de los alumnos (capítulo 5).¹¹ Para los daños, esas tareas se comportan como las comprobables.

De ahí se sigue algo que no ocurre en los otros dos informes. Si la premisa dice que la máquina resolverá lo comprobable, y casi todo lo escolar es comprobable o se comporta como si lo fuera, en la escuela el

mundo de la premisa ya se puede observar. Los datos de 2025 y 2026 que usa este informe no son, por tanto, la foto de una IA todavía floja que se corregirá sola cuando mejore. Son la primera medición de ese mundo en su parte escolar.

La adopción tampoco espera: ya es mayoritaria en la secundaria, en la universidad y en España. En otoño de 2025, el 54 % de los adolescentes de Estados Unidos había usado chatbots para los deberes, y el 10 % hacía con ellos todo o casi todo su trabajo escolar.² En el Reino Unido, el 94 % de los universitarios usa la IA en trabajos que se califican. La proporción que mete texto generado directamente en uno de ellos ha pasado del 3 % en 2024 al 12 % en 2026.³

En España, según PISA 2025, el examen internacional de la OCDE a alumnos de 15 años, el 54 % usa chatbots para ayudarse a aprender al menos una vez por semana, más que el 46 % de la media de los países de la organización.¹

Para qué se usa lo muestra Anthropic, que analizó en 2025 más de medio millón de conversaciones de cuentas universitarias con su modelo, Claude. En torno al 47 % eran «directas»: el alumno pedía la solución o el trabajo hecho, con poca implicación propia.¹² Son datos de una sola empresa, que es parte interesada, y de la primavera de 2025, pero no la favorecen.

Lo que la premisa añade a lo que ya hay agrava los daños.

- **Sin errores que delaten.** Hoy, una respuesta copiada todavía se equivoca a veces, y ese error es la última huella que avisa al profesor, y al propio alumno, de que algo no cuadra. Con la premisa, la respuesta copiada es correcta y no hay nada que revisar.
- **Agentes.** Hacer un curso a distancia o un examen en línea deja de exigir siquiera copiar y pegar: lo hace entero un agente, en nombre del alumno. Esto sí depende de la otra mitad de la premisa, la del trabajo sostenido, y golpea sobre todo a la certificación a distancia.
- **Modelos en casa.** El modo tutor pasa a ser un ajuste que cualquiera quita, y el proveedor deja de poder hacer de guardián, por la razón que da el capítulo 3 del Informe I: lo que corre en casa no lo ve nadie.

Salvo el agente que trabaja semanas solo, nada de esto necesita la otra mitad. Cada ejercicio es una tarea corta que se pega de una en una, y

resolverla no exige decidir qué merece la pena estudiar.

1.4 La mitad que falta

La otra promesa es la que más se repite: un tutor para cada niño, mejor que cualquier profesor. Exige dos cosas que la primera no exige.

La primera es conducta. Un tutor solo enseña a quien lo usa, y usarlo cuesta: hay que sentarse, equivocarse y preguntar. Philip Oreopoulos, economista de la educación de la Universidad de Toronto, y un coautor dieron durante dos cursos el tutor de Khan Academy a alumnos de Tennessee con bajo rendimiento, dentro de las horas de refuerzo y con el profesor en el aula. Casi todos lo probaron, y casi nadie le preguntó cuando se equivocaba. Los autores concluyen que el límite no es el tutor, sino que el alumno lo use; el capítulo 3 cuenta el ensayo entero.¹³ El ensayo de OpenAI apunta a lo mismo por cómo se hizo: sus universitarios estudiaban en sesiones cronometradas, con el tutor delante y sin tener que elegirlo frente al modo normal.

David Strömberg, de la Universidad de Estocolmo, y sus coautores de Hong Kong siguieron durante treinta meses a los alumnos de secundaria de un condado chino, el seguimiento más largo de los que usa este informe. Llegan a la misma conclusión: los alumnos prefieren las herramientas generalistas, y el problema es de incentivos, no de técnica.¹⁴ Cuando alguien protege el tiempo y vigila el aula, la misma plataforma de Khan Academy, incluso sin IA generativa, da mucho más (capítulo 4). Para el alumno que quiere usarlo, el tutor que ya existe basta. Lo que hoy limita es la conducta, y un tutor más capaz cambia lo que se ofrece al alumno, no lo que el alumno quiere hacer esta tarde.

La segunda es criterio, y pesa en la promesa grande: la de un tutor que decide por el alumno qué no ha entendido, si conviene insistir o cambiar de camino, cuándo dejarle equivocarse y cómo sostener su esfuerzo durante meses. Nada de eso tiene un juez automático, y es la mitad de la premisa que no tiene fecha. En educación no hay ningún dato sobre ella. El indicio más cercano viene de la investigación: en una de las pocas medidas que hay, hecha por METR, una organización independiente que evalúa estos sistemas, los agentes internos de los grandes laboratorios acertaron

entre febrero y marzo de 2026 el 59 % de sus decisiones de juicio, como qué camino tomar o qué resultado creerse. Los investigadores humanos acertaron el 90 %.¹⁵

1.5 Dos velocidades

Arvind Narayanan y Sayash Kapoor, informáticos de Princeton, sostienen que la IA transforma despacio porque su uso tiene que pasar por organizaciones que se adaptan a su propio ritmo, y que eso vale aunque los métodos se aceleren sin límite.¹⁶ El capítulo 2 del Informe I concluye que aciertan para los beneficios y se equivocan para los daños que dependen de una sola decisión. La escuela es el caso más limpio de esa asimetría.

El daño va deprisa porque es una decisión individual. Pegar los deberes en un chatbot no exige que nadie apruebe nada, compre nada ni cambie un horario. En el condado chino que estudian Strömberg y sus coautores, casi ningún alumno usaba IA generativa en septiembre de 2022, y en junio de 2025 la usaba cerca del 80 %. Las herramientas más usadas eran generalistas, sin ninguna función educativa.¹⁴ Ese es el nivel al que llega el uso que nadie organiza.

El beneficio va despacio porque pasa por instituciones, y Estonia lo muestra dentro de una misma población. En agosto de 2025 dio un tutor a los dos primeros cursos de su secundaria superior. La aplicación, hecha por investigadores estonios con OpenAI, está pensada para guiar el razonamiento en lugar de dar respuestas hechas. Antes del programa, entre el 64 y el 90 % de los alumnos estonios ya usaba herramientas comerciales para sus tareas. En marzo de 2026 habían activado la cuenta 7.700 de los unos 20.000 alumnos a los que iba destinada. El 47 % de ellos la usaba cada semana: en torno al 18 % del total, según la cuenta de este informe.⁴ Tras siete meses de programa nacional, la herramienta pensada para enseñar llega a menos de uno de cada cinco; la que da la respuesta la usaba la mayoría antes de empezar.

Corea del Sur muestra que ir deprisa no lo arregla. Implantó libros de texto con IA en marzo de 2025, la adopción se hundió en un semestre entre quejas de familias y profesores, y en agosto, cinco meses después, el Parlamento les retiró el rango de libro de texto.¹⁷ La institución fue

rápida al implantar y al retirar, y el beneficio tampoco llegó, porque lo que falló fue el uso.

La premisa no acerca los dos ritmos, y hace más hondo el rápido. Con una máquina mejor, la adopción no tiene por qué ir más deprisa, porque la decisión sigue siendo de cada alumno y de cada tarde. Pero cada uso pesa más: lo copiado no tiene errores, un agente entrega tareas enteras y lo que corre en casa no lo corta nadie. La institución que tendría que organizar el buen uso no se adapta antes porque el modelo sea mejor: sus tiempos son los de sus claustros, sus presupuestos y sus leyes. Lo que crece es el daño que llega antes de que reaccione. La pregunta útil no es, por tanto, cuándo llegará el tutor perfecto, sino quién organiza el tiempo en que se usa.

1.6 Qué se deduce y qué no

Lo de la primera columna de la tabla solo necesita la mitad de la premisa que ya se cumple y la conducta que ya se observa. Lo de la segunda necesita algo que la premisa no trae, casi siempre la presencia o la ausencia de una estructura adulta alrededor del alumno. La tercera recoge afirmaciones que circulan y que ni la premisa ni los datos sostienen.

Tabla 1 Qué se deduce de la premisa para aprender, y qué no

Se deduce de la premisa	Se deduce solo si se añade algo	No se deduce
Quien usa la IA sin estructura la usa sobre todo para acabar antes, y aprende menos sin notarlo (capítulos 2 y 3)	Que el tutor de IA mejore al alumno medio: solo si alguien garantiza el tiempo y la presencia, como en la India o en Nigeria (capítulo 4)	Que la escuela sobre: su tiempo protegido pasa a ser lo escaso (capítulo 4)

continúa

Se deduce de la premisa	Se deduce solo si se añade algo	No se deduce
La nota de lo que se hace fuera de la vista deja de medir y de separar a unos alumnos de otros (capítulo 5)	Que se abra la distancia entre alumnos: solo donde la estructura depende de la familia (capítulo 3)	Que un tutor sobrehumano, por sí solo, dé los saltos de aprendizaje que prometen sus anuncios (capítulo 4)
Certificar vuelve a la sala, cuesta horas de profesor y, con modelos que funcionan sin conexión, exige detectar dispositivos (capítulos 5 y 6)	Que una generación pierda la capacidad de comprobar a la máquina: solo si no se reponen la práctica y el examen presencial (capítulo 7)	Que una generación quede «incapaz de pensar» (capítulo 2)
El saber propio pasa de servir para producir a servir para comprobar lo que produce la máquina (capítulo 7)	Que crezca la brecha por renta: solo donde el uso intensivo se concentre en los hogares pobres, como en Estados Unidos (capítulos 2 y 3)	Que prohibir la IA en el aula resuelva un daño que ocurre en casa (capítulo 6)

La primera columna no depende de las fechas de los laboratorios. Salvo la última fila, que solo importa donde la máquina siga fallando, nada de lo que contiene necesita la mitad del criterio, y por eso los capítulos 2, 5 y 6 pueden dar veredictos altos hasta 2031 sin apostar por el calendario de nadie.

La segunda es donde se decide casi todo. Cada «solo si» es una estructura que alguien puede poner o quitar: un horario protegido, un adulto en la sala, un examen presencial.

La tercera reúne tres errores con nombre. La escuela montada sobre el tutor, como Alpha School en Estados Unidos, cara y sin datos publicados, da por hecho que el tutor vuelve innecesaria la escuela (capítulo 4).¹⁸ El alarmismo que da a una generación por incapaz de pensar descansa en un estudio de electroencefalograma con 18 personas en su sesión decisiva (capítulo 2).¹⁹ Y la prohibición en el aula, como la que plantea Cataluña para menores de 14 años, no llega a los deberes que se hacen en casa (capítulo 6).²⁰ Los tres miran la máquina y no lo que el alumno hace con

ella ni quién tiene alrededor.

1.7 Veredicto sobre la premisa en educación

La posición de este informe es que, en la escuela y para los daños, la premisa no es una apuesta, sino una consecuencia que ya se puede observar. La premisa dice que la máquina resolverá lo comprobable; casi todo lo que la escuela manda es comprobable o se comporta como si lo fuera; y la mayoría de los alumnos ya usa la máquina para hacerlo. Los daños de los capítulos 2, 5 y 6 no necesitan nada más. Lo que la premisa añade —respuestas sin errores, agentes, modelos en casa— no los crea: los agrava. Los beneficios del capítulo 4 necesitan una conducta que la técnica no cambia y, en su versión grande, la otra mitad, la del criterio, que no tiene fecha.

Con Narayanan y Kapoor en los beneficios, que pasan por instituciones y van despacio; contra ellos en los daños, que dependen de una decisión de cada alumno y van deprisa.

La posición tiene un límite. La Education Endowment Foundation, la fundación que evalúa con ensayos lo que funciona en las escuelas de Inglaterra, reconoció en junio de 2026, por boca de su consejera delegada, Becky Francis, que las pruebas sólidas sobre el efecto real de estas herramientas en el aprendizaje apenas han empezado, sobre todo por debajo de los 16 años.²¹

QUÉ CAMBIARÍA ESTE JUICIO

Hacia el optimismo, un ensayo grande en el que un tutor de IA de uso voluntario, sin un adulto que organice el tiempo, mejore de forma duradera al alumno medio. No hay ninguno en marcha con ese diseño. Los pendientes dirán si un tutor bien hecho enseña, no si se usa sin nadie que obligue: el segundo ensayo de la plataforma británica Eedi con Google DeepMind se hace en el instituto y un humano revisa cada mensaje del tutor; el fondo de la Education Endowment Foundation estudia qué pasa cuando se delega el esfuerzo en la máquina, con primeros resultados en 2027.^{21, 22} En sentido contrario, lo reforzaría que el uso semanal de la herramienta estonia siguiera por debajo del 20 % de los alumnos a los que va destinada cuando el programa se amplíe, en el curso 2026-2027.⁴

— *Si la máquina ya resuelve todo lo que la escuela pide, la primera pregunta es qué le pasa a quien la usa para resolverlo.*

La máquina hace el ejercicio, y el ejercicio era el aprendizaje

**+18 % y
—20 %**

nota de los deberes y del examen mensual a libro cerrado de 26.811 alumnos chinos, seis meses después de empezar a usar IA¹⁴

—17 %

examen sin IA de los alumnos turcos que practicaron con un chat como ChatGPT; no notaron que rendían peor²³

4 de cada 5

usuarios chinos que acaban los deberes más deprisa que el más rápido de los que no usan IA¹⁴

2.1 Treinta meses en un condado chino

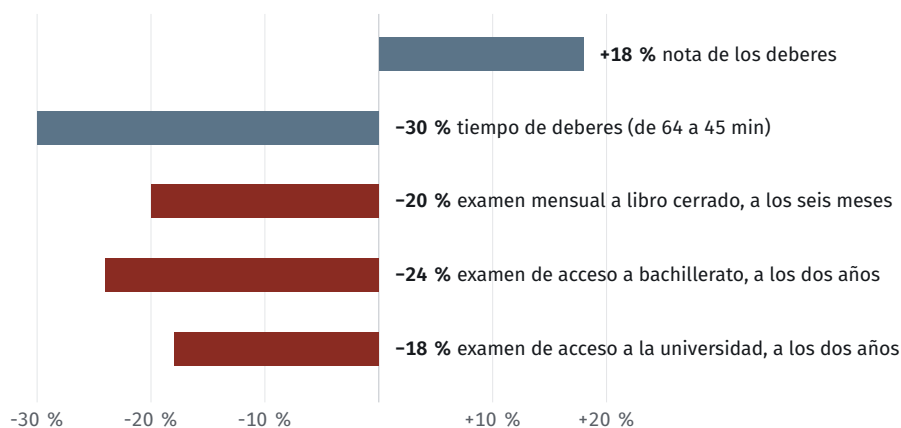
Entre septiembre de 2022 y junio de 2025, las escuelas de un condado del centro de China guardaron lo que hacían sus 26.811 alumnos de 7.º a 12.º. Quedó registrado cuánto tardaba cada uno en los deberes de cada semana y qué nota sacaba, cómo le iba en los exámenes mensuales a libro cerrado de nueve asignaturas y, al final de cada etapa, en los exámenes de acceso al bachillerato y a la universidad. Tres economistas de las universidades de Estocolmo y de Hong Kong, Strömberg, Lei y Wu, cruzaron esos registros con una encuesta que repartieron los profesores: cada alumno miró en sus aplicaciones la fecha en que se había dado de alta. Al empezar el periodo casi nadie usaba IA generativa. Al acabar la usaba cerca del 80 %, sobre todo dos asistentes generalistas, Doubao y DeepSeek, sin ninguna función educativa.¹⁴

Como cada alumno empezó en un momento distinto, se puede hacer lo que los economistas llaman diferencias en diferencias: comparar cómo

cambian las notas de quien empieza a usar la IA con cómo cambian, esos mismos meses, las de quien todavía no ha empezado. Antes de adoptarla, los dos grupos tenían notas casi idénticas y evolucionaban de forma parecida. No hubo sorteo, así que quien empieza antes puede ser distinto de quien espera; por eso importa que la misma pauta aparezca después en diseños que sí lo tienen.

A los seis meses de empezar, el tiempo de deberes pasa de 64 a 45 minutos. La nota de esos deberes sube un 18 %. En el examen mensual, hecho sin la herramienta, la nota baja un 20 %. Los dos porcentajes se calculan sobre la nota media que esos alumnos tenían antes de empezar. En los exámenes de acceso, que recogen lo aprendido en varios cursos, la pérdida tarda unos dos años en completarse. Llega al 24 % en el de bachillerato y al 18 % en el de universidad. Los autores advierten que los estudios de pocas semanas «pueden subestimar sistemáticamente» el coste a largo plazo.¹⁴

Figura 1 Lo que cambia cuando un alumno chino empieza a usar IA generativa



Strömberg, Lei y Wu (2026), 26.811 alumnos de 7.º a 12.º de un condado del centro de China, 2022-2025. Porcentajes sobre la media anterior a la adopción. Documento de trabajo sin revisión por pares.

Casi todo lo que se manda para casa tiene una respuesta que el profesor puede comprobar con una solución o una rúbrica, y pertenece por eso a la mitad de la premisa que, como se vio en el capítulo 1, ya se cumple. Para que haya daño no hace falta una IA mejor que la de ese condado: basta

con que saque la nota de los deberes. La sacaba. Cuatro de cada cinco usuarios terminan los deberes en menos de 50 minutos, más deprisa que el más rápido de quienes no usan IA, con notas que los autores describen como «a la altura de lo que hacen las herramientas de IA corrientes» (traducción del informe).¹⁴ En los deberes, ese condado es ya el mundo de la premisa.

La posición de este informe es la del título de este capítulo: la máquina hace el ejercicio, y el ejercicio era el aprendizaje. Los deberes no enseñaban por la nota que sacaban, sino por el rato de pensar que obligaban a pasar, y quien los entrega hechos por la máquina se queda con la nota y sin el rato. Los límites del estudio son reales: es un documento de trabajo sin revisión por pares, de un solo condado, y el uso de la IA lo declaran los propios alumnos. Lo que lo sostiene es que la misma pauta aparece con otros dos métodos, en otros dos países.

2.2 Con sorteo y con vigilancia, lo mismo

En el otoño de 2023, el equipo de Hamsa Bastani, de la escuela de negocios Wharton de la Universidad de Pensilvania, repartió por sorteo las aulas de un instituto turco entre tres maneras de practicar matemáticas. Eran cerca de mil alumnos de 9.º a 11.º. Unos practicaron con el libro y sus apuntes. Otros, con una interfaz como la de ChatGPT, basada en GPT-4. Y otros, con un tutor montado sobre el mismo modelo, con las soluciones correctas cargadas y la orden de dar pistas en vez de respuestas. Tras cada una de las cuatro sesiones había un examen sin ayuda.²³

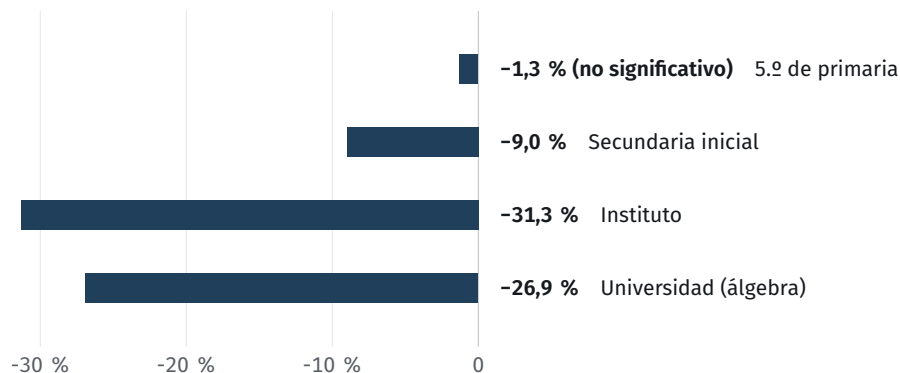
Con el chat libre, los alumnos acertaron en la práctica un 48 % más que el grupo del libro; en el examen, ya sin él, un 17 % menos. Con el tutor, la práctica subió un 127 % y el examen quedó igual que el del grupo del libro: el daño desapareció, pero no hubo mejora. Las conversaciones explican la diferencia. Con el chat libre, lo que los alumnos pedían sobre todo era la respuesta, y la copiaban.²³ Es un solo centro, con cuatro sesiones y un examen inmediato, pero el sorteo quita la duda que deja el estudio chino: aquí nadie eligió la herramienta.

La tercera prueba no mira a nadie usar la IA: mira su huella. ALEKS, una plataforma de matemáticas de la editorial McGraw Hill que usan

más de cuatro millones de alumnos al año, desde 5.º de primaria hasta la universidad, guarda desde 2015 cuánto tarda cada uno en cada problema. Sus problemas son de dos clases: los de enunciado, que se pueden copiar en un chatbot, y los de gráficas, que hay que resolver manipulando la pantalla. Un equipo de la Universidad de California en Irvine y de la propia empresa comparó unos con otros antes y después de ChatGPT.²⁴

Desde 2023, el tiempo dedicado a los problemas de enunciado cae respecto al de los de gráficas, trimestre a trimestre. La caída es casi nula en 5.º de primaria y del 9 % en los primeros cursos de secundaria. En el instituto llega al 31 %, y en la universidad, al 27 %.²⁴

Figura 2 Caída del tiempo por problema en los ejercicios que se pueden pegar en un chatbot, por nivel



Rismanchian y otros (2026), plataforma ALEKS, 3,2 millones de interacciones, 2015-2025. Variación acumulada en once trimestres respecto a los problemas de gráficas. Preprint sin revisión por pares.

Lo decisivo son las pruebas de nivel, unas sin vigilancia y otras con alguien delante. Sin vigilancia, los alumnos también van más deprisa en los problemas de enunciado; con vigilancia, el tiempo vuelve a ser el de siempre. Y en las vigiladas aciertan menos esos problemas: las probabilidades de acertar, contadas como aciertos frente a fallos, bajan un 25 %, mientras en las no vigiladas suben un 85 %. La bajada está al límite de lo que se distingue del azar: si no hubiera ningún efecto, una diferencia así saldría menos de cuatro veces de cada cien. Y los autores no ven la pantalla de nadie: miden una huella en toda la población, no lo que hizo cada alumno.²⁴ Pero en los problemas de gráficas, que un chatbot no

resuelve por el alumno, no cae nada, y eso descarta la pandemia, una peor preparación o la distracción como explicación. Uno de los autores trabaja para McGraw Hill, y el resultado no favorece a su empresa: muestra que la plataforma da por aprendido lo que el alumno ha copiado.²⁴

Tres diseños, tres países y una misma dirección. La síntesis que la OCDE publicó en enero de 2026 sobre la IA generativa en la educación lo resume en una frase: «hacer bien una tarea con IA generativa no lleva automáticamente a aprender». Con las herramientas generalistas, añade, la ventaja en el trabajo entregado «desaparece, y a veces se invierte, en los exámenes, cuando se retira el acceso» (traducción del informe).²⁵

2.3 Rendimiento no es aprendizaje

La psicología cognitiva tenía nombre para esta pauta antes de que existiera ChatGPT. En 2015, los psicólogos Nicholas Soderstrom y Robert Bjork revisaron décadas de experimentos y resumieron una distinción: lo que alguien consigue mientras practica, su rendimiento, predice mal lo que retiene después, su aprendizaje. Lo que facilita la práctica a menudo empeora lo que queda. Y las personas juzgan cuánto han aprendido por lo fácil que les ha resultado, así que se engañan.²⁶

Las tres pruebas anteriores son esa distinción medida con IA: sube la nota de lo que se hace con la máquina y baja la de lo que se hace sin ella. La tercera parte, el autoengaño, también aparece. Los alumnos turcos del chat libre «no percibieron» que rendían peor ni que aprendían menos, y los del tutor creyeron haberlo hecho mejor aunque su examen no se movió; eso se midió en un ensayo de cuatro sesiones.²³ En el condado chino el examen mensual sí mostraba la caída, y cuatro de cada cinco siguieron igual durante meses: no notarlo explica una parte, y el capítulo 3 busca el resto.

