

INDICADORES EN LA ERA DE LA IA

Medir con criterio

Memoria de la conferencia magistral y guía para usar
inteligencia artificial en los indicadores de impacto de la
biblioteca académica

Santiago Villegas-Ceballos

Director de la Unidad de Innovación e IA (iIA) de eLibro y docente principal

IV Encuentro Interamericano de Bibliotecarios, de Información y Alfabetización · Universidad Nacional
de Cuyo · Mendoza, 1 de octubre de 2026

Marco ACEQUIA, propuesta para Argentina · Marco PAPAYAS, tercera edición · Unidad iIA de eLibro ·
elibro.com/iia

Medir con criterio

Memoria de la conferencia magistral «Indicadores en la era de la IA: cómo medir el impacto de la biblioteca universitaria para la acreditación» y guía para usar inteligencia artificial en los indicadores de impacto de la biblioteca académica.

Autor: Santiago Villegas-Ceballos, Director de la Unidad de Innovación e IA (iIA) de eLibro Corp. y docente principal; bibliotecario; miembro del Standing Committee de la Sección de Tecnología de la Información de la IFLA (a título personal). santiago@elibro.com · @MedeJean

Edita: Unidad de Innovación e Inteligencia Artificial (iIA) de eLibro · <https://www.elibro.com/iia> · iia@elibro.com

Primera edición: octubre de 2026. Fecha de corte de la información: 1 de octubre de 2026.

Ocasión: IV Encuentro Interamericano de Bibliotecarios, de Información y Alfabetización (EIBIA), «Hacia la transformación y conexión de las bibliotecas», organizado por el Sistema Integrado de Documentación de la Universidad Nacional de Cuyo, Mendoza, Argentina, 30 de septiembre al 2 de octubre de 2026. eLibro es auspiciante del encuentro.

Licencia

El texto propio de esta obra se publica bajo Creative Commons Atribución 4.0 Internacional (CC BY 4.0). Las láminas de la presentación y la infografía del Marco PAPAYAS se publican bajo Creative Commons Atribución-NoComercial 4.0 Internacional (CC BY-NC 4.0). Las normas, leyes, estudios, gráficos y citas de terceros conservan los derechos de sus titulares y se citan con su fuente. Los gráficos de las figuras 2 y 3 son elaboración propia a partir de las fuentes indicadas.

Cómo citar

Villegas-Ceballos, S. (2026). *Medir con criterio: Memoria de la conferencia magistral y guía para usar inteligencia artificial en los indicadores de impacto de la biblioteca académica*. Unidad iIA de eLibro. <https://www.elibro.com/iia>

Aviso de encuadre

Esta obra es una propuesta de organización del trabajo. No cuenta con aval de la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU), de la Comisión Nacional de Acreditación de Chile (CNA), de ISO, de la IFLA ni de ninguna universidad. Toda cifra sobre inteligencia artificial lleva modelo, fecha y fuente; lo autorreportado por un proveedor, lo publicado por la industria sin revisión por pares y los preprints se señalan como tales.

Identidad y producción

Identidad visual de la Unidad iIA de eLibro: tipografía DM Sans; composición y producción editorial en HTML y CSS de impresión. Los archivos de la demostración, el guion de auditoría y la matriz ACEQUIA en una página se publican con esta obra.

Índice

Presentación	4
PRIMERA PARTE Memoria de la conferencia	5
1 La acequia: el caudal y el oasis	6
2 Siete propiedades estructurales, en la biblioteca académica	6
3 ¿Mejoran las tasas de alucinación? Sí, entre generaciones, con condiciones	9
4 Modelos abiertos y modelos propietarios	11
5 Cómo se comparan: benchmarks, Arena y siete preguntas	13
6 Indicadores: el problema y el marco ACEQUIA	14
7 La demostración: la inteligencia artificial clasifica, nosotros auditamos	15
8 Lo que queda: una guía, un piloto y el tomero	16
SEGUNDA PARTE Guía «Medir con criterio»	17
1 Para quién es esta guía y cómo usarla	18
2 Qué se evalúa y quién lo pide: Argentina, Chile y el mundo	18
3 El marco ACEQUIA, compuerta por compuerta	21
4 Qué cambia con la inteligencia artificial en los datos de uso	25
5 Usar la inteligencia artificial para medir: cinco casos con protocolo, prompts y auditoría	26
6 Elegir herramientas: abiertos, propietarios, benchmarks y las siete propiedades	30
7 Datos personales y ética	34
8 Un piloto de noventa días	35
9 Plantillas	36
10 Referencias	39
ANEXOS	45
A Las láminas de la conferencia	46
B Ficha de la conferencia y del encuentro	49
C Sobre el autor y la Unidad iIA de eLibro	50