Un ensayo que parece decir lo contrario precisa el mecanismo. Lira y otros, un equipo de Pensilvania, Harvard y Microsoft, pusieron a adultos a aprender a escribir cartas de presentación. Quienes practicaron con IA se esforzaron menos y, aun así, escribieron mejor después sin ella; bastaba ver la carta corregida por la máquina para mejorar.²⁷ Eran adultos que querían aprender, en una tarea corta y con la prueba justo después: el

ensayo demuestra que el mecanismo existe, no que funcione con un alumno de 14 años en casa. Lo que enseñó fue ver un buen ejemplo y tener que producir después, sin ayuda, algo parecido. El alumno que pega la solución de los deberes también ve el ejemplo, pero nadie le pide luego reproducirlo: lo entrega. Como escriben los autores del estudio de ALEKS, quien copia el enunciado en un chatbot y pega la solución se salta la práctica de recuperar de memoria, que es lo que convierte lo visto en recuerdo.²⁴

Justin Reich, que dirige en el Instituto Tecnológico de Massachusetts un laboratorio dedicado a cómo se enseña, pone el problema en el mismo sitio. Si los profesores aportan el contenido y el contexto, y los alumnos «ya no hacen los ejercicios» que lo contienen, «es un problema serio» (traducción del informe).²⁸

2.4 El quinto alumno

El estudio chino también dice quién se libra de la pérdida. Uno de cada cinco usuarios sigue dedicando a los deberes el mismo tiempo que quienes no usan IA. Saca mejor nota en ellos y, según el resumen oficial del estudio, sufre «pérdidas pequeñas» en los exámenes: pierde poco.¹⁴ Los autores añaden una cautela. Quien mantiene el tiempo puede ser, ya de entrada, un alumno más motivado, y obligar a los demás a dedicar más tiempo no garantiza que recuperen lo perdido.¹⁴

Con esa cautela, la lectura más compatible con los datos es que el daño viene de saltarse la práctica. La herramienta es la misma para los cinco; lo que cambia es si el rato de trabajo sigue ahí. El tutor de Turquía apunta a lo mismo desde el otro lado: como obligaba al alumno a hacer los pasos, evitó el daño, pero tampoco produjo mejora.²³ Las pruebas de que la IA puede enseñar, como el ensayo de Lira o las herramientas diseñadas con intención pedagógica, que según la OCDE tienden a mejorar el aprendizaje de forma sostenida, vienen de condiciones más controladas; el capítulo 4 mira cuáles.²⁵ Si esta lectura es correcta, lo que decide el resultado es el tiempo de trabajo del alumno, no qué herramienta usa ni si está prohibida, y lo perdido debería poder repararse reponiendo la práctica. Nadie lo ha medido, y los propios autores dudan de que baste

con obligar a dedicar más tiempo.

2.5 Qué añade la premisa

Lo que la premisa añade a ese condado no reduce el daño del uso libre: lo agrava, por dos razones, y además lo hace menos visible.

La primera es que desaparece el único aviso que quedaba. En los problemas del ensayo turco, GPT-4 se equivocaba en casi la mitad cuando no tenía las soluciones cargadas: el 49 %, con un modelo de 2023.²³ La cifra es de un modelo viejo y no sirve para decir que la IA falla; sirve para ver qué se pierde cuando deja de fallar. En 2023, el alumno que copiaba entregaba a veces respuestas erróneas, y el error era una señal para él y para su profesor. Con la premisa, la respuesta copiada es correcta y no queda nada que revisar.

La segunda es la que ya apuntó el capítulo 1: como modelo que funciona en un móvil sin conexión, la máquina de la premisa queda fuera del alcance de cualquier proveedor que pudiera obligarla a enseñar en vez de resolver. El modo tutor que hoy ofrecen los laboratorios pasa a ser un ajuste que cualquiera quita, y lo que decide el resultado es la configuración que elige el alumno.

Además, el daño se nota menos. En Turquía, la mayor subida en la práctica, el 127 %, fue la del grupo del tutor, y su examen no se movió: cuanto más acierta la máquina, menos dice la nota de la práctica sobre lo aprendido.²³ En el condado chino, entre los usuarios, una nota más alta en los deberes llega a anunciar un examen peor; el capítulo 5 lo desarrolla.¹⁴

La posición de este informe es que, con la premisa, el daño de este capítulo crece en vez de menguar. La premisa mejora lo que la máquina ofrece, pero no cambia lo que el alumno elige cuando nadie organiza su tiempo, y abarata todavía más la opción de acabar antes. Cambiaría esta posición que, en un mismo diseño, quienes empiezan con modelos más capaces pierdan menos que quienes empezaron con los de 2023, o que un modo tutor activado por defecto siguiera en uso sin un adulto que lo sostenga.

2.6 ¿Se ve ya en los grandes exámenes?

PISA 2025, publicado en septiembre de 2026, dio las medias más bajas de la historia de la OCDE en las tres materias.²⁹ La posición de este informe es que la IA no explica esa caída. La lectura baja desde su máximo de 2012, las ciencias desde hace una década y las matemáticas desde 2018, años antes de ChatGPT.²⁹ El recuadro en que la OCDE analiza por qué cae la lectura no nombra ni una vez la inteligencia artificial, los chatbots ni ChatGPT. Las fuerzas que señala son otras: la capacidad o la voluntad de mantener la atención, el clima de las aulas, un gusto por leer que cae desde 2009, también entre los padres, y la pandemia como acelerador. Su conclusión es que no hay «una explicación simple».³⁰

Que PISA 2025 no lo muestre tampoco prueba que no esté pasando. Los alumnos que examinó tenían 15 años y nacieron en 2009 o 2010. Cuando apareció ChatGPT, en noviembre de 2022, tenían 12 o 13; la cuenta es de este informe. Tuvieron la IA a mano solo en una parte de la secundaria, y con una adopción todavía parcial: la usaba casi a diario uno de cada cinco.³¹ La OCDE encuentra que quienes no usan la IA para tareas escolares concretas sacan más nota que quienes la usan, pero advierte que esa relación puede reflejar «una mezcla compleja de quién adopta la IA y cómo la usa».^{29,31} Con una encuesta no se puede separar la causa de la selección.

El tiempo de deberes que declaran los alumnos de la OCDE cayó un 11 % entre 2022 y 2025, sobre todo porque hay menos que pasen de dos horas; la OCDE no lo relaciona con la IA, y aquí tampoco se hace (capítulo 3).³¹

Donde sí se ve ya es en la nota de lo hecho en casa. En una universidad israelí estudiada por Hausman, Rigbi y Weisburd, las asignaturas que se evalúan con trabajos hechos en casa suben sus notas tras ChatGPT, y sus suspensos caen un tercio; lo que eso hace con la nota como medida es asunto del capítulo 5.³²

El primer gran examen que medirá a una cohorte que hizo toda la secundaria con la IA ya generalizada será PISA 2029. Sus alumnos nacieron en 2013 o 2014. En 2025 tenían 11 o 12 años; la cuenta es de este informe. Sus resultados saldrán hacia finales de 2030, dentro del plazo de los veredictos. Aun entonces, PISA solo mostrará asociaciones; la prueba

de causa seguirá viniendo de estudios con grupo de comparación, como el chino.

2.7 Quién pierde más

En China, la pérdida no se reparte por igual. Es mayor en ciencias sociales, un 27 %, que en lengua china, un 9 %. Y es mayor en los alumnos más jóvenes, en los chicos y en los que antes iban mejor.¹⁴

En Estados Unidos, una encuesta del Pew Research Center a 1.458 adolescentes, hecha en otoño de 2025, apunta a otro sitio. El 10 % hace todo o casi todo su trabajo escolar con ayuda de un chatbot, pero la proporción cambia con la renta. En los hogares que ganan menos de 30.000 dólares al año es uno de cada cinco; en los de más ingresos, el 7 %.² Si hacer todo o casi todo con un chatbot equivale a lo que en China era acabar los deberes mucho antes y con nota altísima, cosa que nadie ha medido, esa diferencia abriría una brecha por renta en lo que se aprende justo cuando la explicación se ha vuelto gratis para todos. Es una hipótesis de este informe.

Las dos cosas pueden ser ciertas a la vez, porque miden cosas distintas. La de China dice quién pierde más entre los que ya usan la IA; la de Estados Unidos, quién usa la IA de la manera que, según el estudio chino, parece dañar más. De España no hay datos de ninguna de las dos. La pregunta queda abierta, y el capítulo 3 sugiere que la respuesta depende menos de la máquina que de si el alumno tiene cerca a alguien que le haga sentarse.

2.8 Dos maneras de equivocarse

Frente a estos datos hay dos errores simétricos, y los dos tuvieron mucha difusión. El optimismo se apoyaba en un metaanálisis muy citado que atribuía a ChatGPT un efecto «grande» sobre el rendimiento y que su revista retractó en abril de 2026 por discrepancias en el análisis; el capítulo 4 lo cuenta.³³ El alarmismo, en un estudio con electroencefalogramas de personas que escribían con un chatbot, sin revisión por pares, cuya

sesión decisiva tenía 18 participantes y que mide conectividad cerebral, no aprendizaje.¹⁹

La posición de este informe no está a medio camino entre los dos. La IA no atonta: el quinto alumno chino pierde poco, y el ensayo de Lira muestra que, en condiciones controladas, puede enseñar. Pero tampoco enseña sola. Usada como la usaban cuatro de cada cinco alumnos del condado chino, quita la práctica, y la práctica era el aprendizaje.

2.9 Por qué estos veredictos

El primero es «muy probable» porque el mecanismo ya se mide en tres países con tres diseños distintos, y su causa no es técnica: quien usa la IA sin nadie que organice su tiempo tiende a acabar antes. La proporción solo se ha medido en el condado chino, pero allí es de cuatro de cada cinco, lejos de la mitad. La premisa lo agrava, porque borra el error que servía de aviso y convierte el modo tutor en un ajuste que se quita. En 2031 lo confirmaría un estudio con registros de tiempo, fuera de China, en el que más de la mitad de quienes usan la IA para los deberes acaben mucho antes con nota alta y rindan peor en un examen sin máquina.

El segundo es «probable» y no «muy probable» porque descuenta dos incertidumbres distintas. La primera es cuánto se pierde a escala de país: un condado no es un país, la evaluación ya está volviendo a la sala, como cuenta el capítulo 5, y los deberes se pueden rediseñar para que la práctica vuelva a hacerse delante de alguien. La segunda es si alguien lo mide y lo publica a tiempo: el estudio chino necesitó treinta meses de datos y un año más para publicarse. El umbral, un 5 % de la media anterior, es una cuarta parte de lo medido en China y lo fija este informe: separa una pérdida que se nota de una diferencia mínima que, con muestras nacionales, cualquier análisis distinguiría del azar. PISA 2029 no sirve para este veredicto, porque no compara a quien usa la IA con quien no la usa en igualdad de condiciones.

El tercero está en el tramo más bajo por el tamaño de lo medido. Un 20 % menos en un examen mensual es aprender menos, no dejar de saber leer, y el quinto alumno pierde poco. El nivel básico de PISA es el nivel 2, el que la OCDE toma como mínimo; en 2025 quedaba por debajo en

las tres materias a la vez el 19,7 % de los alumnos de la OCDE, frente al 16 % de 2022.^{29,34} Pasar del 30 % en 2029 exigiría una subida casi tres veces mayor que la del último ciclo; la cuenta es de este informe. Ni la caída de 2025, que empezó antes de la IA, ni ningún dato de este capítulo apuntan a algo de ese tamaño.

VEREDICTO Hasta 2031, si se cumple la premisa

La mayoría de quienes usan la IA por su cuenta para los deberes la usan para acabar antes, y aprenden menos que sin ella

■ ■ ■ ■ ■ **Muy probable**

Grave; solo se ve en un examen sin máquina, y ni siquiera eso basta para cambiar la conducta

Fuera de China, un estudio con grupo de comparación sobre una prueba externa nacional sin máquina (acceso a la universidad o evaluación censal) mide, entre quienes usan la IA para los deberes, una pérdida de al menos el 5 % de su media anterior

■ ■ ■ ■ ■ **Probable**

Grave; no se sabe si es recuperable

Una generación incapaz de leer o de razonar por la IA: en PISA 2029, más del 30 % de los alumnos de la OCDE queda por debajo del nivel básico en las tres materias, y la caída es mayor donde más se usa la IA para los deberes

■ ■ ■ ■ ■ **Muy improbable**

Catastrófico; los datos no apuntan ahí

Escala de los veredictos, siempre de aquí a 2031. Se explica al final del informe.

QUÉ CAMBIARÍA ESTE JUICIO

Bajaría el primer veredicto una réplica en otro país, con uso libre de la IA, que no encontrara la pérdida en un examen sin máquina. Bajaría la gravedad de los dos primeros cualquier prueba, que todavía no existe, de que lo perdido se recupera al volver el examen presencial o la práctica vigilada. Subiría el segundo a «muy probable» una segunda medición de varios años, fuera de China, que encontrara la misma pauta de deberes rápidos con nota alta y examen peor. El tercero solo se movería si PISA 2029, con la primera cohorte que hizo toda la secundaria con la IA a mano, mostrara una caída mayor que la de toda la década anterior.

— *Si la pérdida es tan clara, queda por explicar por qué la eligen, y por qué casi nadie usa el tutor que no da la respuesta, que existe y es gratis.*

El alumno elige la respuesta porque no nota lo que pierde

**96 % y
17 %**

alumnos de Tennessee que probaron el tutor Khanmigo, y sesiones con errores en las que el alumno mediano le escribió algo, a lo largo de dos cursos¹³

0,06-0,08

desviaciones típicas al año que ganaron esos alumnos: parecido a lo que da la misma plataforma sin IA¹³

1,4 horas

deberes diarios que declara el alumno medio de la OCDE en 2025, diez minutos menos que en 2022; es una media, que suben quienes pasan de dos horas³¹

3.1 Dos cursos con el tutor a un clic

En agosto de 2026, Philip Oreopoulos, economista de la educación en la Universidad de Toronto, publicó con N. Low el resultado de un ensayo con sorteo que duró dos cursos. En 18 institutos de secundaria inferior de Tennessee, alumnos flojos en matemáticas fueron asignados al azar a practicar, en sus horas de refuerzo, con Khan Academy y su tutor de IA, Khanmigo, configurado para orientar y no dar la solución. El grupo de comparación siguió con otros programas digitales y con clases en grupo pequeño.¹³

Las condiciones eran las mejores posibles para un tutor así: horario lectivo, un profesor en el aula y un tutor que sugería qué preguntarle. Al alumno solo le quedaba pedir ayuda cuando se atascaba.

Casi todos lo probaron: el 96 %. Después, el alumno mediano le escribió uno de cada tres días de práctica, y solo en el 17 % de las sesiones en las que se había equivocado. De los mensajes enviados durante los

ejercicios, el 14,5 % contenía una pregunta de matemáticas o un paso de razonamiento. Los demás eran respuestas escuetas a las preguntas del propio tutor, clics en las ayudas que él mismo proponía o mensajes de poco esfuerzo, como «no sé» o pedirle directamente la respuesta.¹³

El programa en conjunto sirvió, aunque poco: los alumnos ganaron entre 0,06 y 0,08 desviaciones típicas al año. La desviación típica, la unidad en que se comparan estos efectos, mide cuánto se separan normalmente las notas de los alumnos de la media, y permite poner en la misma escala pruebas distintas. El ensayo no tenía un grupo que practicara en Khan Academy sin tutor, así que no puede decir cuánto añadió este; los autores señalan que la ganancia se parece a la de practicar allí sin IA.¹³

Lo que sí creció fue la práctica. A lo largo del segundo curso, los alumnos activos pasaron de 42 a 48 minutos semanales de ejercicios. Sus conversaciones con el tutor, mientras, se quedaron en unas 2,6 por semana.¹³ La parte del programa que venía impuesta, la sesión de práctica, se hizo. La parte que dependía de la iniciativa del alumno, preguntar, casi no. La conclusión de los autores es que lo que manda es la implicación: hacer realidad la promesa de la tutoría con IA «exigirá conseguir que los alumnos la usen, no solo darles acceso» (traducción del informe).¹³

Sal Khan, fundador de Khan Academy, es parte interesada: el tutor es su producto. Aun así, reconoció que para la mayoría había sido «un no acontecimiento», y que lo que el ensayo ponía en primer plano era «cuánto importa la motivación».³⁵

En Tennessee no había ninguna máquina de respuestas a mano, y aun así casi nadie pidió ayuda. Fuera del aula, esa máquina está en la misma pantalla que el tutor. El ensayo no mide el tutor de la premisa, que sería mucho mejor; mide lo que la premisa no toca: si el alumno lo usa. La posición de este informe es que el límite del tutor de IA está en la demanda, no en la oferta, y que tiene dos causas que se refuerzan: el alumno no nota lo que pierde cuando elige la respuesta, y aunque lo notara, pagaría hoy lo que solo ganaría mañana.

3.2 El mismo patrón en todas partes

Tennessee no es una rareza. Donde las dos opciones conviven, en Estonia, entre el 64 y el 90 % de los alumnos ya usaba herramientas comerciales antes del programa. El tutor público, hecho para no dar la respuesta, lo usa cada semana en torno al 18 % de sus destinatarios (cuenta de este informe; capítulo 1).⁴ OpenAI, que admitió al lanzar su modo estudio que quien quiera tomar el camino fácil puede tomarlo, no ha publicado cuánto se usa.⁵

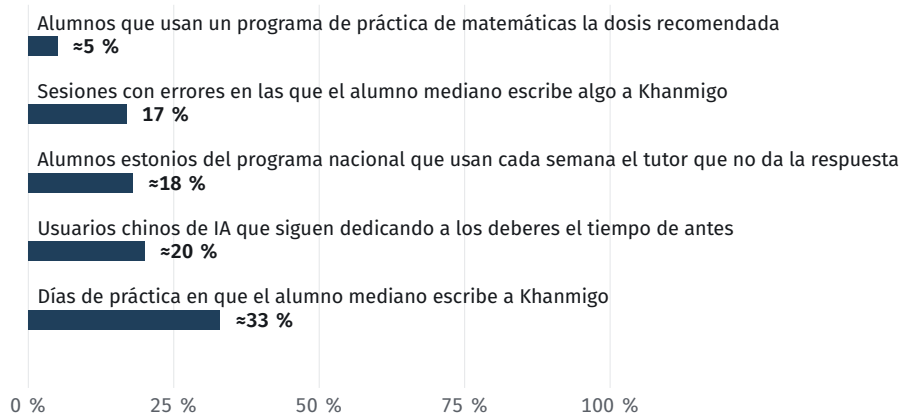
El condado chino del capítulo 2 muestra qué hacen los alumnos con una herramienta generalista cuando nadie organiza nada. Cuatro de cada cinco usuarios acababan los deberes más deprisa que el más rápido de los que no usaban IA.¹⁴ Strömberg y sus coautores lo leen así: las herramientas de tutoría bien diseñadas «ya existen, a menudo casi gratis», pero la mayoría de los alumnos «simplemente no las usan» y prefiere las generalistas, que facilitan externalizar los deberes. Por eso la pregunta esencial «no es tecnológica, sino organizativa e institucional» (traducción del informe).¹⁴

El patrón es anterior a la IA generativa. Los buenos resultados de los programas de práctica de matemáticas en línea más usados en Estados Unidos se medían en quienes cumplían la dosis recomendada, en torno a media hora semanal, y esa dosis la cumplía alrededor del 5 % de los alumnos.³⁶

Las cinco barras no se pueden comparar como si fueran una sola encuesta. Lo que comparten es la forma: en alumnos, en días o en sesiones, el uso que exige trabajo al alumno es minoritario, lo mismo con un programa de ejercicios que con un tutor que conversa.

La promesa de un tutor para cada alumno, la de las «dos sigmas» que examina el capítulo 4 y la de las empresas que ofrecen estos tutores, da por hecho que lo escaso era el tutor. Aquí tienen más razón Oreopoulos y Strömberg, por tres motivos. El abandono ocurre antes de que el alumno pueda comprobar si el tutor es bueno: el 96 % lo prueba y luego casi no le escribe. Se repite con herramientas de calidad muy distinta, de un programa de ejercicios a un tutor que conversa. Y solo cambia, como se verá más abajo, cuando cambia el entorno, no el modelo. El propio Oreopoulos lo resume así: si un alumno motivado quiere usar la tecnología

Figura 3 Cinco medidas del uso de tutores, plataformas de práctica e IA generalista, cada una con su denominador



Holt (2024); Oreopoulos y Low (2026); Fundación TI-Hüpe (marzo de 2026; el ≈18 % es cuenta de este informe); Strömberg, Lei y Wu (2026). Medidas distintas, con poblaciones y denominadores distintos; ninguna mide directamente el esfuerzo.

como un buen tutor, «está lista»; el problema es que la mayoría no la usa así (traducción del informe).³⁵

3.3 No nota lo que pierde

La primera causa viene del capítulo 2: lo que alguien consigue mientras practica guía mal lo que retiene después, y quien aprende juzga cuánto ha aprendido por lo fácil que le resultó. La consecuencia para la elección es que el alumno decide con una señal falsa.

El ensayo turco del capítulo 2 la mide por los dos lados: ni los alumnos del chat libre notaron que rendían peor, ni los del tutor que su examen no había mejorado.²³ En los dos casos la sensación de haber aprendido fue por delante del aprendizaje.

Los experimentos de S. D. Shaw y Gideon Nave, de la escuela de negocios Wharton, en la Universidad de Pensilvania, muestran lo mismo en adultos y en el momento de decidir. Con una IA a mano que los investigadores hacían acertar o fallar, la seguridad con que contestaban subió aunque la máquina se equivocara: la sensación de acertar no dependía de acertar.³⁷ Los autores lo llaman «rendición cognitiva», adoptar lo que dice

la máquina sin apenas examinarlo, y la distinguen de delegar una parte de la tarea sin perder el control. Lo que se sabe de estos experimentos viene de la nota de Wharton y de otras fuentes secundarias, no del artículo original; el capítulo 7 los cuenta enteros. Como los adultos tampoco lo notan, no notarlo no es cosa de la edad.

La pérdida, además, llega cuando ya nadie la relaciona con la elección. En el condado chino, el examen mensual a libro cerrado tardó unos seis meses en caer un 20 %, y la caída en los exámenes de acceso tardó unos dos años en completarse.¹⁴ El alumno que pega los deberes en un chatbot un martes no recibe ninguna señal ese martes. La recibe, si la recibe, en un examen lejano, mezclada con todo lo que le ha pasado entre medias.

3.4 Paga hoy lo que ganaría mañana

La segunda causa actuaría aunque el alumno viera la pérdida. En 2016, Adam Lavecchia, Heidi Liu y el propio Oreopoulos resumieron, en el manual de referencia de la economía de la educación, lo que la economía del comportamiento sabe de los estudiantes. Parten de «la propensión de los jóvenes a tomar malas decisiones a largo plazo», y la primera de las cuatro implicaciones que sacan es que «algunos estudiantes se centran demasiado en el presente» (traducción del informe).³⁸ El coste de esforzarse hoy es nítido; el beneficio de saber algo dentro de un año, vago. Entre los remedios que revisan, los que encajan con ese sesgo actúan sobre la situación inmediata: estructura, plazos fijados de antemano, un tiempo reservado. La disciplina los llama dispositivos de compromiso: atarse, o dejarse atar, antes de que llegue la tentación.³⁸

Aplicado a la elección, el reparto que sigue es de este informe. Con la máquina que responde, el alumno gana hoy: en el condado chino, los deberes pasaron de 64 a 45 minutos. Su nota, además, subió un 18 %. Lo pierde meses después, en el examen, sin aviso.¹⁴ Con el tutor pierde tiempo hoy: en otro ensayo del equipo de Oreopoulos, que se cuenta más abajo, con un tutor que no daba la respuesta los alumnos contestaron en torno a una pregunta menos en el primer ejercicio y tardaron 1,64 minutos más en cada una.³⁹ Lo que gana, sobre todo no perder y como mucho aprender un poco más, llega tarde y no se ve. Con el presente

pesando de más, quien compara las dos en el momento de elegir se queda con la primera.

Los autores del ensayo de Tennessee llegan a la misma explicación. En las mejores condiciones posibles, escriben, «pedir ayuda con la propia confusión siguió siendo una elección, y la mayoría de los alumnos la rechazó la mayor parte de las veces» (traducción del informe). Lo consideran coherente con el sesgo hacia el presente, la rutina y la tendencia a evitar pedir ayuda, y remiten a su manual de 2016.^{13,38} La tercera barrera añade otro coste inmediato: preguntar es reconocer que no se sabe, y pegar el enunciado en otra ventana no obliga a reconocer nada.

El tutor estonio del capítulo 1 intenta otra cosa: cuando un alumno le pide una redacción, le pregunta si va justo de tiempo y para qué la necesita.⁴ De lo anterior se deduce, y es deducción de este informe, que una pregunta en el momento pesa poco frente a una herramienta que está en la misma pantalla; no hay datos de qué hacen los alumnos después de contestarla.

3.5 Lo que sí cambia la conducta

Si la causa está en la situación y no en el carácter, cambiar la situación debería cambiar la conducta.

El primer indicio es el ensayo que acaba de citarse. En marzo de 2026, el equipo de Oreopoulos sorteó a más de 6.000 alumnos de Chattanooga, en Tennessee, para una sesión de práctica en la plataforma NUMI. Se sorteó por separado si aparecía, sin que lo pidieran, un tutor de IA que no daba la respuesta, y si regía una regla que no dejaba avanzar hasta acertar tres seguidas; unos tuvieron las dos cosas, otros una y otros ninguna. Con el tutor fueron más despacio, pero tras un error necesitaron menos intentos para volver a acertar: la IA hizo la recuperación «más productiva, pero no más rápida». Una semana después, tutor y regla juntos dieron unos 3 puntos porcentuales más de acierto en lo practicado, con una significación que los autores llaman marginal y sin respaldo, advierten, para hablar de ganancias transformadoras.³⁹ Su conclusión de fondo vale para todo el capítulo: la misma tecnología que explica un error puede ahorrar al alumno pensar el problema, y cuál de las dos cosas ocurre depende «no

solo de las capacidades del modelo», sino del diseño del entorno y de la conducta que provoca (traducción del informe).³⁹

El segundo está en los experimentos de Shaw y Nave. La rendición bajaba cuando se pagaba por acertar y cuando cada respuesta se corregía en el acto.³⁷ La traducción a la escuela es de este informe: esos dos ingredientes son los de un examen corto y frecuente sin la máquina, que premia saber y dice enseguida si se sabe.

El tercero es el de los internados de la India que cuenta el capítulo 4: cuando unos coordinadores protegen un tiempo de práctica y vigilan el aula, el uso de Khan Academy, sin IA generativa, se multiplica.⁴⁰ A los tres se suma un anuncio, que no es un dato. Tras el ensayo de dos cursos, Sal Khan explicó que la IA «no podía estar al lado del contenido»: tenía que ir dentro, para que el esfuerzo productivo fuera más difícil de esquivar.³⁵

Ninguna informa mejor al alumno. Todas cambian la situación: el tutor aparece sin que lo llamen, no se avanza sin acertar, hay un examen pronto, alguien vigila el tiempo. Y todas le cuestan al alumno tiempo o comodidad en el momento, justo lo que evita cuando elige. El cambio no puede salir de él en ese momento; tiene que fijarlo alguien antes.

3.6 La objeción de siempre

Profesores y familias suelen plantear la objeción de fondo con otras palabras: el alumno medio no quiere aprender y no dedica ni una hora en casa, así que un chatbot no le hará aprender, sino que le volverá más ocioso. Acierta en el mecanismo: el chatbot no hace aprender a nadie, acorta el camino del menor esfuerzo, y con el presente pesando de más ese camino gana. Se equivoca en el marco, y el marco importa para lo que se haga después.

El dato no la desmiente, pero la matiza. En PISA 2025, el alumno medio de la OCDE declara 1,4 horas diarias de deberes, diez minutos menos que en 2022 (la caída que vio el capítulo 2).³¹ Es una media, y la suben quienes pasan de dos horas; no sabemos cuántos no hacen ninguna. Es además tiempo declarado, de deberes y no de estudio, y nadie sabe qué parte consiste ya en copiar y pegar.

El marco es otro asunto. Lo que describen los datos no es un carácter,

sino una conducta que cambia cuando cambia la situación: los mismos alumnos de Tennessee hicieron la práctica que venía impuesta y dejaron el tutor que era opcional, y en la India el uso se multiplicó cuando alguien protegió el tiempo. No es pereza como defecto moral, sino un sesgo predecible, que se corrige cambiando la situación.

3.7 Quién decide

Dylan Wiliam, profesor emérito del Instituto de Educación del University College de Londres y autoridad mundial en evaluación formativa, propone, según el resumen de Craig Barton, una prueba para decidir si se delega una tarea en la IA: «¿Es esto algo en lo que quiero mejorar?». ⁴¹ Para un adulto es el mejor criterio que conoce este informe, porque separa lo que conviene delegar de lo que conviene practicar.

Pero la prueba supone que quien la aplica contesta mirando lejos. Un chico de 14 años ante diez ecuaciones, una tarde cualquiera, contestará con sinceridad que no quiere mejorar en eso, porque el beneficio de saber álgebra llega años después y no lo ve. Es lo que predice el sesgo hacia el presente: el fallo no es de la regla, sino del momento en que se aplica. Ese sesgo se atenúa con la edad, pero no desaparece. ³⁸ Por eso al adulto la regla le sirve sobre todo si la aplica antes, no en mitad de la tarea, y a un chico de 14 años le sirve menos. La posición de este informe es que la decisión no puede quedar solo en manos del alumno en el momento de la tentación, sino que tiene que tomarse antes: el propio alumno en un día tranquilo, fijando dónde y cuándo trabaja sin la opción fácil, o su familia, su escuela o el diseño de la herramienta.

Ahí entra la desigualdad. Si lo escaso es alguien que fije el sitio y la hora, la IA ayuda a quien ya lo tiene, sea una familia que vigila, unas clases de refuerzo o la costumbre de organizarse, y deja solo a quien no. Becky Francis, que dirige la Education Endowment Foundation, advirtió en junio de 2026, en la misma declaración que cita el capítulo 1, que hay que evitar que estas herramientas ensanchen las desigualdades. ²¹

Los datos todavía apuntan a sitios distintos. En Estados Unidos, uno de cada cinco adolescentes de hogares con menos de 30.000 dólares al año hace todo o casi todo su trabajo escolar con un chatbot. En los hogares con

más ingresos lo hace el 7 %.² En el condado chino, en cambio, perdieron más los alumnos que iban mejor.¹⁴ Las dos cosas pueden ser ciertas a la vez, y cuál pesa más es la pregunta que queda abierta.

3.8 Qué añade la premisa

La premisa mejora la oferta: un tutor más sabio que cualquier profesor, con paciencia sin límite y en cualquier idioma. No toca la demanda, porque el sesgo hacia el presente no depende de la calidad de la máquina. Strömberg y sus coautores lo dicen de su propio dato: este problema de incentivos «no desaparecerá con los avances de la tecnología» (traducción del informe).¹⁴ Y la premisa empeora la elección.

También mejora la máquina que da la respuesta: la premisa dice que resolverá lo comprobable, y todo lo que la escuela pide lo es. En los experimentos de Shaw y Nave, consultar a la IA subía mucho el acierto cuando la IA acertaba (capítulo 7).³⁷ Con una máquina que casi nunca falla, rendirse deja de costar nada en la nota. Y sale más caro para el aprendizaje: cuando la máquina fallaba, copiar aún obligaba a revisar la respuesta; cuando casi nunca falla, no queda nada que revisar, y la pérdida no avisa (capítulo 2).

Y con los modelos que funcionan en el móvil sin conexión, el modo tutor queda en un ajuste que cualquiera quita (capítulos 1 y 2). Un tutor opcional mejor compete en la misma pantalla con una máquina de respuestas mejor, y la elección se decide antes de que la calidad del tutor llegue a notarse.

Por eso la promesa de un tutor mejor que cualquier profesor descansa en dos cosas: la mitad del criterio, que sabe qué necesita cada alumno y sostiene su esfuerzo durante meses y es la que no tiene fecha (capítulo 1), y una conducta que ninguna mejora del modelo produce. Aunque llegara la primera, faltaría la segunda.

3.9 Por qué estos veredictos

Que la mayoría siga usando la máquina que da la respuesta aunque tenga el tutor a mano no va en el bloque de veredictos. Es un beneficio que no

llega, no un daño, y el daño que produce, aprender menos sin notarlo, ya lo recoge el primer veredicto del capítulo 2.

Que, con la IA a mano, los alumnos de hogares con menos renta pierdan más o ganen menos que los demás es «posible». El mecanismo apunta ahí: si lo escaso es alguien que fije el sitio y la hora, lo tiene más quien tiene más. En Estados Unidos el uso intensivo se concentra en los hogares pobres.² La enseñanza para evaluar lo que dice la IA llega más a los alumnos favorecidos.³¹ En Nigeria, con un tutor supervisado, ganaron más los que ya iban mejor.⁴² Pero en China perdieron más los que iban mejor, y ninguno de los estudios con grupo de comparación que reúne este informe separa todavía la pérdida por la renta del hogar.

VEREDICTO Hasta 2031, si se cumple la premisa

Los estudios con grupo de comparación publicados hasta 2031 encuentran que, con la IA a mano, los alumnos de hogares con menos renta pierden más o ganan menos que los demás

■ ■ ■ ■ ■ **Possible**
Grave para la equidad; se nota tarde

Escala de los veredictos, siempre de aquí a 2031. Se explica al final del informe.

QUÉ CAMBIARÍA ESTE JUICIO

El veredicto subiría a «probable» si los primeros resultados del fondo de la Education Endowment Foundation, previstos para 2027, separaran a los alumnos por nivel socioeconómico y encontraran pérdidas mayores entre los de hogares con menos renta; bajaría si la pérdida resultara igual en todos los grupos. La posición sobre el uso del tutor cambiaría con un tutor proactivo e integrado en el contenido, no opcional, como el que Khan Academy anunció en 2026, que diera ganancias duraderas fuera del aula vigilada; lo más cercano hoy es una sola sesión con un efecto pequeño medido a la semana.

— Si el límite no es la calidad del tutor, sino que se use, la promesa central de la premisa, un tutor mejor que cualquier profesor, vale menos de lo que parece. Cuánto menos es la pregunta siguiente.

Un tutor mejor que cualquier profesor no sube el techo si no sube el tiempo

0,16-0,21

desviaciones típicas que da la tutoría humana en programas de más de 400 alumnos de Estados Unidos, medidos con pruebas estandarizadas; en el conjunto de ensayos, los de menos de 100 alumnos dan 0,55⁴³

de 7,2 a 47,4

minutos semanales de uso de la misma plataforma cuando unos coordinadores protegen el tiempo en internados de la India; las matemáticas suben unas 0,5 desviaciones típicas en 31 semanas, sin IA generativa⁴⁰

0,23

desviaciones típicas en inglés con un tutor de IA en Nigeria, con profesor, horario fijo, examen en papel y alumnos voluntarios⁴²

4.1 La escuela del tutor

A finales de agosto de 2026, Alpha School, una cadena privada estadounidense, empezaba el curso con unas cincuenta sedes y pruebas en tres escuelas públicas. Es lo más parecido a la premisa que existe en una escuela de verdad: sus alumnos pasan dos horas diarias con aplicaciones que se adaptan a cada uno gracias a la IA, y el resto de la jornada en talleres. La matrícula cuesta entre 40.000 y 75.000 dólares al año. La escuela asegura que sus alumnos aprenden «el doble». No ha publicado los datos ni se ha sometido a una evaluación independiente, y los investigadores consultados por Scientific American señalan la selección de los alumnos como explicación posible.¹⁸

Gerald LeTendre, sociólogo de la educación en la Universidad Estatal de Pensilvania, dijo a la revista que no hay pruebas de que un modelo de tutoría pura con IA consiga «todos los efectos» que produce estudiar en un aula.¹⁸

Lo que sí se sabe es cómo organiza la jornada. No es un tutor que sustituya a la escuela. Es una escuela que reserva dos horas de cada día para el tutor, decide qué se hace el resto del tiempo y trabaja con alumnos cuyas familias pagan decenas de miles de dólares por estar ahí. El sitio, la hora y quién se sienta en el aula los decide la escuela, no la máquina.

El capítulo anterior concluyó que el límite del tutor es que se use, no su calidad, y dejó abierto cuánto vale entonces la promesa. Se prometen dos desviaciones típicas de mejora. En los programas grandes de Estados Unidos, la tutoría humana da entre 0,16 y 0,21. Lo que falta para acercarse a lo prometido no es un tutor mejor, sino que alguien consiga que el alumno pase más tiempo trabajando con él.

4.2 De dónde sale la promesa

La idea de un tutor para cada niño se apoya en una cifra. En 1984, el psicólogo Benjamin Bloom sostuvo que los alumnos con un tutor particular rendían dos desviaciones típicas más que los de una clase corriente, un salto enorme en la unidad que explicó el capítulo 3. El obstáculo era el precio: nadie podía pagar un tutor a cada alumno. Si la IA lo hace gratis, el obstáculo desaparece. Ese razonamiento sostiene la premisa del capítulo 1 aplicada a la escuela y está detrás del «el doble» de Alpha.

En 2024, Paul von Hippel, que estudia métodos de evaluación en la Universidad de Texas en Austin, volvió a los experimentos de los que salió la cifra. Los hicieron alumnos de Bloom en los años ochenta, duraban tres semanas y tras cada unidad había una prueba de dominio, una corrección y otra prueba; los alumnos con tutor recibían, además, cerca de una hora semanal más de enseñanza. Su conclusión es que aquellos efectos fueron «reales y notables», pero se obtuvieron «en una prueba estrecha y especializada, y no solo con tutoría» (traducción del informe). El efecto típico de la tutoría, estima, ronda un tercio de desviación típica, y le parece «temerario» prometer dos con IA cuando la tutoría humana rara

vez los ha dado.⁴⁴

Lo que von Hippel desmonta no es que la tutoría funcione, sino de qué estaba hecha la cifra. Tenía tres ingredientes y una inflación. El primero, la calidad del tutor, la máquina de la premisa lo iguala. El segundo, una prueba tras cada unidad que obligaba a recuperar de memoria lo aprendido, que es, según el capítulo 2, la parte de la práctica que enseña, también puede imponerlo la máquina, si el alumno se queda delante. El tercero, la hora semanal añadida con el alumno presente, no lo pone ninguna máquina. Y una parte de las dos sigmas la ponía la prueba estrecha, ajustada a lo que se acababa de enseñar: eso no es aprendizaje que un tutor tenga que igualar. De la premisa no se deduce, por tanto, que un tutor sobrehumano suba el aprendizaje medio dos desviaciones típicas.

4.3 El techo a escala

Cuánto da la tutoría cuando pasa de un programa pequeño a uno grande lo midieron en 2024 Matthew Kraft, economista de la educación en la Universidad Brown, y dos colegas, con más de 260 ensayos con sorteo. En el conjunto de los ensayos, los programas de menos de 100 alumnos dan 0,55 desviaciones típicas. Los de 100 a 399 dan 0,32. Los autores prefieren una muestra más exigente, la de programas de Estados Unidos medidos con pruebas estandarizadas. En ella, los de 400 a 999 alumnos dan 0,21, y los de 1.000 o más, 0,16. Son dos muestras, y parte de la distancia puede venir del tipo de prueba, pero los autores atribuyen la caída sobre todo al tamaño. La literatura está hecha de programas pequeños: el 59 % de los estudios tiene menos de 100 alumnos. En cambio, el 76 % de los alumnos de la escuela pública estadounidense vive en distritos donde cualquier programa pasaría de 1.000.⁴³

Los autores encuentran un paquete de rasgos que protege en parte contra esa caída: tutoría en persona, dentro del horario escolar, varias veces por semana y en grupos muy pequeños.⁴³ Tres de los cuatro tratan de cuánto tiempo tiene el alumno al tutor para él y con qué regularidad. Ninguno mide lo que sabe el tutor, así que el dato no dice que eso no cuente.

La lectura que sigue es de este informe. Al crecer, un programa pierde sobre todo lo que depende de organizar: que el alumno vaya, que vaya a menudo y que el grupo siga siendo pequeño. Puede que pierda también calidad, porque para miles de alumnos hacen falta cientos de tutores y no todos son buenos, y esa parte la premisa sí la arregla. Los datos son compatibles con que el techo lo ponga la dosis, pero no lo prueban. Lo distinguiría un programa grande con tutores excelentes y poca dosis, comparado con otro de tutores corrientes y mucha.

Un tutor sobrehumano al alcance de millones de alumnos es, por construcción, un programa de más de 1.000, con la mejor calidad posible y la dosis que cada alumno quiera darle. Se parece al ensayo de Tennessee del capítulo 3, donde, según los autores, la ganancia con el tutor a un clic se pareció a la de practicar en la plataforma sin IA.¹³

4.4 Lo que sube el techo es el tiempo

Philip Oreopoulos, autor del ensayo de Tennessee, firma otro, publicado en enero de 2026, en el que lo añadido no fue un tutor, sino organización. En 83 internados públicos de secundaria inferior de Uttar Pradesh, todos con la plataforma clásica de Khan Academy y sin IA generativa, una parte de los centros recibió coordinadores. Arreglaban la conexión y las altas, protegían un tiempo de práctica, vigilaban el aula y seguían el progreso de cada alumno. El uso de la plataforma pasó de 7,2 a 47,4 minutos semanales. En 31 semanas, las matemáticas subieron unas 0,5 desviaciones típicas, en todos los niveles. Según los autores, estas plataformas rinden poco a escala «no porque la tecnología sea ineficaz, sino porque las escuelas se enfrentan a fricciones de aplicación considerables» (traducción del informe).⁴⁰

Dentro de ese ensayo la comparación es limpia: los dos grupos tenían la misma plataforma y el mismo contenido. Lo que movió el aprendizaje fue la organización que llevó al alumno a ella. El uso multiplicado por seis es su huella, no una causa aislada, porque los coordinadores hacían más cosas. Los minutos solos tampoco bastan: en Tennessee, los alumnos activos practicaban entre 42 y 48 minutos semanales, con un efecto mucho menor, aunque allí el grupo de comparación también tenía programas y

un profesor.¹³

En Nigeria se añadió otra cosa. Un equipo del Banco Mundial, encabezado por los economistas De Simone y Tiberti, sorteó a los alumnos que se apuntaron en nueve institutos públicos del estado de Edo. Los elegidos hicieron doce sesiones de 90 minutos, dos tardes por semana, después de clase. Trabajaban por parejas con la versión gratuita de Copilot, el asistente de Microsoft. Un profesor abría y cerraba cada sesión, con un tema del programa de inglés. La prueba final, en papel y sin IA, dio 0,31 desviaciones típicas en el conjunto, que incluía conocimiento de IA y destrezas digitales. En inglés, el resultado principal, dio 0,23. También mejoraron los exámenes ordinarios de fin de curso, y cuantas más sesiones, más ganancia.⁴²

Los autores no atribuyen el resultado a la máquina, sino a «la intervención en su conjunto», y advierten que, sin ese entorno supervisado, los beneficios pueden quedarse en «los alumnos que ya están motivados para aprender» (traducción del informe).⁴² El ensayo no separa la máquina del profesor ni de las tres horas semanales añadidas. Muestra que el resultado llegó con todo ese envoltorio: un sitio, una hora fija, un adulto que dirigía, un examen en papel y voluntarios.

El capítulo 3 explicó por qué hace falta ese envoltorio: por el sesgo hacia el presente, el coste de esforzarse hoy es nítido y el beneficio de saber dentro de un año, vago. En sentido estricto, un dispositivo de compromiso lo elige quien se ata, como el plazo que alguien se fija de antemano para no aplazar.³⁸ Los coordinadores de la India, las dos tardes con profesor de Nigeria y el paquete de Kraft los ponen otros, pero hacen el mismo papel: fijan de antemano el sitio y la hora que el alumno, solo, no fijaría. La escuela lo hace cada día y con todos, y además pasa lista.

Tabla 2 Qué se añadió en cada programa y frente a qué se midió (lectura del informe)

Programa	Qué se añadió	Frente a qué se compara	Efecto de lo añadido, en desviaciones típicas
Tutoría humana, más de 260 ensayos ⁴³	Un tutor humano	Alumnos sin esa tutoría en cada ensayo	0,55 en programas de menos de 100 alumnos; 0,16 en los de 1.000 o más de Estados Unidos
Tennessee, dos cursos ¹³	Khan Academy con un tutor de IA opcional, en horario lectivo	Otros programas digitales y grupos pequeños con un profesor	0,06-0,08 al año
Internados de la India ⁴⁰	Coordinadores que llevan al alumno a la plataforma, sin IA generativa	La misma plataforma, sin coordinadores	≈0,5 en 31 semanas
Nigeria, inglés ⁴²	Copilot por parejas, con un profesor, dos tardes por semana	Nada esas tardes	0,23 en doce sesiones

Los grupos de comparación, las pruebas y los plazos son distintos, así que la tabla no sirve para ordenar programas. Donde lo añadido fue organización que llevó al alumno a trabajar, en la India, el efecto fue grande. Donde fue un tutor opcional dentro de un tiempo ya fijado, en Tennessee, fue pequeño y, según los autores, parecido al de la plataforma sin IA. Lo que acompaña al efecto es el uso obligado, no quién explica.

4.5 Dónde sí rinde la IA

El primer sitio donde rinde es detrás del adulto. En un ensayo prerregistrado de la Universidad de Stanford, con 874 tutores y 1.787 alumnos de

un distrito del sur de Estados Unidos, la IA no hablaba con el alumno: le sugería al tutor humano, en directo, qué preguntar. En dos meses, la probabilidad de superar la pequeña prueba con que acaba cada lección subió 4 puntos. Con los tutores peor valorados, la subida fue de 9 puntos, y sus alumnos alcanzaron a los de los tutores buenos. La herramienta solo se usó en el 29 % de las sesiones, y no se detectó mejora en el examen de fin de curso.⁴⁵ En Inglaterra, en un ensayo de la Education Endowment Foundation, los profesores de ciencias que preparaban clases con ChatGPT ahorraron un 31 % del tiempo sin perder calidad.⁴⁶ Y en un ensayo exploratorio de Google y la plataforma Eedi con 165 alumnos británicos, pequeño y firmado por quien fabrica el modelo, un tutor de IA revisado por un humano igualó o superó al tutor humano solo, medido dentro de la misma sesión.⁴⁷ En los tres casos, alguien organizaba el tiempo y la máquina mejoraba lo que pasaba dentro: más acierto al final de la lección, menos horas de preparación. Falta ver si dura: que el ensayo de Stanford no moviera el examen de fin de curso es el patrón del capítulo 2, en el que sube la medida cercana y no la lejana.

El segundo es donde no hay nadie. En Nigeria, el grupo de comparación no tenía nada esas tardes.⁴² Allí faltaban a la vez la explicación y la hora, y el ensayo dio las dos, así que no dice cuál pesó más. Un tutor que explica bien tiene más que aportar donde no compite con un profesor. Ganaron más quienes ya partían con mejor rendimiento. Es un solo ensayo, evaluado por el mismo organismo que lo diseñó y sin seguimiento a largo plazo.

El tercero es el alumno que ya se obliga solo. En un ensayo de Harvard publicado en 2025, estudiantes de Física hicieron una lección con un tutor de IA en casa y otra en clase, con métodos activos, y luego se invirtió el orden. El tutor, hecho a medida y con soluciones escritas por los profesores, dio más del doble de ganancia en menos tiempo.⁴⁸ Eran alumnos de Harvard que hacían la lección porque contaba para la asignatura, con una prueba inmediata y solo dos lecciones. Greg Kestin, que diseñó ese tutor, advierte a propósito de Alpha del riesgo de un conocimiento «quebradizo».¹⁸ Estos alumnos sí traen un dispositivo de compromiso en sentido estricto, elegido por ellos: una asignatura exigente, una nota que les importa, la costumbre de estudiar. Para ellos, el tutor de la premisa es una ganancia real. Son también quienes menos lo necesitaban, y de ahí

puede nacer la brecha que el capítulo 3 considera posible.

4.6 Cómo leer un anuncio

Un resultado muy superior a los de este capítulo, en muchos alumnos, exige saber a quién se midió, frente a qué y con qué prueba.

El caso más citado ya se retiró. En mayo de 2025, una revista del grupo Springer Nature publicó un metaanálisis de 51 estudios que atribuía a ChatGPT un efecto «grande» sobre el rendimiento. Lo cifraba en 0,867 desviaciones típicas y recomendaba usarlo en varios niveles educativos. Acumuló 433 citas antes de que la revista lo retractara, el 22 de abril de 2026, por discrepancias en el análisis.³³ Aun sin esos errores, esa literatura está hecha sobre todo de estudios pequeños y de pocas semanas, medidos con la herramienta delante o con pruebas del propio investigador.

El segundo es de septiembre de 2026. El Gobierno de El Salvador anunció que las 171 escuelas de su programa piloto de modernización, que incluía IA, habían alcanzado un nivel «equiparable a Alemania y Suecia». La prueba era PISA for Schools, una herramienta de la OCDE para centros, no el estudio PISA. Se hizo a unos siete alumnos voluntarios por escuela, en escuelas elegidas, según el propio Gobierno, por sus «condiciones de infraestructura, acceso a internet y disposición». No hubo medida previa ni grupo de comparación, y el piloto sumaba dispositivos, conexión, un cambio de método y apoyo a los docentes, así que nada se puede atribuir a la IA.⁴⁹ El PISA nacional de esa misma semana no desmiente el piloto, porque mide a otros alumnos. Con una muestra representativa, sitúa al país en 346 puntos en matemáticas, unos 120 por debajo de la media de la OCDE. Respecto a 2022, es una mejora pequeña.⁵⁰

Los tres casos fallan por sitios distintos. Alpha y El Salvador miden a alumnos elegidos y sin grupo de comparación; El Salvador, además, mezcla la IA con muchas otras piezas. El metaanálisis tenía errores de análisis y descansaba en estudios medidos cerca del uso, el mismo defecto que von Hippel encontró detrás de las dos sigmas de Bloom. Un sorteo arregla la comparación, pero no quién entra en la muestra. La selección solo se arregla midiendo a alumnos que no se han elegido a sí mismos, y por eso el resultado de Nigeria, sorteado entre voluntarios, puede no

valer para los demás.

4.7 Qué añade la premisa

La premisa da un tutor mejor que cualquier profesor, gratis, paciente y a cualquier hora. No da tiempo. Lo que añade cuando el tiempo ya está garantizado se puede medir con los ensayos de este capítulo: más acierto al final de la lección cuando el tutor humano es flojo, casi un tercio menos de horas de preparación para el profesor y una explicación donde no había ninguna. Ninguno ha mostrado todavía un efecto duradero comparable al de los internados de la India, que no tenían IA generativa, y todos quedan lejos de las dos desviaciones típicas prometidas.

El único tutor cuyo diseño cambió algo la conducta, el de NUMI que cuenta el capítulo 3, no dio más tiempo. En una sola clase de duración fija, los alumnos hicieron menos ejercicios y más despacio, y la ganancia vino de usar mejor ese tiempo. El efecto fue pequeño y de significación justa.³⁹ Si se confirmara a escala y durase, el diseño también subiría el techo, y este juicio tendría que cambiar.

La posición de este informe es que, con la premisa, la escuela se hace más necesaria, no menos. La explicación abunda; lo escaso pasa a ser un sitio y una hora en que se trabaja sin la opción fácil, con un adulto que nota quién falta. En el choque sobre el tutor para cada niño, tienen más razón von Hippel, Kraft y Oreopoulos que quienes prometen las dos sigmas, por una razón concreta: la premisa mejora la calidad del tutor, y los datos son compatibles con que el techo a escala lo ponga el uso organizado, no la calidad. Quienes prometen aciertan en una mitad. Donde alguien obliga a usar el tutor, como en la tutoría organizada o con un profesor detrás, o donde el alumno ya se obliga solo, el tutor rinde en las medidas cercanas; si dura, está por ver. En clase y como opción, en Tennessee, rindió poco. La otra mitad, la de que la escuela sobra, no se sigue de ningún dato: la escuela que más se parece a la premisa sigue siendo una escuela, y cara.

4.8 Por qué estos veredictos

El primer veredicto se formula como un daño que dejaría rastro en boletines oficiales y presupuestos: un programa retirado o recortado sin haber mostrado efecto. Es «probable» porque ya ha pasado una vez. Corea implantó libros de texto con IA en marzo de 2025 y, cinco meses después, el Parlamento les retiró el rango de libro de texto.¹⁷ Los anuncios sin grupo de comparación, como el de El Salvador o el de Alpha, que ya se prueba en escuelas públicas, empujan en esa dirección.^{18,49} No llega a «muy probable» porque un programa sin efecto demostrado también puede seguir sin que nadie lo retire; entonces el daño existe, pero no deja este rastro. Es moderado: se pierden dinero y horas de clase, que se recuperan, aunque sean justo el recurso que este capítulo encuentra escaso.

En el segundo, un gobierno regional es un Estado federado, una comunidad autónoma o una provincia, y cuenta una ley o un presupuesto. Un distrito escolar suelto no cuenta. Tampoco las tres escuelas públicas que prueban el modelo de Alpha: son pruebas, no recortes de plantilla. Es «improbable» porque esos países van en la dirección contraria. España va a gastar unos 2.800 millones de euros al año en bajar el número de alumnos por aula (capítulo 6).⁵¹ Corea dio marcha atrás con sus libros de texto.¹⁷ Lo harían más probable dos precursores: un resultado grande de un tutor de IA en un ensayo con sorteo, que alguien pudiera invocar, y recortes educativos generales por presión sobre los presupuestos públicos. Sería grave donde ocurriera, porque quitaría lo que los datos señalan como escaso: el adulto que pone la hora.

VEREDICTO Hasta 2031, si se cumple la premisa

Entre 2027 y 2031, un programa nacional o regional de IA para alumnos de un país de la OCDE, extendido sin ensayo con grupo de comparación, se retira o se recorta sin haber mostrado efecto en una evaluación externa, después de ocupar horas lectivas

■ ■ ■ ■ ■ **Probable**

Moderado: dinero y tiempo lectivo perdidos, recuperables

Un gobierno nacional o regional de un país de la OCDE aprueba reducir plantillas docentes o programas de tutoría humana citando la IA como sustituto

■ ■ ■ ■ ■ **Improbable**

Grave donde ocurra

Escala de los veredictos, siempre de aquí a 2031. Se explica al final del informe.

QUÉ CAMBIARÍA ESTE JUICIO

La posición del capítulo cambiaría con un ensayo grande con sorteo en el que un tutor de IA, usado sin un adulto que organice, haga trabajar bastante más al alumno mediano y aprender más en una prueba sin máquina meses después. También cambiaría si un diseño como el de NUMI, que mejora el uso del tiempo sin alargarlo, diera a escala un efecto claro y duradero. El primer veredicto bajaría si los gobiernos exigieran un ensayo con grupo de comparación antes de extender un programa de IA en las aulas. El segundo subiría con cualquiera de sus dos precursores.

— *Si la escuela pasa a ser sobre todo el lugar y la hora en que se trabaja sin la opción fácil, tiene que saber quién ha aprendido. Esa medida es lo siguiente que rompe la premisa.*

La nota deja de medir: lo que se evalúa fuera de una sala mide a la máquina

94 %

exámenes escritos por la IA que los correctores no detectaron en las pruebas para casa de Psicología de la Universidad de Reading, con más nota que los alumnos reales (2024)¹¹

un tercio

caída de los suspensos en las asignaturas evaluadas con trabajos en una universidad israelí tras ChatGPT; el puesto en las de examen predice cada vez menos el de las de trabajos³²

42 %

universitarios británicos que creen que un trabajo hecho por IA sacaría buena nota en su carrera³

5.1 Dos carriles en Sídney

En noviembre de 2024, la Universidad de Sídney anunció que, desde el segundo semestre de 2025, sus evaluaciones pasarían a ser de dos tipos. Las «seguras» se hacen en persona y sin ayudas, y son las que certifican lo aprendido. Las «abiertas» lo permiten todo, inteligencia artificial incluida.⁵² Adam Bridgeman, el vicerrector adjunto de innovación educativa que diseñó el sistema, lo justificó en una frase: «No es ningún secreto que cualquier evaluación que se hace en casa la puede completar la IA con un nivel alto, y no es práctico, ni aplicable, ni deseable prohibir estas herramientas en bloque».⁵²

Sídney no prohíbe la máquina ni compra un detector: cambia de sitio la certificación. El trabajo hecho en casa sigue sirviendo para aprender, pero ya no para acreditar lo que sabe el alumno.

La decisión tiene dos razones. En los ejercicios con respuesta comprobable, resolverlos es lo que la premisa da a la máquina (capítulo 1). En las redacciones y los trabajos no hace falta la premisa: en 2024, en una prueba ciega de la Universidad de Reading, la corrección ya no distinguía lo escrito por la máquina.

Quien usa la máquina para acabar antes aprende menos y no lo nota (capítulos 2 y 3), y un tutor mejor no lo evita si nadie protege el tiempo (capítulo 4). Alguien tendría que notarlo por él: el profesor y la familia mientras estudia, la universidad o el empleador después. Todos miran la nota, y la nota de lo que se hace fuera de la vista sale del mismo trabajo que la máquina hace por él.

5.2 El termómetro que marca al revés

La nota de los deberes y de los trabajos hacía tres papeles a la vez: para el alumno era el objetivo; para el profesor y la familia, el termómetro con el que sabían si iba aprendiendo; para la universidad o el empleador, un resumen de lo hecho. Convivían por una razón que nadie diseñó: sacar buena nota sin saber costaba casi tanto como saber. Copiar exigía un compañero dispuesto o un riesgo. Era una de las defensas por coste que describe el capítulo 2 del Informe I, y protegía a la vez contra la ignorancia propia y contra el engaño.

Que una medida usada como objetivo acaba midiendo mal tiene nombre. El economista Charles Goodhart, del Banco de Inglaterra, advirtió en 1975 que cualquier regularidad estadística tiende a venirse abajo en cuanto se usa para controlar.⁵³ La antropóloga Marilyn Strathern le dio en 1997 la forma que se cita hoy, y su ejemplo era justo una nota de examen universitaria: «Cuando una medida se convierte en objetivo, deja de ser una buena medida. Cuanto más se da por hecho un 2.1 en los exámenes, peor sirve para distinguir el rendimiento de cada uno» (traducción del informe).⁵⁴ El 2.1 es el notable de los grados británicos: si casi todos lo sacan, no separa a nadie.

Mientras torcer una medida cuesta, la medida se estropea poco a poco: hay que aprender a producir la cifra sin la cosa. La máquina quita ese coste de una vez, porque produce el objeto que se califica, el problema

resuelto o la redacción, sin el saber que la nota quería medir. La medida ya no se desvía: pierde de golpe lo que decía, y puede llegar a decir lo contrario.

Strömberg, Lei y Wu pudieron ver qué le pasa al termómetro, porque en el condado chino del capítulo 2 tenían a la vez la nota de los deberes semanales y la de los exámenes mensuales a libro cerrado de 26.811 alumnos. Entre quienes no usaban IA, mejor nota en los deberes anunciaba mejor examen, como siempre. Entre quienes la usaban, se invertía: más nota en los deberes, peor examen. Los autores sacan la consecuencia práctica, que una subida rápida de la nota de los deberes es «una señal de alarma» de que el aprendizaje está cayendo.¹⁴

En una universidad israelí, los economistas Hausman, Rigbi y Weisburd compararon, para los mismos 36.000 alumnos, las asignaturas evaluadas con trabajos hechos en casa y las evaluadas con examen vigilado, antes y después de ChatGPT. En las primeras, la nota media subió entre uno y un punto y medio sobre cien. Los suspensos cayeron un tercio. Antes, el puesto de un alumno en las asignaturas de examen anunciaba su puesto en las de trabajos; después, cada vez menos.³² Es el 2.1 de Strathern con otra escala: cuando la nota del trabajo se aprieta hacia arriba, deja de separar a quien sabe de quien no.

Se entera solo quien sigue examinando sin máquina. En el condado chino, el examen mensual a libro cerrado, hecho en el propio centro, marcó la caída a los seis meses; los exámenes externos de acceso la vieron entera unos dos años después (capítulo 2).¹⁴ Donde no hay un examen frecuente sin máquina, la primera señal fiable es la externa, y llega cuando ya no queda tiempo para corregir. Ese es el daño propio de este capítulo: no que el alumno aprenda menos, que ya se ha visto, sino que la nota que ven a diario el profesor y la familia deja de avisarles.

5.3 La señal que deja de separar

Fuera de la escuela, la nota y el trabajo escrito sirven para otra cosa: que alguien que no conoce al alumno decida si lo admite o lo contrata. Es una señal en el sentido del economista Michael Spence, que explica el capítulo 6 del Informe I: un escrito informa sobre quien lo hace mientras

producirlo le cueste más a quien sabe menos, y deja de separar cuando la máquina iguala ese coste para todos.⁵⁵

La prueba más directa con datos la publicaron en noviembre de 2025 dos economistas, Galdin, de Dartmouth, y Silbert, de Princeton, con 2,7 millones de solicitudes a encargos de programación en la plataforma Freelancer. Antes de ChatGPT, una propuesta bien ajustada a lo que pedía el cliente pesaba mucho: valía tanto como pedir 26 dólares menos, en encargos que se pagaban entre 30 y 250. El ajuste anunciaba el esfuerzo del trabajador, y el esfuerzo, que el trabajo se terminara bien. Con los modelos de lenguaje, esas relaciones «se debilitan de forma significativa o desaparecen del todo». En las propuestas redactadas con el botón que la plataforma añadió en abril de 2023, esfuerzo y ajuste van en sentido contrario, y la propuesta ya no anuncia quién hará bien el trabajo.⁵⁶

Con un modelo estimado sobre los datos anteriores, los autores simulan un mercado en el que el escrito ya no dice nada. Es una simulación, no algo observado: el quinto de trabajadores más capaces sería contratado un 19 % menos, y el quinto menos capaz, un 14 % más. Su conclusión es que, sin señales costosas, a los empleadores les cuesta más reconocer a los mejores, y el mercado se vuelve «significativamente menos meritocrático». Ellos mismos citan las redacciones de admisión en la universidad como el mismo caso.⁵⁶ Es un documento de trabajo sin revisión por pares y mira un solo mercado.

Cuando la señal deja de separar, no pierde el flojo: pierde el bueno, que se distinguía por un escrito que otros no podían producir. La universidad israelí apunta al mismo sitio desde dentro. El alumno del percentil 25, el que solo deja por debajo a un cuarto de sus compañeros, ganó entre dos y tres puntos, el doble que la media.³² Las notas se juntan porque suben las de abajo, y quien sabía por sí mismo queda confundido con quien ha entregado lo mismo sin saberlo.

5.4 Por qué no basta buscar a la máquina en el texto, y por qué la sospecha castiga al bueno

La primera respuesta fue buscar a la máquina dentro del trabajo, con programas detectores o con profesores atentos al estilo. La única prueba

ciega hecha en un sistema real de exámenes es de 2024, y mide a los profesores. En la Universidad de Reading, un equipo metió respuestas escritas enteras por GPT-4 en los exámenes para casa de cinco asignaturas del grado de Psicología. Los correctores, que no sabían que había exámenes de máquina, no vieron el 94 %. De media, esas respuestas sacaron medio tramo de calificación más que los alumnos reales.¹¹ Era un modelo de 2023, y la corrección ya no distinguía.

Lo que sí se detecta crece, pero es poco. The Guardian, con peticiones de información a las universidades británicas, contó los casos probados de copia con IA: pasaron de 1,6 a 5,1 por cada mil alumnos entre los cursos 2022-23 y 2023-24. En el segundo fueron casi 7.000.⁵⁷ La fuente cuenta lo probado, no el uso real, y si la detección se parece a la de Reading, lo probado es una parte pequeña. Los alumnos tampoco esperan que la corrección los distinga: en torno al 42 % de los universitarios británicos cree que un trabajo hecho por IA sacaría buena nota en su carrera, según la encuesta del Higher Education Policy Institute, un centro de estudios británico.³

De los programas detectores no hay en las fuentes de este informe una prueba ciega equivalente. El argumento contra ellos es de la premisa, y vale también para el profesor atento. Buscar a la máquina en el texto es buscar huellas: una referencia inventada, un error que un alumno no cometería, un estilo demasiado liso. Son rasgos de una máquina que todavía falla, y la mejora de los modelos consiste en buena parte en que dejen de aparecer. Es la deducción del capítulo 6 del Informe I: la sanción solo alcanza a quien deja huellas, y la premisa las borra. Si en 2024 los correctores no vieron el 94 %, con la premisa no queda en el texto nada que encontrar.

La segunda respuesta es comprobar después: si un trabajo despierta dudas, se llama al alumno para que lo explique. En la Universidad de Málaga, el delegado del rector para la estrategia digital lo planteaba así en septiembre de 2026: cuando haya dudas sobre un trabajo, prueba de autoría con el estudiante (capítulo 6).⁵⁸ Tiene un defecto de diseño: se pone en marcha por sospecha, y la sospecha la despierta la calidad. Un trabajo flojo no hace pensar en una máquina; uno bien escrito, sí. En la encuesta británica, algunos alumnos lo contaban así: a la gente que escribe bien «a menudo se la acusa» de usar IA.³ Es un testimonio, no

una medida, pero el mecanismo no necesita más: si lo que dispara la comprobación es escribir bien, su coste y el de los errores recaen sobre quien escribe bien. El resultado se parece al de Galdin y Silbert, aunque por otro camino: allí el bueno pierde porque su escrito deja de distinguirse; aquí, porque le hace sospechoso.

La posición de este informe es que ni los detectores ni la prueba de autoría pueden ser la defensa principal: los primeros buscan una huella que la premisa borra, y la segunda carga el coste sobre el alumno que más merecía la nota. Se sigue una regla menor, también deducción del informe: si una institución sigue calificando trabajos hechos fuera, a quién se comprueba debe decidirlo un sorteo, no el estilo. Una defensa oral a unos pocos alumnos tomados al azar no castiga a quien escribe bien y devuelve a la copia un riesgo, que es lo que la hacía cara. No certifica a nadie: lo que certifica tiene que ocurrir en otro sitio.

5.5 La vuelta a la sala

Las instituciones que responden de lo que certifican ya han elegido. La asociación británica de contables colegiados acabó en marzo de 2026 con sus exámenes a distancia, como cuenta el capítulo 6 del Informe I.⁵⁹ En Estados Unidos, las ventas de las libretas azules de examen en papel subieron en dos cursos un 80 % en Berkeley, según *The Wall Street Journal*.⁶⁰ En Inglaterra, Ian Bauckham, director de Ofqual, el regulador de los exámenes oficiales, anunció en julio de 2026 un escrutinio «mucho, mucho mayor» de los trabajos que cuentan para la nota: «Para que los títulos valgan más que el papel en que están escritos, no podemos normalizar la idea de que lo que genera la IA sustituye al esfuerzo humano genuino».⁶¹

La objeción más seria la hizo Arthur Spirling, politólogo de Princeton: es raro prohibir en el examen algo que al alumno le resultará natural durante toda su carrera.⁶⁰ Acierta en el diagnóstico: la máquina estará en casi todo lo que ese alumno haga después. Se equivoca en la conclusión, por una razón concreta: certificar en la sala no es prohibir la máquina. En Sídney, el carril abierto la permite entera, y ahí se aprende a trabajar con ella. Lo que un título promete a un tercero es otra cosa: lo que su

titular sabe por sí mismo, incluido lo que hace falta para juzgar lo que la máquina le da (capítulo 7). La sala no exige quitar la máquina de todo: exige que lo que se califica no lo pueda hacer ella sola. Juzgar, en la sala y delante de alguien, una respuesta suya con errores se puede certificar con la máquina encima de la mesa.

Bauckham va más lejos, y ahí el informe se separa de él. La IA en el aula, dijo, trae «menos aprendizaje, menos satisfacción, menos utilidad en el trabajo» y es, en último término, «un asalto encubierto a la realización del potencial humano».⁶¹ Los datos de 2025-2026 dicen algo más preciso. En el ensayo turco del capítulo 2, en el mismo instituto y en las mismas sesiones de 90 minutos, el chat que daba la respuesta bajó un 17 % el examen posterior sin IA, y el tutor que solo daba pistas lo dejó igual. En Nigeria, con profesor y horario fijo, el tutor lo mejoró (capítulo 4).^{23,42} Lo que decide no es que la IA esté en el aula, sino si hace el ejercicio por el alumno. Es lo que sostiene Justin Reich, citado en el capítulo 2: el problema serio es que los alumnos dejen de hacer los ejercicios en los que vive el contenido. Y pide cautela a quien afirma demasiado: «Nos hemos equivocado mucho con la tecnología en el pasado» (traducción del informe).²⁸ La certeza de Bauckham sobre el aula va más allá de lo que se sabe. Lo que hay que blindar es la certificación.

La posición de este informe es la de Sídney y la de Bauckham en lo que toca a certificar: separar lo que certifica, presencial y hecho por la persona, de lo que enseña, abierto y con la máquina. Goodhart y Strathern explican por qué se rompe una medida que es objetivo: quien la persigue encuentra la forma más barata de alcanzarla. Los dos carriles no lo evitan; hacen que, en lo que certifica, la forma más barata vuelva a ser saber, porque en la sala no hay nada que haga el ejercicio por el alumno. En el abierto, el uso de la máquina se puede enseñar a la vista, porque la acreditación no depende de él. Y el mismo principio, llevado a pruebas cortas y frecuentes sin máquina, devuelve al profesor y a la familia el termómetro que habían perdido, mucho antes del examen externo.

5.6 Qué añade la premisa

Dos de los rasgos de la «IA potente» que el capítulo 1 traducía al pupitre alcanzan a lo que hoy todavía se salva: manejar todo lo que maneja una persona que trabaja a distancia y aceptar encargos de días o semanas.⁸ Un curso a distancia es eso: leer, entregar tareas, escribir en un foro y hacer cuestionarios durante semanas. La deducción es de este informe: con la premisa, un agente puede hacer un curso a distancia entero en nombre de su alumno, y lo que ese curso certifique sin una prueba presencial acredita al agente.

Mientras la IA vive en los servidores de una empresa, la empresa puede apagarla. Es lo que ocurrió en China durante el gaokao, su examen de acceso a la universidad, en junio de 2025, con unos 13,4 millones de candidatos: Alibaba, Tencent y Moonshot suspendieron el reconocimiento de fotos de sus asistentes en las horas de examen.⁶² En el lenguaje del capítulo 3 del Informe I, el proveedor hizo de guardián. Con un modelo que funciona en un móvil o en unas gafas sin conexión, no hay nadie que pueda apagarlo.

La selectividad española de 2026 se preparó con detectores de frecuencias en al menos diez comunidades.⁶³ Un detector de frecuencias encuentra lo que transmite, como un pinganillo conectado a un teléfono. Un modelo que funciona dentro del propio aparato no necesita transmitir nada. También esto es deducción del informe: con la premisa en local, la defensa tiene que pasar de buscar señales de radio a revisar los objetos que entran en la sala.

La defensa física sigue funcionando: una persona sin aparatos, en una sala y delante de alguien, no puede consultar a la máquina. Pero deja de ser gratis. Una sala y un vigilante costaban poco; una sala con detección de aparatos, revisión a la entrada y personal que sepa qué buscar cuesta bastante más.

5.7 El cuello de botella, en el profesor

La brecha de verificación del capítulo 2 del Informe I, producir casi gratis y comprobar no, tiene en la escuela una forma particular. En un ejercicio

de respuesta cerrada, que la respuesta es correcta sí tiene juez automático, el solucionario. Lo que no lo tiene es si la sabe quien la entrega, y comprobarlo exige lo único que no se multiplica: horas de un profesor delante del alumno.

La cuenta es de este informe. Un examen escrito de una hora para treinta alumnos cuesta una hora de vigilancia y unas cinco de corrección, a diez minutos por examen, y esas cinco el profesor las pone cuando puede. Treinta defensas orales de diez minutos cuestan también cinco horas, pero todas dentro del horario, con cada alumno delante, uno a uno. La diferencia no está en el total, sino en dónde caen las horas, y el horario no crece. Además, lo que lleva semanas, como un trabajo de investigación, no cabe en un examen de una hora: solo se certifica pidiendo al alumno que lo defienda o viéndolo trabajar.

El cuello de botella se muda de quien produce a quien verifica. Un sistema que quiera seguir certificando tareas largas tiene tres salidas: pagar esas horas, reducir lo que certifica a lo que cabe en un examen escrito o seguir certificando con lo que ya no mide. La tercera es la más barata, y por eso la más probable allí donde nadie decide otra cosa. El sorteo no es una cuarta: disuade, pero no certifica.

5.8 Por qué estos veredictos

El primero es «muy probable» porque ya se mide en dos países con diseños distintos: en Israel, la nota de lo hecho fuera anuncia cada vez menos la del examen; en China, entre quienes usan la IA, la anuncia al revés. La premisa lo agrava, porque borra las huellas que quedaban, y con dos países en apenas tres cursos, un tercero antes de 2031 es lo esperable.

El segundo, los cursos a distancia hechos por un agente, es «probable». La premisa da al agente lo que ese curso pide, semanas de tareas, foros y cuestionarios, y la asociación de contables ya cerró sus exámenes a distancia por el fraude con IA. No sube más porque alguien tiene que documentarlo, y un curso bien hecho por una máquina no deja rastro.

El tercero, el fraude con un aparato que lleva el modelo dentro, es «probable» porque la selectividad española de 2026 ya se preparó contra relojes y gafas, esos aparatos existen y la premisa los hace capaces sin

conexión. No llega a «muy probable» porque revisar los objetos a la entrada funciona mientras los aparatos se vean.

Las acusaciones injustas no llevan veredicto. El mecanismo está descrito más arriba, pero hoy no hay registro ni encuesta que las cuente, y un suceso que nadie mide no se puede comprobar en 2031.

La contratación y las admisiones menos meritocráticas quedan en «posible». El mecanismo es sólido, pero la prueba es un solo mercado y una simulación, y los empleadores ya buscan señales que el texto no da, como la entrevista en persona del capítulo 6 del Informe I.

VEREDICTO Hasta 2031, si se cumple la premisa

En los estudios que comparan a los mismos alumnos, la nota de lo hecho fuera de la sala predice la del examen vigilado claramente peor que antes de 2022, o al revés, en al menos tres países

■ ■ ■ ■ ■ Muy probable

Grave: el profesor y la familia pierden el aviso; ya ocurre en dos

Casos documentados de cursos o títulos a distancia completados por un agente de IA en nombre del alumno, en al menos dos países

■ ■ ■ ■ ■ Probable

Grave para quien tiene un título a distancia honrado, que pierde valor

Fraude documentado con un aparato que lleva un modelo sin conexión (gafas, reloj, auricular) en un gran examen presencial nacional, de acceso o profesional, en al menos dos países

■ ■ ■ ■ ■ Probable

Moderado: encarece la defensa física

Contratación o admisiones menos meritocráticas porque el escrito deja de separar, medido en un mercado distinto del de Galdin y Silbert

■ ■ ■ ■ ■ Posible

Grave e injusto

Escala de los veredictos, siempre de aquí a 2031. Se explica al final del informe.

QUÉ CAMBIARÍA ESTE JUICIO

Todo el capítulo cambiaría con una forma de acreditar a distancia la autoría y el saber que resista a un modelo local; no hay ninguna a la vista. El primer veredicto bajaría si algún rediseño del trabajo en casa hiciera

que su nota volviera a anunciar la del examen, medido con los mismos alumnos. El fraude con aparatos bajaría si, hasta 2031, los capaces de llevar un modelo potente siguieran siendo fáciles de ver a simple vista. Las acusaciones injustas entrarían en los veredictos si una encuesta seria empezara a contarlas. Y la contratación menos meritocrática subiría a «probable» si un segundo mercado, de empleo o de admisiones, mostrara con datos lo mismo.

— *Donde un examen en la sala pesa mucho, lo que se certifica queda a salvo, aunque no lo aprendido. En España, la mayor parte de lo que certifica a un alumno se decide fuera de ella.*

España certifica sobre todo con lo que ya no mide

60 %

peso de la media del bachillerato en la calificación de acceso a la universidad; el examen externo y presencial pesa el 40 %⁶⁴

0,8

puntos que la media del bachillerato superaba a la de la prueba de acceso, de media, entre 2015 y 2024, casi todo antes de la IA⁶⁵

**54 % y
35 %**

alumnos españoles de 15 años que usan chatbots cada semana para ayudarse a aprender, y profesores de secundaria que dicen haber usado IA en su trabajo^{1, 66}

6.1 Málaga, septiembre de 2026

El 21 de septiembre de 2026, Joaquín Canca, delegado del rector de la Universidad de Málaga para la estrategia digital, resumió en la prensa local lo que ve su universidad: «En casi todos los trabajos hay trazos de IA». El programa antiplagio, que busca coincidencias con textos ya existentes, «venía funcionando bien, pero con la IA ya no ocurre eso». Su propuesta tiene tres partes. La evaluación «tiene que pasar del resultado al proceso». Cuando haya dudas sobre la originalidad de un trabajo, se hará una prueba de autoría con el estudiante. Y habrá más exámenes orales, aunque Canca reconoce que exigen más tiempo.⁵⁸

Dos meses antes, en julio, la universidad había aprobado un reglamento de evaluación que, según la prensa, recoge esa prueba de autoría cuando haya indicios fundados de fraude y prohíbe que una prueba aislada pese más del 70 % de la calificación final.⁶⁷ Al menos un 30 % de cada nota

tiene que salir, por tanto, de otras pruebas, que pueden ser trabajos como aquellos en los que, según Canca, casi siempre hay trazos de IA.

Lo que se entrega hecho fuera de la vista ya no dice quién lo ha hecho, y la herramienta que lo delataba ha dejado de servir. Cada una de las dos salidas de Canca trae el coste que anticipaba el capítulo 5. El examen oral devuelve la prueba a la presencia, pero se paga en horas de profesor. La prueba de autoría es más barata, pero solo se hace cuando hay dudas sobre la originalidad, es decir, cuando alguien sospecha, y la sospecha, como se vio en el capítulo 5, tiende a caer sobre quien escribe bien.⁵⁸

Pero el problema de Málaga es el de cualquiera que evalúe con trabajos. Lo que hace de España un caso aparte está antes: quién termina la secundaria y con qué nota entra en la universidad lo decide sobre todo el propio centro, no un examen externo. La posición de este informe es que por eso la rotura de la medida deja expuesta aquí una parte mayor de la certificación, y que de las respuestas de 2026 solo las que llevan a la presencia la parte que certifica van al sitio del daño.

6.2 Dónde se certifica en España

Antes de la universidad, lo que sabe un alumno español se certifica dos veces, y en las dos manda su centro.

La primera es el título de la ESO, el que cierra la enseñanza obligatoria. Lo concede el propio profesorado. El real decreto que ordena la etapa lo dice así: «Las decisiones sobre la obtención del título serán adoptadas de forma colegiada por el profesorado del alumno o la alumna», al final de una evaluación que la norma quiere «continua, formativa e integradora». No hay examen externo final. Las evaluaciones de diagnóstico sirven para informar y orientar, y no cuentan para el título.⁶⁸

La segunda es la nota con que se entra en la universidad. Hoy la fija otro real decreto, de 2024: la calificación de acceso es la media ponderada «del 60 por ciento de la nota media de Bachillerato» y «del 40 por ciento de la calificación de la prueba de acceso».⁶⁴ En toda la secundaria, el único examen externo y presencial que cuenta para algo es esa prueba, y pesa el 40 % de la calificación de acceso. En las carreras con nota de corte, la nota de admisión puede sumar exámenes de esa prueba, y allí

pesa algo más.

Esa nota del centro ya se separaba de la prueba antes de la IA. Según la verificadora Newtral, con datos del Ministerio de Ciencia de 2015 a 2024, la media del bachillerato superaba a la de la prueba de acceso en ocho décimas, de media, en todos los tipos de centro. La serie es casi toda anterior a los chatbots, que llegaron a finales de 2022. En los privados, una cuarta parte de los alumnos tenía más de un 9 de media en el bachillerato. En la prueba de acceso, el 7 %. Juan Manuel Moreno, catedrático de Educación de la UNED, lo explica por el propio peso de la nota: el 60 % empuja a los centros a mejorar el punto de partida de sus alumnos, y lo que empezó en la privada lo imitó después la pública.⁶⁵

Es la ley de Goodhart en la forma de la antropóloga Marilyn Strathern, presentada en el capítulo 5: cuando una medida se convierte en objetivo, deja de ser una buena medida.^{53, 54} La media del bachillerato es un objetivo porque decide el 60 % de la calificación de acceso, y quien la pone es quien tiene interés en que salga alta. Pero esa inflación dependía de quien corregía. La que añade la IA no depende de nadie: un profesor exigente que corrige con rigor un trabajo impecable le pone buena nota, aunque el alumno no sepa hacerlo. Las dos empujan hacia el mismo lado, y la segunda no se frena con más exigencia.

Con la premisa, todo lo que se evalúa fuera de una sala mide a la máquina (capítulo 5). La nota del centro mezcla dos cosas: exámenes en clase, que siguen midiendo aunque los ponga quien tiene interés en que salgan altos, y trabajo hecho fuera, que ya mide a la máquina. En ese sentido España certifica sobre todo con lo que ya no mide: con una nota que el incentivo había inflado antes de la IA y que ahora lleva dentro una parte que no dice nada del alumno. Esa parte pesa más que en un sistema en que un examen externo decide casi todo, porque de la nota del centro dependen el título de la ESO y el 60 % del acceso. Cuánto más depende de qué proporción de la nota sale de trabajo hecho fuera de clase: no hay datos publicados, y es la cifra que más falta hace.

Hay además un perjudicado que no se ve desde dentro. Como explicó el capítulo 6 del Informe I con la teoría de las señales de Michael Spence, un escrito deja de separar a unos de otros cuando cuesta lo mismo a todos.⁵⁵ Con la nota del centro pasa igual: si sube para todos, el alumno que hizo el trabajo por sí mismo queda igualado con el que no lo hizo, y

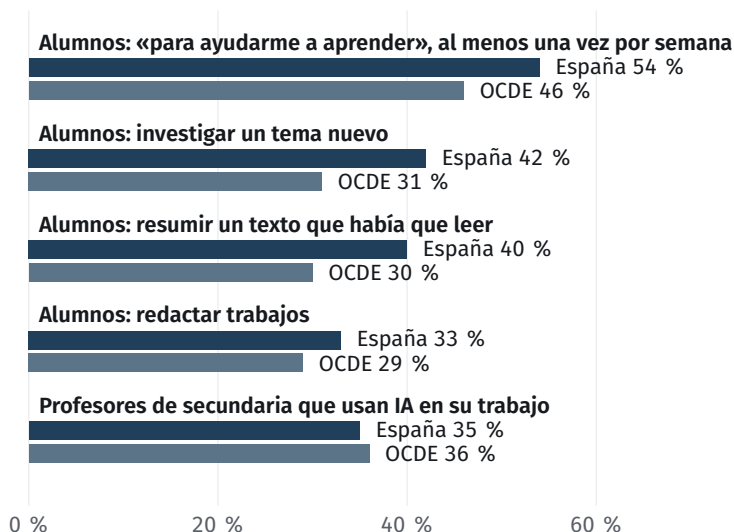
solo los distingue lo que se examina en una sala: en el acceso, el 40 %; en la ESO, lo que el centro evalúa en clase.

6.3 Lo que ya se ve

Los alumnos españoles usan la IA más que la media. Como se vio en el capítulo 1, en PISA 2025 el 54 % de los de 15 años usa chatbots cada semana para ayudarse a aprender. Están por encima de la media en cada uso por el que pregunta la prueba. Cada semana, el 42 % los usa para investigar un tema nuevo y el 40 % para resumir un texto que tenía que leer. Para redactar trabajos, el 33 %.¹

Los dos últimos importan aquí: leer el texto mandado y escribir el trabajo eran tareas de casa que acababan en la nota del centro. Estas cifras dicen cuántos alumnos pasan ya cada semana por la máquina tareas de ese tipo; no qué parte de la nota sale de ellas, ni cuánto aprenden.

Figura 4 Uso de chatbots por los alumnos españoles de 15 años y uso de IA por sus profesores, frente a la media de la OCDE



OCDE, PISA 2025, nota de país de España (alumnos, al menos una vez por semana); TALIS 2024, informe español (profesores de secundaria).

En la universidad, una encuesta de la Fundación Conocimiento y Desarrollo (CYD) a 800 universitarios españoles, de julio de 2026, da un 90 % de uso. El 86 % de quienes la usan cree que mejora su rendimiento académico. El 27 % reconoce que va menos a clase desde que la usa. Y solo el 18 % cree que sus profesores tienen buena formación en IA.⁶⁹ Son respuestas autodeclaradas. La de la asistencia importa porque allí donde ir a clase no es obligatorio, lo que se aprende en ella pasa a depender de la voluntad del alumno, que es lo que el capítulo 3 mostró que falla.

Los profesores españoles la usan como la media. Según la encuesta internacional de la OCDE al profesorado (TALIS), el 35 % de los de secundaria dice haberla usado en su trabajo, frente al 36 % de la OCDE, sobre todo para aprender un tema y preparar clases. De los que no la usan, el 75 % dice no tener formación: casi la mitad del profesorado (la cuenta es de este informe).⁶⁶ Las dos cifras del principio del capítulo no son comparables (una es uso semanal de alumnos en 2025; la otra, uso sin frecuencia de profesores un año antes), así que no dicen quién la usa más.

6.4 Las respuestas de 2026

En 2026 ha habido cuatro respuestas, y lo que las distingue es el sitio al que llegan. La prueba de acceso se blindó en la sala, con detectores de aparatos, como contó el capítulo 5: lo primero que se protegió fue lo que ya se examinaba allí, el 40 %.⁶³

Cataluña fue más allá. Para el curso 2026-2027, su Departamento de Educación, que dirige la consejera Esther Niubó, recomendó que los menores de 14 años no usen herramientas de IA que no estén diseñadas para la educación. Entre los 14 y los 18, con autorización familiar. Pidió también que el trabajo de investigación del bachillerato se defienda de forma oral, para comprobar que es del alumno y que lo entiende. El 22 de septiembre, el presidente Salvador Illa anunció una ley para vetar la IA en las aulas antes de los 14 años: «La consigna es papel, lápiz y libros».²⁰

El Ministerio de Educación actualizó el 14 de septiembre su guía sobre el uso de la IA en la educación. Según la reseña publicada (este informe no ha podido leer la guía entera), trata de alfabetización en IA, privacidad

e integridad académica, y lo que propone para la evaluación es un modelo de «declaración de uso de inteligencia artificial» en los trabajos; la reseña no menciona exámenes presenciales ni orales.⁷⁰ Málaga, como se vio, apuesta por la prueba de autoría y por más orales.

Son tres familias de respuesta: blindar la sala donde ya se examinaba, llevar a la presencia una parte de lo que certificaba el trabajo hecho fuera, y regular la herramienta. Solo la segunda, la de la defensa oral y los orales de Málaga, llega a la parte de la nota que se está degradando.

6.5 Qué respuesta sostienen los datos

La posición de este informe es la de la defensa oral catalana y la de la Universidad de Sídney, cuyos «dos carriles» presentó el capítulo 5: lo que certifica se hace en presencia y sin ayudas, y lo que enseña puede quedar abierto, con la IA dentro.⁵² La defensa oral del trabajo de investigación es eso mismo aplicado a una pieza del bachillerato. El trabajo puede hacerse con la máquina, pero la nota sale de lo que el alumno sabe explicar delante de alguien. Si se hace a todos, y no solo a quien despierta sospechas, no castiga a quien escribe bien, que es el defecto de la prueba de autoría.

La prohibición por edad tiene a su favor la prudencia: como recordaba en junio la Education Endowment Foundation (capítulo 1), la prueba sólida sobre el efecto real de estas herramientas en el aprendizaje apenas ha empezado, sobre todo por debajo de los 16 años.²¹ Pero no llega al daño, por tres razones. El daño mejor medido ocurre en casa: en el condado chino del capítulo 2 subía la nota de los deberes y bajaba la del examen a libro cerrado.¹⁴ Con la premisa, tampoco sirve un filtro por edad en casa, porque un modelo que funciona en un móvil sin conexión no pregunta la edad a nadie (capítulo 3 del Informe I). Y España ya hizo la prueba con el móvil: entre 2022 y 2025, los alumnos en centros que lo prohíben pasaron del 67 % al 86 %, y los resultados en matemáticas y lectura bajaron.¹ La OCDE no encuentra relación, a escala de país, entre los cambios en esas prohibiciones y los del rendimiento.³¹ El móvil no es la IA, pero muestra que prohibir un aparato en el centro no cambia lo que pasa fuera.

La declaración de uso falla por la lógica de Spence: una declaración

solo informa si mentir le cuesta más a quien tiene algo que ocultar, y marcar «no he usado IA» le cuesta lo mismo al que la usó que al que no. Puede servir para enseñar a citar la herramienta. No sirve para certificar.

La respuesta que parece más directa es subir el peso de la prueba de acceso, la parte que sigue midiendo. Tiene un coste que explicaron en 1991 los economistas Bengt Holmström y Paul Milgrom, más tarde premios Nobel, con un principio que ya usó el capítulo 1 del Informe I. Cuando un trabajo tiene varias partes y solo algunas se miden, premiar lo medido hace abandonar lo demás; su propio ejemplo eran los maestros pagados por los resultados de sus alumnos en exámenes tipo test, que dejarían de cultivar la curiosidad, la creatividad y la expresión oral y escrita.⁷¹ Un acceso decidido por un solo examen enseñaría para ese examen y dejaría de lado lo que no cabe en él: escribir un texto largo, defender un argumento en voz alta, trabajar en un laboratorio.

Por eso la posición de este informe no es «más examen», sino más evaluación presencial de tareas variadas: escribir en clase, defender oralmente, hacer prácticas supervisadas. No exige tocar el reparto del 60 y el 40. Exige cambiar de dónde sale la nota del centro: que la parte que certifica, en la ESO y en el bachillerato, se produzca delante de un profesor, y que lo que se hace en casa sirva para aprender y no para calificar.

La presencia corrige lo que añade la máquina, no lo que añade el incentivo: quien examina en clase sigue siendo quien tiene interés en que la nota salga alta. Para eso, lo que se deduce de Goodhart y Strathern es hacer visible la distancia: publicar cada año, centro por centro, cuánto se separa su media de bachillerato de la que sacan sus alumnos en la prueba de acceso.

El coste está en las horas. Una defensa oral cuesta un tiempo que no costaba corregir un trabajo, como admitía Canca en Málaga, y ese coste cae sobre un sistema que ya anda corto. El 49 % de los alumnos españoles de 15 años va a centros cuyo director dice que la falta de profesorado dificulta la enseñanza; en la OCDE, el 39 %.¹ Es el cuello de botella del capítulo 5: comprobar lo que sabe quien entrega el trabajo exige horas de profesor, y esas horas no se multiplican.

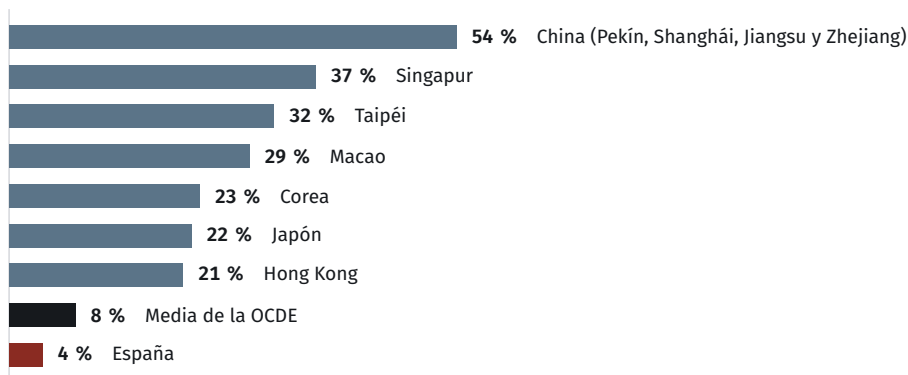
6.6 Dónde va el dinero

El dinero de la gran apuesta educativa española va a otra parte. El anteproyecto de ley presentado en noviembre de 2025 baja el máximo de alumnos por aula de 25 a 22 en primaria. En la ESO, de 30 a 25. El centro de políticas económicas de Esade calcula su coste neto en al menos 2.818 millones de euros al año, de media entre 2027 y 2036. El metaanálisis más completo que cita, con 66 estudios causales, no encuentra un efecto agregado significativo sobre el aprendizaje.⁵¹ Es casi lo que las familias gastan en clases particulares: 2.782 millones en el curso 2023-2024, según Juan Manuel Moreno y Lucas Gortazar. Casi todo el aumento es rebote de la pandemia, y lo que más crece son los idiomas y las enseñanzas artísticas. En secundaria, donde más se usan los chatbots, la proporción de familias que paga clases bajó cinco puntos desde 2019-2020.⁷² La hipótesis de este informe, que los autores no plantean y sus datos no pueden comprobar, es que el chatbot se come la parte de explicar y la academia conserva la de estructura: un adulto que obliga a sentarse dos tardes por semana. Y menos alumnos por profesor son más horas de profesor por alumno, lo que encarece certificar en presencia.

6.7 Una debilidad propia

España no tiene más alumnos por debajo del mínimo que la media de la OCDE: en matemáticas alcanza el nivel básico el 67 %, frente al 65 %. Su déficit está arriba. Solo el 4 % llega en matemáticas a los dos niveles más altos de la prueba, la mitad que en la OCDE.¹

Dos hipótesis hacen pensar que la IA puede apretar justo ahí. En el condado chino del capítulo 2, los que más perdían eran los que iban mejor.¹⁴ Si el mecanismo actúa igual aquí, golpea donde el sistema ya flojea. Y la señal que debería avisar, la nota alta, ya distinguía poco antes de la IA, al menos en la privada, donde una cuarta parte de los alumnos superaba el 9 de media en el bachillerato: es lo que Strathern describió con las notas británicas.⁵⁴ España podría perder alumnos excelentes y, a la vez, la capacidad de verlo en sus propias notas. No hay ningún dato español sobre quién pierde más.

Figura 5 Alumnos de 15 años en los dos niveles más altos de matemáticas en PISA 2025

OCDE, PISA 2025, volumen I y nota de país de España.

6.8 Lo que no se sabe

No hay ningún estudio causal español sobre IA y aprendizaje: solo encuestas, como PISA, la de la Fundación CYD o TALIS. Los datos publicados de la distancia entre la nota del bachillerato y la de la prueba de acceso terminan en 2024, así que nadie sabe todavía si ha crecido. No se sabe qué parte de la nota del centro sale de trabajo hecho fuera de clase. Y no hay datos de cómo certifican lo aprendido las universidades españolas a distancia, como la UNED, la UNIR o la UOC.

6.9 Por qué estos veredictos

El primero es «posible». El mecanismo está medido fuera: en una universidad israelí la nota de los trabajos sube y deja de ordenar a los alumnos, y en China la de los deberes sube mientras baja la del examen.^{14, 32} Y el 60 % ya inflaba la nota antes de la IA, sin exámenes externos que lo frenen. El umbral pide un salto, de ocho décimas a un punto, porque una serie que fluctúa supera su propia media sin que nadie haga nada. No llega a «probable» por tres razones: la media de referencia ya incluye dos convocatorias con chatbots, las de 2023 y 2024; una parte de la media del bachillerato, que nadie ha medido, sale de exámenes en clase, más difíciles de burlar; y ya hay pasos hacia la presencia, como la defensa oral

que recomienda Cataluña. Si ocurre, no bastará para atribuirlo a la IA: la distancia también puede moverse porque cambie la prueba de acceso.

La pérdida de alumnos excelentes es «posible»: descansa en una hipótesis con un solo apoyo directo, el condado chino. Una tendencia previa la haría más fácil: entre 2015 y 2025, un periodo en su mayor parte anterior a la IA, el rendimiento español cayó más entre los alumnos favorecidos que entre los desfavorecidos.¹ Pero con una proporción tan pequeña, la OCDE solo dará por significativa una bajada grande, y si llega, PISA 2029 no permitirá separar lo que añade la IA de lo que ya venía.

Prohibir por edad o pedir declaraciones de uso no lleva veredicto, porque son políticas y no daños; el daño que producen si se quedan solas es el del primer veredicto. Tampoco lo lleva el título de la ESO, y no por falta de riesgo. No hay ningún examen externo con efectos; los que hay, las evaluaciones de diagnóstico y PISA, no se cruzan con el título, y este informe no tiene la serie de titulación del Ministerio con la que compararlos.

VEREDICTO Hasta 2031, si se cumple la premisa

La distancia entre la media del bachillerato y la de la prueba de acceso llega a un punto de media en las convocatorias de 2026 a 2029, según el Ministerio de Ciencia

■ ■ ■ ■ ■ Posible

Grave: es el 60 % de la calificación de acceso

La proporción de alumnos españoles en los dos niveles más altos de matemáticas baja en PISA 2029, respecto de 2025, con una diferencia que la OCDE da por significativa

■ ■ ■ ■ ■ Posible

Grave y lento: es donde España ya flojea

Escala de los veredictos, siempre de aquí a 2031. Se explica al final del informe.

QUÉ CAMBIARÍA ESTE JUICIO

El primer veredicto bajaría a «improbable» si los datos del Ministerio de Ciencia de 2025 a 2027 muestran que la distancia no crece, o si una reforma lleva a la presencia la parte de la nota del centro que certifica. Subiría a

«probable» si en 2025 y 2026, las primeras convocatorias de alumnos que hicieron todo el bachillerato con chatbots a mano, la distancia ya supera las nueve décimas. El segundo bajaría con pruebas de que los mejores alumnos ganan con la IA cuando la usan con estructura, como ganaban más los que iban mejor en el ensayo de Nigeria del capítulo 4.⁴² Y la posición sobre la prohibición por edad cambiaría con un estudio español, con grupo de comparación, en que redujera la pérdida que ocurre en casa.

— *Blindar la certificación salva la nota, no el porqué de aprender. Si la máquina sabe, queda la pregunta de fondo: para qué aprender lo que ella sabe.*

Para qué aprender lo que la máquina sabe

—15 puntos

acierto, por debajo de quien no tiene IA, de quienes la consultan cuando se equivoca; aceptan su error en torno al 80 % de las veces³⁷

**59 % y
90 %**

aciertos en decisiones de juicio estratégico, donde no hay juez automático, de los agentes internos de los grandes laboratorios (cerca del azar, según METR) y de los investigadores humanos¹⁵

1,8 %

conversaciones de universitarios con Claude en que la máquina se limita a recordar; en el 39,8 % crea y en el 30,2 % analiza¹²

7.1 Cuando la máquina se equivoca

Los experimentos de Shaw y Nave, de la escuela de negocios Wharton, que el capítulo 3 usó para explicar por qué el alumno elige la respuesta, dejan un dato que aquí es el centro. Sus 1.372 adultos resolvían problemas en los que la primera respuesta que viene a la cabeza es la equivocada, con una IA a mano que los investigadores hacían acertar o fallar. Cuando la máquina fallaba, su acierto caía 15 puntos porcentuales por debajo del de quienes no tenían IA: tenerla al lado les hacía rendir peor que no tenerla. Aceptaron su respuesta equivocada en torno al 80 % de las veces, y su confianza subió igualmente.³⁷

Las cifras vienen de la nota de la propia Wharton y del blog del repositorio en que se publicó el trabajo, no del artículo, que no ha pasado revisión por pares. Son adultos que responden por internet a problemas

cortos, y lo que se mide es una decisión en el momento, no lo que alguien aprende. Lo que muestra es que consultar a la máquina ocupa el lugar de la deliberación: quien no se para a comprobar se lleva el error de la máquina y, además, más confianza. La rendición bajaba cuando se pagaba por acertar y cuando cada respuesta se corregía en el acto (capítulo 3). En estos problemas bastaba con pararse a pensar. En las tareas que piden saber, que son casi todas las del estudio y el trabajo, comprobar exige además un conocimiento propio con el que contrastar; esa deducción es de este informe.

La pregunta de fondo de este informe es para qué aprender lo que la máquina ya sabe. Si se cumple la premisa, saber deja de hacer falta para producir: la redacción, el problema resuelto o el programa salen igual de bien sin él. Sigue haciendo falta para comprobar lo que se produce, y la posición de este capítulo es que ese es ahora su uso principal. El problema es que ese saber se forma haciendo lo mismo que la máquina permite dejar de hacer.

7.2 La ironía de Bainbridge

En 1983, Lisanne Bainbridge, psicóloga que estudiaba a los operadores de procesos industriales automatizados, publicó un artículo breve que sigue siendo la referencia de su campo. Su idea tiene dos partes. Automatizar no quita a la persona todo el trabajo: le deja justo el más difícil, vigilar a la máquina y tomar el mando cuando falla. Y a la vez le quita la práctica que haría falta para hacerlo, porque las destrezas que no se usan se deterioran. De ahí el título, «Ironías de la automatización», y una de sus frases centrales: «cuanto más avanzado es un sistema de control, más crucial puede ser la aportación del operador humano» (la traducción es de este informe, como la de las demás citas en inglés de este capítulo).⁷³

Bainbridge explicó también por qué esa práctica no se sustituye con explicaciones. El conocimiento que permite resolver lo que nadie había previsto, escribió, «se desarrolla solo con el uso y con información sobre su eficacia», y quien lo recibe en clases teóricas, sin ejercicios prácticos adecuados, probablemente no entenderá gran parte de él.⁷³ Su ejemplo viene de la escuela: cualquier asignatura que uno aprobó en un examen y

en la que no ha vuelto a pensar. Hacen falta las dos cosas, usar lo que se sabe y enterarse de si funcionó, y una buena explicación no da ninguna de ellas.

La idea sirve aquí porque describe un sistema en el que la máquina hace bien casi todo y la persona sigue respondiendo de lo que falla, que es el mundo de la premisa. El operador de Bainbridge es, en este informe, el adulto que mañana revisará en su trabajo lo que produce la máquina; los estudios son donde se formaba. Quien estudia todavía no vigila a la máquina: entrega lo que ella hace, y quien corrige es el profesor. Lo que la máquina le quita son las dos cosas con las que se formaba el futuro vigilante. Quien pega los deberes en un chatbot no usa lo que sabe y nunca se entera de si su propio camino habría funcionado.

El experimento de Shaw y Nave no mide esa pérdida, que tarda años en notarse, sino el paso anterior: dejar de deliberar cuando hay una respuesta a mano. La ironía de Bainbridge describe lo que ocurre si ese paso se repite curso tras curso: lo que no se usa no llega a formarse.

7.3 Con la premisa, la ironía se agudiza

La premisa tiene dos mitades, como mostró el capítulo 1 del Informe I: la de los problemas con respuesta comprobable, donde la IA ya hace trabajo de élite, y la del criterio, donde no hay forma automática de saber si se acertó. De la segunda solo hay una medida publicada, la de METR que citaba el capítulo 1: en decisiones de juicio estratégico dentro de su propia investigación, los agentes internos de los cuatro grandes laboratorios acertaron el 59 %, un resultado que METR describe como cercano al azar, y los investigadores humanos, el 90 %.¹⁵ Es una medida sola, sin serie y de un único campo, pero es la que hay. La ironía de Bainbridge se agudiza en las dos mitades, por lados distintos.

En la mitad comprobable, por el lado de la práctica. El capítulo 2 mostró que la premisa borra el último aviso que quedaba, el error de la máquina. Si la respuesta copiada es correcta, lo propio nunca se pone a prueba, y la información sobre su eficacia, que Bainbridge ponía como condición, solo puede llegar en un examen sin máquina.

En la mitad del criterio, por el lado de la responsabilidad. Ahí la

máquina sigue fallando, así que la persona sigue haciendo falta, y justo para lo más difícil: es lo que mide la distancia entre los agentes y los investigadores de METR. Suponer que se repite en otros terrenos sin juez, como decidir si un análisis se sostiene o si una fuente merece crédito, es una extensión de este informe.

Unir las dos mitades da la deducción central de este capítulo. Es una hipótesis de este informe, sin estudio que la pruebe: el criterio para lo que no tiene juez se forma practicando lo que sí lo tiene. Comprobable quiere decir aquí lo mismo que en el Informe I, una respuesta que se puede cotejar con una solución: el resultado de un problema, un programa que pasa sus pruebas. La idea es que quien ha resuelto muchos problemas de un tipo reconoce una solución falsa de ese tipo, y quien ha escrito programas que fallan sabe dónde mirar en uno ajeno. El ejercicio escolar, que tiene solución, era donde se practicaba el criterio para lo que no la tiene. La premisa automatiza ese ejercicio y deja sin automatizar lo que se practicaba en él.

Hay un dato compatible, aunque mide fundamentos y no criterio. Los universitarios israelíes de Hausman y sus colegas, de los capítulos 2 y 5, que tuvieron IA desde el primer curso rinden mejor donde se puede usar y, escriben los autores, «no mejor, y a veces peor, en las asignaturas avanzadas con examen». Lo resumen así: «la práctica con IA da destreza con IA, pero puede desplazar los fundamentos».³² Y recomiendan que los alumnos lleguen a experimentar lo que es presentarse sin preparación por haberse apoyado demasiado en la IA.³² Es, sin nombrarla, la información sobre la eficacia de Bainbridge. El estudio que mida el criterio mismo no existe; el final de este capítulo dice cuál sería.

Es la mudanza del cuello de botella del capítulo 2 del Informe I, ahora dentro de cada persona: lo escaso deja de ser producir y pasa a ser el criterio para comprobar lo producido, y ese criterio se formaba produciendo.

7.4 Qué hace falta saber para verificar

La objeción habitual en el debate educativo va en sentido contrario: si la máquina recuerda, calcula y redacta, la escuela puede dejar de pedir

memoria y dedicarse a lo que la máquina no hace, pensar con criterio, hacer buenas preguntas, juzgar lo que se lee. Se oye a menudo, aunque este informe no ha encontrado a ningún investigador que la defienda con datos.

Los datos de uso muestran que hoy ese reparto no se da. El análisis de Anthropic que citaba el capítulo 1, sobre 574.740 conversaciones de cuentas con correo universitario con su modelo, Claude, las clasificó también según la función que hacía la máquina. En el 39,8 % creaba y en el 30,2 % analizaba. Solo en el 1,8 % se limitaba a recordar.¹² Son datos de la propia empresa, de abril de 2025. La máquina hace sobre todo las funciones altas, justo las que la objeción quería reservar a la persona. Eso no decide qué debería enseñar la escuela, pero muestra que el reparto prometido, la máquina recuerda y la persona piensa, no sale solo.

La respuesta de fondo es otra. Barbara Oakley, Terrence Sejnowski, neurocientífico del Instituto Salk, y otros tres autores sostienen en un ensayo de 2025 que el criterio descansa en el conocimiento propio. Para evaluar y guiar lo que produce la IA hacen falta esquemas internos, conocimiento organizado que se construye recuperando de memoria, corrigiendo errores y relacionando lo aprendido. Herramientas como ChatGPT, escriben, pueden «cortocircuitar los procesos de recuperación, corrección de errores y construcción de esquemas», y «a medida que las ayudas externas se vuelven más capaces», la memoria propia corre el riesgo de atrofiarse.⁷⁴ Es un ensayo teórico, no un experimento, y el informe lo toma como hipótesis. Encaja con los universitarios israelíes, a los que la práctica con IA les desplazaba los fundamentos. Y añade una razón a la de Shaw y Nave: allí la gente se rendía aunque bastara con pararse a pensar; donde además hace falta saber, quien no tiene esquemas propios no tiene con qué comparar.

En PISA 2025, las destrezas lectoras que más caen son las que la OCDE llama decisivas en la era de la IA: evaluar información, relacionar varias fuentes y pensar críticamente. Los «lectores apresurados», los que leen deprisa y responden rápido y mal, pasaron de casi el 7 % al 11 % de media en la OCDE entre 2018 y 2025, lo que la OCDE resume como que casi se duplicaron.^{29,30} La caída no se puede atribuir a la IA, por las razones que da el capítulo 2: empezó antes y la OCDE no la nombra al explicarla.^{29,30} Lo que sí dice el dato es que la generación que tendrá que verificar a

la máquina tiene ya más débiles justo esas destrezas. Si la máquina las debilita más es, por ahora, una hipótesis.

De aquí no se sigue que haya que volver a memorizarlo todo, ni que la IA estorbe siempre para aprender. El ensayo de Lira y otros del capítulo 2 muestra que ver un buen ejemplo hecho por la máquina y tener que reproducirlo después sin ayuda enseña, al menos a adultos motivados en una tarea corta.²⁷ Y el saber que hace falta para verificar no es todo lo que la máquina sabe. Es, en la propuesta de este informe, lo suficiente para hacer uno mismo, despacio y a pequeña escala, lo que se le va a encargar a ella.

7.5 La generación siguiente

Bainbridge escribió en 1983 una frase que hoy se lee al pie de la letra. Los sistemas automáticos de su época los vigilaban antiguos operadores manuales, que iban «montados sobre sus destrezas, que no cabe esperar que tengan las generaciones siguientes de operadores».⁷³ Quienes hoy comprueban lo que hace la IA, sean los investigadores del 90 % de METR, los profesores que corrigen o los programadores veteranos que revisan código, aprendieron su oficio sin ella. Verifican con un saber que formaron haciendo lo que ahora hace la máquina. Los que vienen detrás tendrán que formarlo de otra manera, o no lo tendrán.

Hasta ahora había un segundo lugar donde se formaba: el primer empleo, en el que el principiante aprendía haciendo el trabajo menor que le encargaban. El capítulo 3 del Informe II estudia si ese primer empleo escasea. En Estados Unidos, el laboratorio de economía digital de Stanford sigue el empleo de los jóvenes de 22 a 25 años en las ocupaciones más expuestas a la IA. Frente a sus iguales de ocupaciones poco expuestas, la brecha ha llegado al 19 %, abierta por menos contratación y no por despidos.⁷⁵ El dato es descriptivo: no prueba que la causa sea la IA, y el Informe II discute por qué. En España, las ofertas para programadores júnior cayeron un 33 % en un año, mientras las de sénior subían un 13 %; son ofertas publicadas, no contrataciones.⁷⁶ Seyed Hosseini y Guy Lichtinger, en un documento de trabajo, advierten de que menos empleo júnior puede «erosionar la oferta futura de la misma pericia que la IA

complementa». ⁷⁷

Si la causa es la que sospecha el Informe II, el ejercicio escolar y el primer empleo, los dos sitios donde se practicaba el criterio, se estrechan a la vez y por la misma razón: la máquina hace el trabajo comprobable con que se aprendía. La unión es de este informe, y los datos de empleo cubren solo dos o tres años: no dicen si esos jóvenes encontrarán otro sitio donde aprender el oficio.

7.6 Ya no hacer falta

Todo lo anterior supone que la persona sigue haciendo falta para verificar. El capítulo 9 del Informe I estudió la posibilidad contraria con la pérdida gradual de poder de Jan Kulveit y otros: los grandes sistemas sociales sirven a las personas en parte porque necesitan su participación, y si dejan de necesitarla, sostienen, pueden debilitarse tanto los controles explícitos, como el voto, como esa alineación implícita. ⁷⁸ Luke Drago y Rudolf Laine llegan a lo mismo por otro camino: quien vive de la IA pierde el incentivo a invertir en las personas. ⁷⁹

Aplicado al saber, el mecanismo tiene dos caras. Si saber deja de hacer falta para producir, es de esperar que baje el incentivo para adquirirlo, tanto para quien estudia como para quien paga su formación, que es el caso de Drago y Laine. No hay datos de que esa expectativa reduzca ya el esfuerzo de los estudiantes. La otra cara llegaría si la máquina verificara también con fiabilidad en la mitad del criterio. Vigilarla sería entonces lo que Bainbridge llamó, para quien vigila un sistema que decide porque el juicio humano no basta, «una tarea imposible». ⁷³ A la población le quedaría juzgar los efectos de esas decisiones y a quienes despliegan la máquina: leer un dato, pesar un argumento, notar una promesa sin prueba. Es el saber que forma la escuela, y sin él el control que Kulveit y otros ven en la participación se queda sin quien lo ejerza.

Es una hipótesis, y el informe la sitúa en el tramo más bajo hasta 2031. La pérdida sería gradual y podría verse venir en los exámenes externos y en la contratación. Nadie ha medido si lo perdido se recupera cuando vuelve el examen presencial de los capítulos 5 y 6, pero ese examen devuelve al menos el motivo para practicar. Y, sobre todo, quienes aprendieron sin IA

seguirán trabajando décadas. La tesis es de Kulveit y de Drago y Laine; lo que añade este capítulo es su cara educativa, que no tiene datos propios.

7.7 Con quién está el informe

La posición de este informe es la del principio del capítulo. Está con Bainbridge en el mecanismo y con Oakley y Sejnowski en la tesis, como hipótesis con apoyo parcial en los datos: adultos que se rinden cuando la máquina falla, universitarios a los que se les resienten los fundamentos, agentes que todavía fallan donde no hay juez.^{15, 32, 37} Y está contra el «ya no hace falta memorizar», porque lo que se delega no es la memoria y porque, si Oakley y Sejnowski tienen razón, sin conocimiento propio no hay criterio.

Ian Bauckham, que dirige Ofqual, el regulador de los exámenes de Inglaterra, y cuyo juicio sobre la IA en clase discute el capítulo 5, dijo en julio de 2026 otra frase que aquí encaja mejor. Hablaba de corregir con IA, que le parece útil aunque se equivoca, se inventa cosas y arrastra sesgos: «la IA tendrá un lugar importante, pero tenemos que seguir al mando».⁶¹ Seguir al mando exige justo el saber que defiende este capítulo. Justin Reich, el investigador del capítulo 2, pide a las escuelas humildad ante lo que aún no se sabe: «corramos hacia respuestas correctas, no hacia las primeras».²⁸

La prueba de Dylan Wiliam que recogió el capítulo 3, preguntarse antes de delegar «¿es esto algo en lo que quiero mejorar?», necesita con la premisa una segunda pregunta: ¿voy a tener que comprobar esto?⁴¹ A los 14 años nadie lo sabe, porque nadie sabe todavía en qué trabajará, y por eso esa respuesta la tiene que dar el currículo, no el alumno.

Es la tesis menos medida de este informe. No existe ningún estudio que compare cómo verifican, en tareas sin juez automático, quienes aprendieron con IA y quienes aprendieron sin ella. Lo que el capítulo propone como formación del que verifica, en cambio, se arregla con estructura: práctica sin máquina, examen en presencia y un primer empleo en el que se aprenda haciendo. Las tres cosas se deciden en casa, en el aula y en quien contrata.

7.8 Por qué estos veredictos

El primero es «posible» y no «probable» porque hay mecanismo y datos parciales, pero ninguna medida directa: Shaw y Nave miden a adultos en decisiones de minutos, y Hausman y sus colegas, fundamentos que se resienten, no la capacidad de detectar errores. El suceso exige además que alguien haga esa comparación antes de 2031, y hoy nadie la ha anunciado. No es «improbable» porque todo lo medido apunta en esa dirección, la premisa lo agrava y la comparación es sencilla de montar con universitarios de la misma edad.

El segundo es «muy improbable» hasta 2031 sobre todo porque quienes hoy revisan el trabajo de la máquina aprendieron sin ella y seguirán en activo mucho después de esa fecha. La pérdida, además, sería gradual y se vería antes en los exámenes y en la contratación. Nadie ha medido si la vuelta al examen presencial la corrige. El riesgo crece cuando esa generación empieza a retirarse, más allá del horizonte de este informe.

VEREDICTO Hasta 2031, si se cumple la premisa

Un estudio con grupo de comparación o una evaluación externa encuentra que jóvenes formados con la IA a mano detectan peor los errores de la IA en su propio campo que otros de la misma edad formados antes de 2023

■ ■ ■ ■ ■ **Posible**
Grave; se nota tarde

Un regulador o un colegio profesional de un país rico reconoce que ya no tiene personas formadas para revisar el trabajo de la IA en su campo y deja de revisarlo

■ ■ ■ ■ ■ **Muy improbable**
Muy grave; difícil de revertir

Escala de los veredictos, siempre de aquí a 2031. Se explica al final del informe.

QUÉ CAMBIARÍA ESTE JUICIO

Subiría el primer veredicto que otros estudios como el de Hausman y sus colegas, fuera de Israel, encontraran que la práctica con IA desplaza los fundamentos en exámenes sin máquina, o que PISA 2029 registrara otra caída en evaluar información y relacionar fuentes. Lo bajarían pruebas

de que practicar con IA y examinarse después sin ella deja los mismos fundamentos que practicar sin máquina; el fondo de la Education Endowment Foundation, que estudia qué pasa cuando se delega el esfuerzo en la máquina, dará sus primeros resultados en 2027.²¹ Y el capítulo cambiaría de eje si la máquina pasara a verificarse a sí misma con fiabilidad medida en los terrenos sin juez, con la distancia de METR cerrándose como señal. Verificar dejaría entonces de ser el uso principal del saber propio, y el peso pasaría a la pérdida de poder del capítulo 9 del Informe I.

— *Casi todo lo que se deduce en este informe se corrige con estructura, y la mayor parte de esa estructura depende de decisiones pequeñas de quien estudia, de quien enseña y de quien paga.*

Qué hacer: un sitio, una hora y alguien que compruebe

—17 %

aciertos en un examen sin IA, frente al grupo de comparación, de los alumnos turcos que habían practicado con un chat como ChatGPT; no notaron que rendían peor²³

de 7,2 a
47,4

minutos semanales de práctica en internados de la India cuando unos coordinadores protegen el tiempo; el aprendizaje sube con claridad, sin IA generativa⁴⁰

94 %

exámenes escritos por la IA que pasaron sin ser detectados en los exámenes para casa de Psicología de la Universidad de Reading, en 2024¹¹

8.1 Dos alumnos con la misma máquina

En un condado del centro de China, entre 2022 y 2025, la mayoría de los alumnos de secundaria empezó a hacer los deberes con un chatbot (capítulo 2). De cada cinco que lo usaban, cuatro los terminaban en menos de cincuenta minutos, con notas altísimas, y en el examen mensual a libro cerrado rendían claramente peor. El quinto, con herramientas parecidas, seguía dedicando a los deberes el tiempo de antes, y perdía poco.¹⁴ Los autores advierten que ese quinto puede diferir de los otros en motivación, y que imponer el tiempo no garantiza recuperar lo perdido. Que el tiempo protege se apoya mejor en los internados de la India del capítulo 4, donde lo impusieron unos adultos y el aprendizaje subió.^{14, 40}

Ninguno de los dos podía saberlo por su nota de deberes, que era buena en ambos casos, y quien pierde tampoco lo nota por sí mismo (capítulo 2).

El capítulo 2 del Informe I distingue tres clases de defensa contra un daño. Las defensas por coste nadie las diseñó: son el precio de hacer daño. Las físicas dependen de un lugar, un cuerpo o un objeto. Las de respuesta, de alguien que vigila y reacciona. En la escuela, el esfuerzo de hacer los ejercicios era una defensa por coste doble. Protegía de la ignorancia propia, porque para tener el ejercicio hecho había que hacerlo, y del engaño ajeno, porque copiar exigía un cómplice o un riesgo. Con la máquina caen las dos a la vez: el alumno puede no aprender, y ni él, ni su familia ni su profesor se enteran a tiempo. Las otras dos clases siguen en pie: ningún modelo abarata un sitio o una hora, ni impide que alguien pida al alumno que explique de palabra lo que ha hecho. La sala, eso sí, ya tiene que buscar dispositivos (capítulo 5).

Donde cayó una defensa por coste, no se trata de volver a encarecer el ejercicio o la copia, que ya no se puede, sino de reponer una física o una de respuesta.

8.2 Si estudia

Separe lo que aprende de lo que entrega. Para lo que quiere saber, siga el orden que propone la OCDE: primero sin IA, después con una herramienta hecha para enseñar y solo al final con la generalista.²⁵ Lo que sale bien con ayuda dice poco de lo que queda después (capítulo 2). En unos ensayos de 2025, adultos que aprendían a escribir cartas de presentación escribieron mejor sin ayuda después de ver la versión corregida por la IA.²⁷ Eran adultos y una tarea corta, y la lectura es de este informe: lo que enseña parece ser ver un buen ejemplo y tener que reproducirlo después, y esa es la mitad que se salta quien copia. Pida el ejemplo resuelto, ciérrelo y rehágalo en una hoja en blanco: la hoja es la defensa de respuesta, porque no se deja engañar.

Examínese sin la máquina, y a menudo. Es el único termómetro que le queda, porque la sensación de haber entendido es justo lo que la máquina mejora. En los experimentos de Shaw y Nave del capítulo 3, la confianza de quienes consultaban a la IA subía incluso cuando les llevaba al error. La rendición bajaba cuando se pagaba por acertar y cuando cada respuesta se corregía en el acto.³⁷ Eran adultos decidiendo en el momento,

no alumnos aprendiendo, y la traducción es de este informe: un examen breve sin ayuda da las dos cosas. Es una defensa de respuesta que puede ponerse usted mismo: exámenes de otros años en papel y con reloj, o un compañero que le pregunta en voz alta.

Decida antes dónde y cuándo trabaja sin la opción fácil. Casi nadie usa el tutor que no da la respuesta porque el esfuerzo se paga hoy y lo aprendido se cobra dentro de meses: es el sesgo hacia el presente del capítulo 3. Contra él no sirve saber lo que conviene, sino atarse antes de que llegue la tentación, lo que la economía del comportamiento llama un dispositivo de compromiso.³⁸ Elija un sitio y unas horas en que la máquina que resuelve no esté a mano: una biblioteca, una sala de estudio, el portátil en otra habitación. Tiene que ser algo físico: el modo tutor es un ajuste, y un ajuste se quita en un clic (capítulo 3). Mejor aún si alguien le espera a esa hora, como en los internados de la India (capítulo 4).

Mire las horas, no la nota. La nota de lo que hace con la máquina mide sobre todo a la máquina, también para usted (capítulo 5). Lleve la cuenta de las horas que dedica cada semana a lo que quiere saber. Si bajan mientras la nota sube, lo que ha ahorrado es práctica.

Antes de abrir el chat, hágase la pregunta de Wiliam (capítulo 3): ¿es esto algo en lo que quiero mejorar?⁴¹ Si la respuesta es sí, hágalo usted; si es no, delegue sin culpa. Hágasela antes de empezar, porque a mitad de la tarea, cuando cansa, contesta por usted el sesgo hacia el presente. Y cuente entre lo que quiere mejorar lo que le hará falta para comprobar a la máquina en su campo: es la ironía de Bainbridge del capítulo 7, y a usted nadie le obligará a aprenderlo.⁷³

Cada etapa pesa distinto. En la ESO, el sitio y la hora los decide sobre todo la familia, y la sección siguiente también le toca. En la universidad, use el carril abierto (capítulo 5) para aprender a manejar la herramienta y prepare el seguro como si no existiera. Ante un examen de memoria y problemas, sea la selectividad o una oposición, pesan sobre todo los simulacros en sus condiciones reales.

La posición de este informe es que quien estudia no necesita renunciar a la IA, sino cambiar el orden y el sitio: primero sin ella y ante algo que no se deje engañar, después con ella. Cambiaría este consejo un ensayo grande en el que un tutor de IA usado libremente, sin nadie que organice el tiempo, mejorara de forma duradera al alumno medio. Ninguno de los

pendientes lo mide: el de Eedi y Google DeepMind se hace en clase, con un humano que revisa cada mensaje (capítulo 1).

8.3 Si tiene hijos

Pague estructura, no explicaciones. La explicación ya es gratis, y con la premisa será mejor que la de cualquier profesor (capítulo 1). El efecto de la tutoría humana cae con la escala, y estar en persona, en horario escolar y varias veces por semana lo amortigua en parte (capítulo 4).⁴³ La lectura es de este informe: lo que limita no es lo que sabe el tutor, sino las horas y la regularidad con que el alumno se sienta delante de él. Si paga clases particulares o una academia, pregunte menos si explican bien y más si su hijo trabaja allí delante de alguien y si le comprueban sin máquina lo que sabe. La hipótesis de este informe es que eso es lo que muchas familias españolas ya compraban: alguien que obliga a sentarse dos tardes por semana (capítulo 6).

Acuerde con él el sitio y la hora antes de que hagan falta. A los catorce años el problema no es contestar bien a la pregunta de Wiliam, sino sostener la respuesta: un compromiso que uno se pone a sí mismo se rompe fácil, y la economía del comportamiento parte de «la propensión de los jóvenes a tomar malas decisiones a largo plazo» (capítulo 3; traducción del informe).³⁸ Por eso el pacto sobre dónde y cuándo se hacen los deberes, y sin qué, conviene que lo ponga un adulto: es una defensa física, y funciona mejor que un filtro por edad, que no alcanza a un modelo instalado en el teléfono. Si a esa hora no hay nadie en casa, busque un sitio con un adulto: una sala de estudio vigilada, una biblioteca, el refuerzo de tarde del centro.

Pídale que rehaga delante de usted un ejercicio de los deberes. El cuaderno ya no informa: en el condado chino, entre quienes usaban la IA, mejor nota en los deberes anunciaba peor examen (capítulo 5).¹⁴ No necesita saber la materia: le basta con ver si lo resuelve sin el teléfono y si llega al mismo resultado que entregó. Si se atasca en lo que entregó perfecto, ya sabe lo que la nota no le decía. Es una defensa de respuesta, y usted es quien comprueba.

8.4 Si enseña

Separe lo que certifica de lo que enseña. La Universidad de Sídney lo hace desde 2025 con dos carriles: las evaluaciones seguras, presenciales y sin ayudas, certifican; las abiertas, donde se permite todo, IA incluida, enseñan (capítulo 5).⁵² El carril abierto responde a la objeción de Arthur Spirling que discute el capítulo 5: certificar sin la máquina no es prohibirla.⁶⁰ La defensa del carril seguro es física: una sala y una hora.

Haga pruebas cortas y frecuentes sin máquina, y que no sean todas iguales. Le devuelven el termómetro que perdió la nota de lo hecho en casa, que dejó de medir al volverse objetivo: es la ley de Goodhart del capítulo 5.^{53,54} Y le dan al alumno las dos cosas de «Examínese». No cargue todo el peso en un examen final: premiar solo lo que se mide hace abandonar lo demás, que es la cautela de Holmström y Milgrom (capítulo 6).⁷¹ Escribir en clase, resolver un problema en la pizarra o hacer una práctica supervisada miden cosas distintas, y todas ocurren a la vista.

Pida que se defienda de palabra lo que se entrega por escrito. Es lo que Cataluña ha pedido para el trabajo de investigación de bachillerato (capítulo 6).²⁰ Es una defensa de respuesta, y cuesta horas de profesor que corregir un trabajo en casa no costaba (capítulo 5). Por eso conviene elegir: menos trabajos escritos, y defendidos, valen más que muchos que nadie puede comprobar.

Ponga la IA detrás de usted, no delante del alumno. Los usos con mejores pruebas son los del adulto. Preparar clases con ella ahorró un 31 % del tiempo en un ensayo inglés. Sugerir en directo a un tutor poco experimentado qué preguntar mejoró la prueba de cada lección, aunque no el examen de fin de curso (capítulo 4).^{45,46} La deducción es de este informe: el tiempo que la máquina le ahorra al preparar es el que le hace falta para comprobar en persona.

No se fíe de los detectores ni de su olfato. Los correctores de Reading ya no distinguían en 2024 lo escrito por la máquina, y la premisa borra la última huella que quedaba, el error (capítulo 5).¹¹ La sospecha, además, cae sobre quien escribe bien: algunos universitarios británicos cuentan que a la gente que escribe bien «a menudo se la acusa» de usar IA (traducción del informe).³ Si la nota sale de lo que usted ha visto hacer, no necesita sospechar de nadie. Y si todavía califica trabajos hechos fuera, decida por

sorteo a quién pide defenderlos de palabra, no por el estilo: así no castiga a quien escribe bien y la copia vuelve a tener un riesgo (capítulo 5).

La posición de este informe es la de Sídney, y no la de la prueba de autoría a posteriori, que depende de sospechar. Cambiaría si apareciera una forma de acreditar a distancia quién hizo un trabajo y qué sabe que resistiera a un modelo local. No hay ninguna a la vista.

8.5 Si decide sobre un centro o un presupuesto

Proteja tiempo de trabajo dentro del horario. En la India, proteger el tiempo subió el aprendizaje sin IA generativa, y en la tutoría a gran escala el horario escolar amortigua en parte la caída del efecto (capítulo 4).^{40, 43} Con la premisa, la escuela se hace más necesaria como defensa física: el sitio y la hora sin la opción fácil. Si introduce un tutor de IA, que sea dentro de ese tiempo y de los que aparecen solos y no dejan avanzar sin acertar: con ellos los alumnos van más despacio y aprovechan más sus errores, aunque es una sola sesión y el efecto es pequeño y apenas significativo. Pruébalo con sorteo antes de extenderlo (capítulo 3).³⁹

Presupueste lo que cuesta certificar en presencia. Horas de profesor para defensas orales y escritura en clase, y detección de dispositivos en los exámenes que deciden algo, como hicieron al menos diez comunidades en la selectividad de 2026 (capítulo 5).⁶³ La defensa física sigue funcionando, pero ha dejado de ser gratis.

No generalice nada sin un ensayo con sorteo. Corea retiró a sus libros de texto con IA el rango de libro de texto a los cinco meses, y el metaanálisis más citado a favor de ChatGPT fue retractado (capítulos 1 y 4).^{17, 33} Antes de comprar una herramienta, pida el resultado de un grupo que no la usó, medido en un examen sin ella.

En España, lleve a la presencia la parte que certifica. El título de la ESO lo decide el profesorado del centro sin examen externo, y el 60 % de la nota de acceso a la universidad es la media del bachillerato (capítulo 6).^{64, 68} Si dirige un centro, haga que la media del bachillerato salga de pruebas y defensas hechas en el centro. Si decide la norma, revise cómo se acredita el título de la ESO y el reparto del 60 y el 40, sin subir sin más el peso de la prueba de acceso, que haría abandonar lo que esa prueba

no mide. Y publique cada año la distancia entre la media del bachillerato y la de la prueba de acceso, que ya era de ocho décimas antes de la IA.⁶⁵

8.6 Lo que no sirve

Los detectores de texto generado: buscan el error que la premisa borra, y la sospecha, con detector o sin él, cae sobre quien escribe bien (capítulo 5). Las prohibiciones que no llegan a casa: el daño medido está en los deberes hechos fuera del centro, adonde no llega una norma del aula (capítulo 6). Las declaraciones de uso, como la que propone la guía del Ministerio de Educación, usadas para certificar: falsearlas no cuesta nada, y una señal que no cuesta no distingue a nadie. Sirven, como mucho, para enseñar a citar la herramienta (capítulo 6).⁷⁰ Los anuncios de resultados sin grupo de comparación (capítulo 4). Y esperar un tutor mejor: el bueno ya existe y casi nadie lo usa, y con la premisa rendirse a la máquina sale todavía más rentable para la nota (capítulo 3).

Tampoco sirve el alarmismo. Que una generación quede incapaz de leer o de razonar por culpa de la IA no se deduce de ningún dato (capítulo 2). Lo que se pierde es práctica, y la práctica se puede reponer; si lo perdido se recupera, nadie lo ha medido todavía (capítulo 2). El quinto alumno del condado chino mantuvo su tiempo por su cuenta, quizá porque era otro tipo de alumno. La apuesta de este informe es que ese tiempo no dependa de ser como él, sino de un sitio, una hora y alguien que compruebe.

Síntesis

En la escuela, el esfuerzo de hacer los ejercicios cumplía dos funciones a la vez, y nadie las separó porque nunca hizo falta: producía el aprendizaje y era la prueba de que se había producido. Era una defensa por coste doble, de las que describe el capítulo 2 del Informe I. Protegía contra la ignorancia propia, porque para tener el ejercicio hecho había que hacerlo, y contra el engaño, porque copiar exigía un cómplice o un riesgo. La premisa de los laboratorios dice que la máquina resolverá lo comprobable, y casi todo lo que la escuela pide lo es. Por eso ese esfuerzo pasa a ser opcional y se lleva las dos funciones de golpe: quien lo esquiva aprende menos sin notarlo, y el profesor y la familia dejan de saberlo, porque la nota sale igual de bien. Un tutor mejor que cualquier profesor no lo arregla por sí solo, porque lo que faltaba nunca fue la explicación, sino algo que obligue a sentarse, y ese tutor está a un clic de la máquina que resuelve. Lo que no cae son las defensas físicas y las de respuesta: el tiempo del alumno en un lugar, un adulto presente, el examen en una sala. De ahí la frontera que este informe ha trazado, y que la tabla del capítulo 1 detalla. Se deduce que lo que se hace fuera de la vista deja de enseñar a la mayoría de quienes trabajan sin nadie que organice su tiempo, y deja de certificar a todos. No se deduce que la escuela sobre, sino que su función se muda de explicar, que ya es gratis, a garantizar el tiempo y comprobar en presencia. El cuello de botella se muda también dentro de cada persona: cuando la máquina hace lo comprobable, lo escaso es el criterio para juzgarla, y ese criterio se forma con el mismo esfuerzo que la máquina permite ahorrar.

Juicios principales

En la escuela, la mitad de la premisa que importa para los daños ya se cumple. Casi todo lo escolar es comprobable o se comporta como si

lo fuera, y la mayoría de los alumnos ya usa la máquina para hacerlo. Los daños no esperan al calendario de los laboratorios. Los beneficios prometidos, un tutor que sepa qué necesita cada alumno y le sostenga el esfuerzo durante meses, necesitan la otra mitad, la del criterio, que no tiene fecha, y una conducta que la técnica no cambia. Por eso van a ritmos distintos: el daño es una decisión de cada alumno cada tarde y va deprisa; el beneficio pasa por instituciones y va despacio (capítulo 1).

La máquina hace el ejercicio, y el ejercicio era el aprendizaje. Tres diseños en tres países dan la misma dirección: sube la nota de lo que se hace con la máquina y baja la del examen sin ella. En un condado chino, la nota de los deberes subió un 18 % y la del examen mensual a libro cerrado bajó un 20 %.¹⁴ La IA no atonta, porque el alumno que mantiene su tiempo de estudio pierde poco. Lo que quita es la práctica, y con la premisa el daño crece, porque desaparece el error que servía de aviso (capítulo 2).

El alumno elige la respuesta porque no nota lo que pierde y paga hoy lo que ganaría mañana. El tutor que no da la respuesta existe y es gratis, y casi nadie lo usa: en un ensayo de dos cursos en Tennessee, el 96 % lo probó y el alumno mediano le escribió solo en el 17 % de las sesiones en que se equivocaba.¹³ No es pereza como defecto moral, sino un sesgo hacia el presente, predecible, que se corrige cambiando la situación: un sitio y una hora fijados antes de que llegue la tentación. Una máquina mejor mejora la oferta, no la demanda (capítulo 3).

Un tutor mejor que cualquier profesor no sube el techo si no sube el tiempo. Las «dos sigmas» de la promesa salieron de experimentos de tres semanas con una prueba estrecha; la tutoría humana a gran escala da mucho menos, y los datos son compatibles con que lo que se pierde al crecer sea sobre todo la dosis, no la calidad del tutor. Donde alguien protegió el tiempo, en unos internados de la India, la misma plataforma, sin IA generativa, dio un efecto grande. Con la premisa, la escuela se hace más necesaria, no menos, como el sitio y la hora en que se trabaja sin la opción fácil (capítulo 4).

La nota de lo que se hace fuera de la sala deja de medir, y lo primero que se pierde es el aviso. Cuando una medida se convierte en objetivo deja de ser una buena medida, y la máquina quita de golpe el coste de torcerla. Entre quienes usan la IA, más nota en los deberes llega

a anunciar un examen peor, y el profesor y la familia se quedan sin el termómetro con el que sabían si el alumno aprendía. Fuera de la escuela, el escrito deja de separar a unos candidatos de otros, y quien pierde es el bueno (capítulo 5).

Certificar vuelve a la sala, y la sala deja de ser gratis. Ni los detectores ni la prueba de autoría pueden ser la defensa principal: los primeros buscan un error que la premisa borra, y la segunda se dispara por sospecha, que cae sobre quien escribe bien. Lo que funciona es separar lo que certifica, en presencia y sin ayudas, de lo que enseña, abierto y con la máquina. Pero cuesta horas de profesor que no se multiplican, y con modelos que caben en unas gafas la sala tiene que revisar lo que entra en ella (capítulo 5).

España certifica sobre todo con lo que ya no mide. El título de la ESO lo decide el profesorado del centro sin examen externo, y el 60 % de la nota de acceso a la universidad es la media del bachillerato, que ya superaba a la de la prueba de acceso antes de la IA. La respuesta que sostienen los datos no es prohibir por edad ni pedir declaraciones de uso. Es llevar a la presencia la parte de la nota del centro que certifica, con tareas variadas y no solo con más examen, y publicar cada año, centro por centro, cuánto se separa su nota de la prueba de acceso (capítulo 6).

El saber propio pasa de servir para producir a servir para verificar, y se forma con lo que la máquina permite dejar de hacer. Es la ironía que la psicóloga Lisanne Bainbridge describió en 1983 para la automatización industrial: la persona se queda con lo más difícil, vigilar a la máquina y tomar el mando cuando falla, y pierde la práctica que haría falta para hacerlo. Quienes hoy comprueban lo que hace la IA aprendieron sin ella. Es la tesis menos medida del informe, porque no existe ningún estudio que compare cómo verifican quienes aprendieron con IA y quienes aprendieron sin ella (capítulo 7).

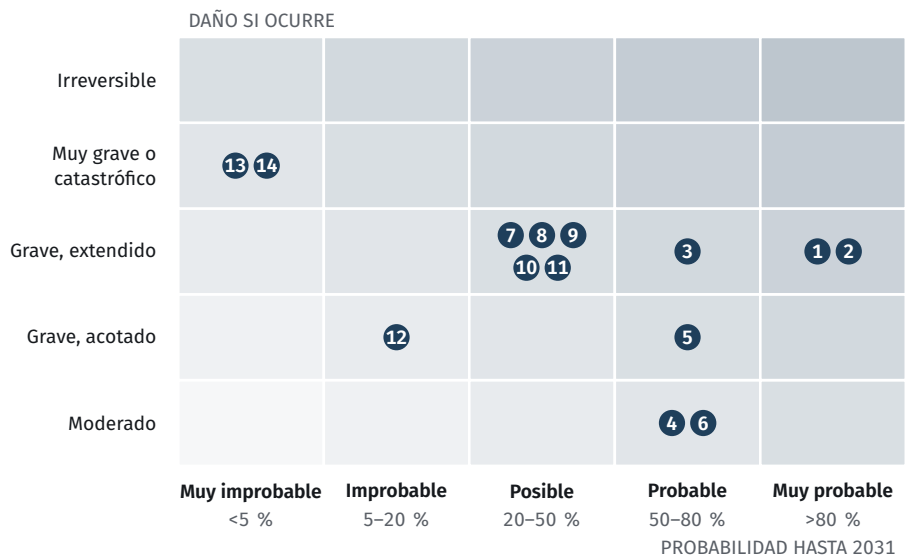
Donde el esfuerzo dejó de ser obligatorio, hay que poner un sitio, una hora y alguien que compruebe. No se trata de volver a encarecer el ejercicio o la copia, que ya no se puede, sino de reponer una defensa física o una de respuesta. Para quien estudia, eso es aprender primero sin la máquina y examinarse sin ella a menudo, en un sitio y a una hora fijados de antemano; para quien enseña, pedir que se defienda de palabra lo que se entrega por escrito; para quien decide, proteger tiempo de

trabajo dentro del horario y pagar las horas que cuesta comprobar en presencia. No sirven los detectores, las prohibiciones que no llegan a casa, las declaraciones de uso ni los anuncios sin grupo de comparación (capítulo 8).

Todos los veredictos

Los capítulos 2 a 7 dan 14 veredictos, todos con el mismo plazo, 2031, y la misma condición, que se cumpla la premisa. Puestos de más a menos probables, repiten la forma de la tesis. Los dos «muy probables» ya se miden y son los más silenciosos: quien usa la IA por su cuenta aprende menos, y la nota de lo que hace fuera de la sala deja de avisar a quien le rodea. Ninguno de los dos se ve en la nota que llega a casa. Los cuatro «probables» son esa misma pérdida medida a escala de país y tres daños del dinero y de la certificación: programas que se retiran sin haber mostrado efecto, cursos a distancia hechos por un agente y aparatos que meten el modelo en la sala del examen. En «posible» se agrupan cinco daños graves cuya prueba todavía falta: la brecha por renta, la señal escrita en la contratación, la nota del bachillerato español, sus alumnos excelentes y la capacidad de los jóvenes de comprobar a la máquina. Lo catastrófico, una generación incapaz de leer o de razonar, o quedarse sin nadie formado para revisar a la máquina, está en el tramo más bajo, porque lo que se pierde es práctica, y la práctica se repone.

Figura 6 Los 14 veredictos, por probabilidad y gravedad



Juicio de este informe. Cada número remite a la tabla siguiente; la fila agrupa la gravedad que da cada capítulo.

La fila de arriba, la de lo irreversible, está vacía, y los dos daños muy graves están en la columna de lo muy improbable: nada de lo que este informe deduce para la educación es a la vez catastrófico y probable. El peso está en la fila de los daños graves y extendidos, repartido entre «posible» y «muy probable». Son los que dependen de que alguien ponga un sitio, una hora y un examen en presencia, y los de la derecha ya ocurren donde nadie los ha puesto. Abajo, lo moderado es dinero y tiempo lectivo perdidos en programas sin efecto y lo que cuesta revisar lo que entra en la sala.

Tabla 3 Los veredictos del informe, de aquí a 2031, si se cumple la premisa

N.º	Suceso	Probabilidad	Gravedad	Cap.
1	La mayoría de quienes usan la IA por su cuenta para los deberes la usan para acabar antes, y aprenden menos que sin ella	Muy probable	Grave; solo se ve en un examen sin máquina	2
2	En los estudios que comparan a los mismos alumnos, la nota de lo hecho fuera de la sala predice la del examen vigilado claramente peor que antes de 2022, o al revés, en al menos tres países	Muy probable	Grave: el profesor y la familia pierden el aviso; ya ocurre en dos	5
3	Fuera de China, un estudio con grupo de comparación sobre una prueba externa nacional sin máquina mide, entre quienes usan la IA para los deberes, una pérdida de al menos el 5 % de su media anterior	Probable	Grave; no se sabe si es recuperable	2
4	Entre 2027 y 2031, un programa nacional o regional de IA para alumnos de un país de la OCDE, extendido sin ensayo con grupo de comparación, se retira o se recorta sin haber mostrado efecto en una evaluación externa, después de ocupar horas lectivas	Probable	Moderado: dinero y tiempo lectivo perdidos, recuperables	4

continúa

N.º	Suceso	Probabilidad	Gravedad	Cap.
5	Casos documentados de cursos o títulos a distancia completados por un agente de IA en nombre del alumno, en al menos dos países	Probable	Grave para quien tiene un título a distancia honrado, que pierde valor	5
6	Fraude documentado con un aparato que lleva un modelo sin conexión (gafas, reloj, auricular) en un gran examen presencial nacional, de acceso o profesional, en al menos dos países	Probable	Moderado: encarece la defensa física	5
7	Los estudios con grupo de comparación publicados hasta 2031 encuentran que, con la IA a mano, los alumnos de hogares con menos renta pierden más o ganan menos que los demás	Posible	Grave para la equidad; se nota tarde	3
8	Contratación o admisiones menos meritocráticas porque el escrito deja de separar, medido en un mercado distinto del de Galdin y Silbert	Posible	Grave e injusto	5
9	La distancia entre la media del bachillerato y la de la prueba de acceso llega a un punto de media en las convocatorias de 2026 a 2029, según el Ministerio de Ciencia	Posible	Grave: es el 60 % de la calificación de acceso	6

continúa

N.º	Suceso	Probabilidad	Gravedad	Cap.
10	La proporción de alumnos españoles en los dos niveles más altos de matemáticas baja en PISA 2029, respecto de 2025, con una diferencia que la OCDE da por significativa	Posible	Grave y lento: es donde España ya flojea	6
11	Un estudio con grupo de comparación o una evaluación externa encuentra que jóvenes formados con la IA a mano detectan peor los errores de la IA en su propio campo que otros de la misma edad formados antes de 2023	Posible	Grave; se nota tarde	7
12	Un gobierno nacional o regional de un país de la OCDE aprueba reducir plantillas docentes o programas de tutoría humana citando la IA como sustituto	Improbable	Grave donde ocurra	4
13	Una generación incapaz de leer o de razonar por la IA: en PISA 2029, más del 30 % de los alumnos de la OCDE queda por debajo del nivel básico en las tres materias, y la caída es mayor donde más se usa la IA para los deberes	Muy improbable	Catastrófico; los datos no apuntan ahí	2
14	Un regulador o un colegio profesional de un país rico reconoce que ya no tiene personas formadas para revisar el trabajo de la IA en su campo y deja de revisarlo	Muy improbable	Muy grave; difícil de revertir	7

Qué vigilar

Ningún veredicto está cerrado. Cada capítulo dice qué dato lo movería, y casi todos esos datos tienen fecha o una fuente que los publica con regularidad. Esta tabla reúne los que más veredictos moverían a la vez.

Tabla 4 Señales que harían cambiar los veredictos

Dato	Qué indicaría	Dónde mirarlo
Un estudio fuera de China, con registros de tiempo, sobre alumnos que usan la IA para los deberes	Si la mayoría acaba mucho antes con nota alta y rinde peor sin máquina, se confirman los veredictos 1 y 3; si no aparece la pérdida, bajan	Estudios con grupo de comparación como el de Strömberg, Lei y Wu (capítulo 2)
Si lo perdido se recupera cuando vuelven el examen presencial o la práctica vigilada	Bajaría la gravedad de los veredictos 1, 3 y 11; nadie lo ha medido	Estudios de seguimiento de más de tres años (capítulos 2 y 7)
Resultados del segundo ensayo de Eedi y Google DeepMind, anunciados para el verano de 2026 y sin publicar a finales de septiembre	Si un tutor restringido y revisado por un humano enseña más que el contenido fijo; no dirá si se usa sin alguien que obligue	Eedi (capítulos 1 y 8)
Primeros resultados del fondo de la Education Endowment Foundation sobre delegar el esfuerzo en la máquina, previstos para 2027	Pérdidas mayores en hogares con menos renta subirían el veredicto 7; que practicar con IA y examinarse sin ella deje los mismos fundamentos bajaría el 11	Education Endowment Foundation (capítulos 3 y 7)
Uso semanal de la herramienta estonia cuando el programa se amplíe en el curso 2026-2027, y el estudio de impacto de la Universidad de Tartu, Stanford y OpenAI	Si el uso sigue por debajo del 20 % de los destinatarios, se confirma que el beneficio va despacio aunque la herramienta sea buena	Programa estonio y OpenAI (capítulo 1)

continúa

Dato	Qué indicaría	Dónde mirarlo
Un tutor que aparezca solo, integrado en el contenido, como el que anunció Khan Academy en 2026 o el de NUMI, medido a escala	Con ganancias duraderas fuera del aula vigilada, el tutor subiría el techo sin más tiempo y cambiaría la posición de los capítulos 3 y 4	Ensayos de Oreopoulos y de Khan Academy (capítulos 3 y 4)
Si los gobiernos exigen un ensayo con grupo de comparación antes de extender un programa de IA en las aulas	Si lo exigen, baja el veredicto 4; un resultado grande de un tutor en un ensayo con sorteo, sumado a recortes presupuestarios, subiría el 12	Boletines oficiales y presupuestos (capítulo 4)
Nota de lo hecho fuera de la sala frente a la del examen vigilado, en los mismos alumnos, en un tercer país	Cumpliría el veredicto 2; un rediseño del trabajo en casa que devolviera la relación lo bajaría	Estudios como el de Hausman, Rigbi y Weisburd (capítulo 5)
Aparatos capaces de llevar un modelo potente sin conexión, y fraudes documentados con ellos	Si siguen siendo fáciles de ver a simple vista, baja el veredicto 6	Organizadores de la selectividad y de exámenes profesionales (capítulo 5)
Una encuesta seria de alumnos acusados de usar IA, o de acusaciones revocadas en recurso	Las acusaciones injustas a quien escribe bien entrarían en los veredictos	Universidades y encuestas como la del Higher Education Policy Institute (capítulo 5)
Un segundo mercado, de empleo o de admisiones, en que el escrito deje de separar a los candidatos	El veredicto 8 subiría a «probable»	Estudios de economía del trabajo (capítulo 5)

continúa

Dato	Qué indicaría	Dónde mirarlo
Distancia entre la media del bachillerato y la de la prueba de acceso, convocatorias de 2025 a 2029	Por encima de nueve décimas en 2025 y 2026, el veredicto 9 sube a «probable»; si no crece, baja a «improbable»	Estadísticas del Ministerio de Ciencia (capítulo 6)
PISA 2029, con resultados hacia finales de 2030: alumnos bajo el nivel básico en las tres materias, alumnos españoles en los niveles altos de matemáticas, destrezas de evaluar información	Resuelve los veredictos 10 y 13 y movería el 11. Es la primera cohorte que hizo toda la secundaria con la IA a mano	OCDE (capítulos 2, 6 y 7)
Acierto de los agentes en decisiones de juicio, hoy del 59 % frente al 90 % de los investigadores	Si se acerca al humano con fiabilidad medida, verificar deja de ser el uso principal del saber propio y el peso pasa a la pérdida de poder del capítulo 9 del Informe I	METR (capítulos 1 y 7)
Un estudio que compare cómo detectan los errores de la IA jóvenes formados con ella y otros formados antes de 2023	Resuelve el veredicto 11; no hay ninguno anunciado	Universidades y evaluadores externos (capítulo 7)

Lo que no se sabe

Los veredictos descansan en datos cortos. El seguimiento más largo son los treinta meses del condado chino. Nadie ha medido si lo perdido se recupera cuando vuelve la práctica o el examen presencial, ni qué pasa pasados tres años. Por debajo de los 16 años, la Education Endowment Foundation reconoce que la prueba sólida apenas ha empezado, y sus primeros resultados llegarán en 2027.²¹ Del tutor que aparece solo y no deja avanzar sin acertar, el único diseño de tutor que ha cambiado algo la conducta, hay una sola sesión con un efecto pequeño. La prueba de que

el escrito deja de separar a los candidatos viene de un solo mercado de encargos cortos de programación. Y no hay datos de si la expectativa de que la máquina haga los trabajos reduce ya el esfuerzo de los estudiantes.

Dos preguntas quedan abiertas del todo. Quién pierde más: en China, los alumnos que iban mejor, los más jóvenes y los chicos; en Estados Unidos, el uso más intensivo se concentra en los hogares pobres; de España no hay datos. Y cómo verifica a la máquina quien aprendió con ella: no existe ningún estudio que lo compare con quien aprendió sin ella, y la tesis del capítulo 7 descansa en un mecanismo, un ensayo teórico y datos indirectos.

España tiene sus propios huecos. No hay ningún estudio causal español sobre IA y aprendizaje, solo encuestas. La serie publicada de la distancia entre la nota del bachillerato y la de la prueba de acceso termina en 2024. No se publica qué parte de la nota del centro sale de trabajo hecho fuera de clase, que es la cifra que más falta para saber cuánto de la certificación española ha dejado de medir. Y no hay datos de cómo acreditan lo aprendido las universidades a distancia.

Cómo leer los veredictos

Cada veredicto es un juicio de este informe sobre un suceso concreto, con un plazo, 2031, y una condición, que se cumpla la premisa. No es una frecuencia ni sale de un modelo matemático: resume las razones de «Por qué estos veredictos», en cada capítulo, para que se pueda discrepar de ellas una a una. Solo llevan veredicto los daños. Que casi nadie use el tutor que no da la respuesta es un beneficio que no llega, y su daño, aprender menos sin notarlo, ya está en el veredicto 1 (capítulo 3).



Un tramo no es una cifra: «posible» quiere decir entre el 20 y el 50 %, sin más. Casi todos los veredictos de este informe nombran, además, el dato publicado que diría en 2031 si el suceso ocurrió: un estudio con grupo de comparación, una convocatoria de la prueba de acceso, un ciclo de PISA. Eso los hace comprobables y tiene un precio. Un daño que ocurre

y nadie mide no cumple el veredicto, y por eso algunos quedan en un tramo más bajo que el daño que describen; lo que hoy no cuenta nadie, como las acusaciones injustas a quien escribe bien, se queda en la prosa. Los veredictos no se suman ni se multiplican, porque comparten causas: el mismo uso sin nadie que organice el tiempo está detrás del 1, el 2, el 3 y el 7. La gravedad es la del daño si ocurre, y no se promedia con la probabilidad. Si la premisa no se cumple, los que menos bajan son los de los capítulos 2, 5 y 6, porque solo necesitan la mitad comprobable, que en la escuela ya se cumple (capítulo 1). Los que más bajarían son los que necesitan agentes o modelos potentes que funcionen sin conexión, el 5 y el 6.

Referencias

Numeradas por orden de primera cita. Bajo cada entrada, el alcance del resultado que se utiliza. Consulta: septiembre de 2026.

- 1 OCDE. *PISA 2025 Results (Volume I). Country notes: Spain*. 2026-09-08. oecd.org ↗
- 2 McClain, C., Anderson, M., Sidoti, O. y Bishop, W. (Pew Research Center). *How Teens Use and View AI*. 2026-02-24. pewresearch.org ↗
PRIMARIA LEÍDA. 1.458 adolescentes de EE. UU. y sus padres, 25 sep-9 oct 2025; margen $\pm 3,3$. Muestra probabilística (Ipsos KnowledgePanel).
- 3 Stephenson, R. y Armstrong, C. (Higher Education Policy Institute). *Student Generative AI Survey 2026 (HEPI Report 199)*. 2026-03-12. hepi.ac.uk ↗
PRIMARIA LEÍDA. Encuesta de Savanta (dic 2025) a 1.054 universitarios del Reino Unido. Autodeclarado.
- 4 TI-Hüpe (AI Leap Foundation). *Estonia Launches Nationwide AI Leap Education Program*. 2026. tihupe.ee ↗
Cifras de activación comunicadas por el propio programa, marzo de 2026.
- 5 MIT Technology Review. *OpenAI is launching a version of ChatGPT for college students*. 2025-07-29. technologyreview.com ↗
SECUNDARIA seria; incluye la cita de Leah Belsky (OpenAI). Página oficial del producto no leída.
- 6 OpenAI. *New tools for understanding AI and learning outcomes*. 2026-03-04. openai.com ↗
Ensayo propio de la empresa sobre su «modo estudio».
- 7 Dario Amodei. *Machines of Loving Grace: How AI Could Transform the World for the Better*. 2024-10. darioamodei.com ↗
Ensayo del consejero delegado de Anthropic. Define «IA potente» y los factores que limitan a la inteligencia.
- 8 Dario Amodei. *The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI*. 2026-01. darioamodei.com ↗
Ensayo de opinión; el autor advierte de que no pretende comunicar certeza ni probabilidad.

- 9 Epoch AI. *Trends in Artificial Intelligence*. 2026. epoch.ai ↗
Tendencias estimadas con intervalos de confianza amplios; se actualizan de forma continua.
- 10 OpenAI. *GPT-6 Astra: A new generation of intelligence*. 2026-09-03. openai.com ↗
Resultados publicados por el desarrollador.
- 11 Scarfe, P., Watcham, K., Clarke, A. y Roesch, E.. *A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study (PLOS ONE 19(6))*. 2024-06-26. journals.plos.org ↗
PRIMARIA LEÍDA (resumen). Revisado por pares. Con GPT-4 y exámenes para casa; anterior a 2025.
- 12 Anthropic. *Anthropic Education Report: How University Students Use Claude*. 2025-04-08. anthropic.com ↗
PRIMARIA LEÍDA de parte interesada: datos de uso de un solo proveedor, cuentas universitarias, abril de 2025.
- 13 Oreopoulos, P. y Low, N.. *One Click Away: AI Tutoring with Khanmigo in a Two-Year School Experiment*. 2026-08. nber.org ↗
NBER Working Paper 35620.
- 14 Strömberg, D., Lei, V. y Wu, Y.. *The Generative AI Learning Penalty: Evidence from Chinese Secondary Education*. 2026-06-02. cepr.org ↗
CEPR Discussion Paper 21577; columna de los autores en VoxEU, 4 de septiembre de 2026. Documento de trabajo.
- 15 METR. *Frontier Risk Report*. 2026-05-19. metr.org ↗
Evaluación piloto de febrero-marzo, con acceso interno. Riesgo de despliegues no autorizados pequeños; no probabilidad calibrada de catástrofe.
- 16 A. Narayanan y S. Kapoor. *AI as Normal Technology*. 2025-04. knightcolumbia.org ↗
Ensayo; la tesis de la difusión lenta como amortiguador.
- 17 Rest of World. *South Korea's AI textbooks fail after rushed rollout*. 2025. restofworld.org ↗
Reportaje; la adopción cayó del 37 al 19 % y el Parlamento retiró a los libros su rango oficial.
- 18 Randolph, M. (Scientific American). *Alpha School's AI teaching model is expanding. Does it work?*. 2026-08-28. scientificamerican.com ↗
SECUNDARIA seria; Alpha no publica datos ni permite evaluación independiente.

- 19 Kosmyrna, N. et al.. *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*. 2025-06. arxiv.org ↗
Muestra pequeña; preprint.
- 20 elDiario.es. *Salvador Illa plantea prohibir la IA en las aulas catalanas antes de los 14 años*. 2026-09-22. eldiario.es ↗
SECUNDARIA seria; las recomendaciones del Departamento (defensa oral, autorización 14-18) en Xataka Móvil (4 sep 2026) y El Debate (31 ago 2026).
- 21 Education Endowment Foundation. *New landmark programme of research to investigate the impact of Generative AI on learning and cognition*. 2026-06-08. educationendowmentfoundation.org.uk ↗
PRIMARIA LEÍDA (nota de prensa). Institución independiente de referencia en evidencia educativa (Inglaterra).
- 22 Eedi. *Launched: our second AI tutor RCT*. 2026-04. eedi.com ↗
Anuncio del segundo ensayo aleatorizado con Google DeepMind: 1.525 alumnos de 8.º a 10.º curso, 10 institutos británicos, cuatro brazos, 12 semanas; tutor restringido que solo actúa cuando el alumno falla. Resultados anunciados para el verano de 2026, sin publicar a 23 sep 2026. Comprobado en la web el 23 sep 2026.
- 23 Bastani et al.. *Generative AI without guardrails can harm learning*. 25-06-2025. doi.org ↗
Experimento escolar de matemáticas en Turquía; distingue rendimiento con ayuda de aprendizaje posterior. No prueba daño universal ni irreversible.
- 24 Rismanchian, S., Uzun, H., Matayoshi, J., Cosyn, E. y Kurd-Misto, E.. *Faster Completion, Less Learning: Generative AI Reduced Study Time on Math Problems and the Knowledge They Build*. 2026-06. arxiv.org ↗
Preprint sin revisión por pares; parte de los autores trabaja en McGraw Hill, dueña de la plataforma.
- 25 OCDE. *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*. 2026-01-19. oecd.org ↗
- 26 N. C. Soderstrom y R. A. Bjork. *Learning Versus Performance: An Integrative Review*. 2015. doi.org ↗
Perspectives on Psychological Science, 10(2). El rendimiento durante la práctica no indica lo que se retendrá; las «dificultades deseables» lo empeoran a corto plazo y lo mejoran a largo.
- 27 Lira, B., Rogers, T., Goldstein, D. G., Ungar, L. y Duckworth, A. L.. *AI assistance can enhance rather than hinder skill development*. 2025-02. arxiv.org ↗

Ensayos aleatorizados prerregistrados con adultos.

- 28 MIT News (sobre J. Reich y el Teaching Systems Lab). *Helping K-12 schools navigate the complex world of AI*. 2025-11-03. news.mit.edu ↗
PRIMARIA institucional con citas literales de Reich. Guía: A Guide to AI in Schools (ago 2025).
- 29 OCDE. *PISA 2025 Results (Volume I): Future-Ready Students. Executive summary y tabla I.1*. 2026-09-08. oecd.org ↗
- 30 OCDE. *PISA 2025 Results (Volume I), capítulo 2: Student performance in PISA 2025 (recuadro I.2.3, «Why has reading performance declined over time?»)* . 2026-09-08. oecd.org ↗
- 31 OCDE. *PISA 2025 Results (Volume I), capítulo 4: Student school life and beyond*. 2026-09-08. oecd.org ↗
Dispositivos, IA, deberes, familia y absentismo.
- 32 Hausman, N., Rigbi, O. y Weisburd, S.. *Generative AI in universities: Grades up, signals down, skills in flux*. 2025-06-17. cepr.org ↗
Columna de los autores en VoxEU.
- 33 Humanities and Social Sciences Communications (Springer Nature). *Retraction Note: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis*. 2026-04-22. doi.org ↗
El artículo original, de J. Wang y W. Fan, se publicó el 6 de mayo de 2025.
- 34 OCDE. *PISA 2025 Results (Volume I)*. 2026-09-08. oecd.org ↗
760.000 alumnos de 15 años en 91 países. Las relaciones entre uso de IA y puntuación son correlaciones.
- 35 Chalkbeat. *Students rarely engaged with Khan Academy's AI-powered tutor Khanmigo, study finds*. 2026-08-25. chalkbeat.org ↗
Estudio de dos años en distritos de Estados Unidos.
- 36 Holt, L.. *The 5 Percent Problem: Online mathematics programs may benefit most the kids who need it least*. 2024. educationnext.org ↗
Education Next, 24(4).
- 37 Shaw, S. D. y Nave, G.. *Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender (SSRN 6097646)*. 2026-01. executiveeducation.wharton.upenn.edu ↗
SECUNDARIAS SERIAS (Wharton Executive Education; blog de SSRN; TNW). Preprint; SSRN bloqueado con verificación antibots. Adultos, Test de Reflexión Cognitiva; no mide aprendizaje.

- 38 Lavecchia, A. M., Liu, H. y Oreopoulos, P.. *Behavioral Economics of Education: Progress and Possibilities (Handbook of the Economics of Education, vol. 5)*. 2016. ideas.repec.org ↗
PRIMARIA: resumen leído (versión NBER WP 20609, 2014). Capítulo de manual de referencia; el propio Oreopoulos lo usa en 2026 para explicar el poco uso de Khanmigo.
- 39 Oreopoulos, P., Liut, M., Sungu, A. y Low, N.. *Making AI Tutoring Productive: Evidence from a Mastery-Based Math Practice Experiment (NBER WP 35621)*. 2026-08. nber.org ↗
PRIMARIA LEÍDA (resumen e introducción). Ensayo aleatorizado prerregistrado, >6.000 alumnos, una sesión y prueba a la semana. Sin revisión por pares; efecto diferido pequeño y marginal.
- 40 Oreopoulos, P., Keyes-Krysakowski, O. y Agarwal, D.. *How In-School Supervised Ed-Tech Support Produces Massive Learning Gains: A Khan Academy Field Experiment in India (NBER WP 34683)*. 2026-01. nber.org ↗
PRIMARIA: resumen leído. 83 centros de Uttar Pradesh; sin IA generativa (Khan Academy clásico). Sin revisión por pares.
- 41 Barton, C. (Eedi). *AI in Education #11: Five Things I Learned from Dylan Wiliam*. 2026-06-30. eedi.substack.com ↗
UNA SOLA FUENTE (boletín de una empresa de tecnología educativa que resume una conversación). Citas cortas de Wiliam; usar con cautela.
- 42 De Simone, M., Tiberti, F., Barron Rodriguez, M., Manolio, F., Mosuro, W. y Dikoru, E. J.. *From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria (Policy Research Working Paper 11125)*. 2025-05-20. openknowledge.worldbank.org ↗
PRIMARIA: resumen oficial leído; PDF no accesible (403). Ensayo aleatorizado con voluntarios; el Banco Mundial diseña y evalúa. Cifras oficiales: 0,31 d. t. global, 0,23 en inglés.
- 43 Kraft, M. A., Schueler, B. E. y Falken, G. T.. *What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability (EdWorkingPaper 24-1031)*. 2024-10. edworkingpapers.com ↗
PRIMARIA LEÍDA (PDF del documento de trabajo). 265 ensayos aleatorizados. Publicado en Review of Educational Research en 2026 (DOI 10.3102/00346543261446660; versión publicada no leída).
- 44 von Hippel, P. T.. *Two-Sigma Tutoring: Separating Science Fiction from Science Fact (Education Next 24(2))*. 2024. educationnext.org ↗
PRIMARIA LEÍDA. Revista de divulgación académica (Hoover/Harvard). Revisa las fuentes originales de Bloom 1984; útil para el techo realista de la tutoría.

- 45 Wang et al. (Stanford). *Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise*. 2024-10. nssa.stanford.edu ↗
Ensayo aleatorizado con 900 tutores y 1.800 alumnos de 3.º a 8.º, durante dos meses. No detecta mejora estadísticamente significativa en el examen de fin de curso, con una exposición breve y escasa.
- 46 Education Endowment Foundation. *ChatGPT in lesson preparation: Teacher Choices trial*. 2024-12. educationendowmentfoundation.org.uk ↗
68 institutos ingleses; profesores de ciencias.
- 47 Google DeepMind y Eedi. *AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms*. 2025-12. arxiv.org ↗
165 alumnos de 13 a 15 años; los mensajes de la IA pasaban por un tutor humano.
- 48 Kestin et al. (Harvard). *AI tutoring outperforms in-class active learning*. 2025-06. nature.com ↗
194 universitarios de Física en Harvard: alumnos muy seleccionados y motivados.
- 49 Infobae. *Programa piloto con IA y tecnología lleva a 171 escuelas salvadoreñas a niveles de países desarrollados en PISA, dice el gobierno*. 2026-09-09. infobae.com ↗
- 50 La Noticia SV. *PISA 2025: El Salvador mejora sus resultados en Ciencias, Matemáticas y Lectura*. 2026-09. lanoticiasv.com ↗
Cifras nacionales de PISA 2025, de prensa.
- 51 Gortazar, L. y Montalbán, J.. *La nueva ley de ratios: estimación de su impacto económico*. 2026-05-26. esade.edu ↗
Centro de políticas económicas de Esade (EsadeEcPol).
- 52 Universidad de Sídney. *University of Sydney AI assessment policy (two-lane approach)*. 2024-11-27. sydney.edu.au ↗
PRIMARIA LEÍDA (nota oficial). En vigor desde el 2.º semestre de 2025.
- 53 C. A. E. Goodhart. *Problems of Monetary Management: The U.K. Experience*. 1975. rba.gov.au ↗
Ponencia de 1975, publicada en Papers in Monetary Economics I (Banco de la Reserva de Australia, 1976). Origen de la ley de Goodhart, formulada para la política monetaria.
- 54 M. Strathern. *'Improving ratings': audit in the British University system*. 1997. cambridge.org ↗
European Review 5(3), p. 308: formulación popular de la ley de Goodhart, tomada de Keith Hoskin.
- 55 M. Spence. *Job Market Signaling*. 1973. doi.org ↗

- Quarterly Journal of Economics 87(3). Una señal informa solo si cuesta más a quien no tiene la cualidad.
- 56 Galdin, A. y Silbert, J.. *Making Talk Cheap: Generative AI and Labor Market Signaling*. 2025-11-11. arxiv.org ↗
 PRIMARIA LEÍDA (resumen, introducción, datos, contrafactual). Preprint (job market paper), sin revisión por pares. Freelancer.com, 2,7 M de solicitudes. El 19 % / 14 % es simulación del modelo, no dato observado.
- 57 The Register (con la investigación de The Guardian). *UK students flock to AI to help them cheat*. 2025-06-16. theregister.com ↗
 SECUNDARIA seria que resume la investigación del Guardian (peticiones de información a 155 universidades). Casos probados, no uso real.
- 58 Málaga Press (con declaraciones a SUR). *La UMA detecta IA en el 90 % de los trabajos de clase*. 2026-09-21. malagapress.com ↗
 SECUNDARIA local; citas de J. Canca. La cifra del 90 % de trabajos no está verificada en documento de la UMA. Reglamento: Novaciencia, 23 jul 2026.
- 59 The Decoder (con el Financial Times). *AI fraud forces world's largest accounting body to end online exams*. 2025-12-29. the-decoder.com ↗
 Recoge la entrevista del FT a Helen Brand; confirmado por The Accountant.
- 60 Entrepreneur (con The Wall Street Journal). *College Professors Are Turning to an Old-School Product From a Family-Owned Business to Combat AI Cheating*. 2025-05-29. entrepreneur.com ↗
 SECUNDARIA que cita al WSJ (de pago). Cifras de ventas de tres universidades.
- 61 Dyson, J. (Schools Week). *More scrutiny of coursework plans to protect exams from AI*. 2026-07-03. schoolsweek.co.uk ↗
 SECUNDARIA seria (prensa educativa especializada) con citas literales del discurso del director de Ofqual.
- 62 Fortune (con Bloomberg). *Alibaba, Tencent freeze AI tools during high-stakes China exam*. 2025-06-09. fortune.com ↗
 SECUNDARIA seria; coincide con The Washington Post (10 jun 2025). La repetición en 2026 no está verificada.
- 63 elDiario.es. *Claves y detalles de la Selectividad 2026: exámenes más prácticos y detectores de frecuencia para combatir el uso de IA*. 2026-06. eldiario.es ↗
 SECUNDARIA seria.
- 64 Boletín Oficial del Estado. *Real Decreto 534/2024, de 11 de junio, por el que se regulan los requisitos de acceso a las enseñanzas universitarias oficiales de Grado, las características básicas de la prueba de acceso y la normativa*

básica de los procedimientos de admisión (texto consolidado). 2024-06-11. boe.es ↗

PRIMARIA LEÍDA (arts. 14.2 y 15). Consolidado a 2 jul 2026.

- 65 Newtral. *En los centros privados la nota de Bachillerato es mayor que la de la PAU, pero en los públicos también*. 2025-06-10. newtral.es ↗

SECUNDARIA seria (verificadora) con datos del SIIU (Ministerio de Ciencia) 2015-2024. Anterior al efecto de la IA.

- 66 Instituto Nacional de Evaluación Educativa. *Resultados de TALIS 2024: informe español*. 2025-10-07. inee.educacion.es ↗

- 67 Novaciencia. *La UMA cambia la evaluación: regula el uso de la IA y limita el peso de un único examen*. 2026-07-23. novaciencia.es ↗

SECUNDARIA; reglamento de evaluación de la UMA aprobado en consejo de gobierno (jul 2026): prueba de autoría ante indicios fundados de fraude y ninguna prueba aislada por encima del 70 % de la calificación final. Comprobado en la web el 23 sep 2026; no leído en el reglamento.

- 68 Boletín Oficial del Estado. *Real Decreto 217/2022, de 29 de marzo, por el que se establece la ordenación y las enseñanzas mínimas de la Educación Secundaria Obligatoria*. 2022-03-29. boe.es ↗

PRIMARIA (arts. 15 y 17 leídos vía herramienta; comprobar el número del artículo de la evaluación de diagnóstico).

- 69 Fundación CYD. *Uso y percepción de la IA en el entorno universitario II*. 2026-07-02. fundacioncyd.org ↗

PRIMARIA LEÍDA (nota de la fundación). Encuesta a 800 universitarios; autodeclarado.

- 70 Objetivo CLM (sobre la guía del INTEF). *La nueva guía de IA para las aulas aborda privacidad, evaluación y bienestar digital*. 2026-09-23. objetivo-castillalamancha.es ↗

SECUNDARIA; la guía (actualización del 14 sep 2026) no se ha leído entera.

- 71 Holmström y Milgrom. *Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design*. 1991. doi.org ↗

Journal of Law, Economics, & Organization 7 (número especial). Premiar lo medible desvía la atención de lo que no se mide; modelo teórico.

- 72 Moreno, J. M. y Gortazar, L.. *El mercado de las clases particulares en España entre 2020 y 2024*. 2026-06-25. esade.edu ↗

EsadeEcPol, con la Encuesta de Gasto de los Hogares en Educación del INE.

- 73 L. Bainbridge. *Ironies of Automation*. 1983. doi.org ↗

- Automatica, 19(6). Cuanto más se automatiza, más depende el sistema de un operador muy cualificado que ya no practica.
- 74 Oakley, B., Johnston, M., Chen, K.-Z., Jung, E. y Sejnowski, T. J.. *The Memory Paradox: Why Our Brains Need Knowledge in an Age of AI*. 2025-05-03. arxiv.org ↗
PRIMARIA: resumen leído. Ensayo teórico, no experimento. No usar su tesis sobre el efecto Flynn.
- 75 Stanford Digital Economy Lab. *No Widespread Displacement, but the AI Employment Gap for Young Workers Has Widened to 19%*. 2026-08. digitaleconomy.stanford.edu ↗
Datos hasta junio de 2026; brecha relativa de 22-25 años, análisis descriptivo, limitaciones de muestra y especificación.
- 76 Infobae. *La IA golpea primero a los más formados: las ofertas para programadores junior caen un 33 %*. 2026-09-17. infobae.com ↗
Datos de ofertas de la Fundación CYD, DigitalES e InfoJobs; ofertas publicadas, no contrataciones.
- 77 S. M. Hosseini y G. Lichtinger. *Generative AI as Seniority-Biased Technological Change: Evidence from U.S. Résumé and Job Posting Data (SSRN 5425555)*. 2025-10-27. papers.ssrn.com ↗
Documento de trabajo sin revisión por pares; currículos de LinkedIn vía Revelio Labs; adopción medida por ofertas de integrador de IA, no por uso.
- 78 Kulveit et al.. *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*. 2025-01. arxiv.org ↗
Argumento teórico, no estimación empírica.
- 79 Drago y Laine. *The Intelligence Curse*. 2025-04. intelligence-curse.ai ↗
Serie de ensayos sin revisión por pares. Aplica a la IA la maldición de los recursos: quien vive de la IA pierde el incentivo a invertir en las personas.