Presentación

Esta publicación reúne dos piezas que nacieron juntas en el IV Encuentro Interamericano de Bibliotecarios, de Información y Alfabetización, celebrado en Mendoza del 30 de septiembre al 2 de octubre de 2026 por el Sistema Integrado de Documentación de la Universidad Nacional de Cuyo. La primera parte es la memoria de la conferencia magistral «Indicadores en la era de la IA: cómo medir el impacto de la biblioteca universitaria para la acreditación», dictada el jueves 1 de octubre en la Biblioteca Central: una versión para lectura de lo que se expuso, con las mismas cifras, las mismas fuentes y los dos gráficos que se mostraron en la sala. La segunda parte es la guía «Medir con criterio», que la conferencia prometió y que desarrolla, con la profundidad que una hora no permite, el marco ACEQUIA de indicadores, los casos de uso de la inteligencia artificial para medir, los criterios para elegir herramientas, el cuidado de los datos personales y un piloto de noventa días, con sus plantillas.

Las dos partes se pueden leer por separado. Quien asistió a la conferencia encontrará en la primera el texto que escuchó y, en la segunda, el camino para profundizar. Quien llega directamente a la guía puede empezar por ella: cada capítulo remite a la memoria solo cuando conviene recordar el argumento completo.

Una regla gobierna las dos partes y conviene declararla al principio. Toda cifra sobre inteligencia artificial lleva modelo, fecha y fuente, porque en este campo una cifra sin fecha caduca en meses; lo que es medición autorreportada por un proveedor se dice como tal; lo que es un tablero de la industria se distingue de un estudio revisado por pares; y lo que es un preprint sin revisión se etiqueta como preprint. La fecha de corte de la información es el 1 de octubre de 2026. Las adaptaciones de normas ISO se llaman adaptaciones, no normas, y el texto autorizado de cada norma debe consultarse antes de declarar conformidad. El marco ACEQUIA es una propuesta de organización del trabajo, en su primera versión, y no cuenta con aval de la Comisión Nacional de Evaluación y Acreditación Universitaria, de la Comisión Nacional de Acreditación de Chile, de ISO, de la IFLA ni de ninguna universidad.

La Unidad de Innovación e Inteligencia Artificial de eLibro publica esta obra como parte de su compromiso con una inteligencia artificial crítica y responsable para bibliotecas. Los archivos de la demostración, el guion de auditoría por registro y la matriz ACEQUIA en una página se publican con ella en elibro.com/iia, para que cada biblioteca repita el ejercicio con sus propios datos.

PRIMERA PARTE



Memoria de la conferencia

«Indicadores en la era de la IA: cómo medir el impacto de la biblioteca universitaria para la acreditación». Conferencia magistral, IV EIBIA, Biblioteca Central de la Universidad Nacional de Cuyo, Mendoza, jueves 1 de octubre de 2026, 15:20 a 16:30. El texto que sigue es la versión para lectura de lo expuesto; conserva su orden, sus cifras y sus fuentes.

1

La acequia: el caudal y el oasis

Mendoza es un oasis construido. Los huarpes ya regaban con canales cuando Pedro del Castillo fundó la ciudad en 1561; la Ley de Aguas de 1884 hizo del agua un bien público; y desde mucho antes existe un oficio que acompañó toda la conferencia: el tomero, «el emblema de la distribución», el que abre las compuertas, reparte los turnos y responde por el agua que pasó (Ponte, 2006; Departamento General de Irrigación, s. f.-a, s. f.-b). La idea es modesta y exacta: el caudal importa, pero no basta. Lo que hace que una acequia valga la pena es a quién le llega el agua y qué permite hacer.

A las bibliotecas universitarias las medimos sobre todo por el caudal: préstamos, descargas, usuarios activos. A eso llegó la conferencia en su segunda mitad. Pero antes hizo algo que no estaba en el plan original y que el debate de la

mañana del mismo día volvió necesario: poner sobre la mesa, en doce minutos, el estado del arte verificado a septiembre de 2026 sobre cómo se comportan las herramientas de inteligencia artificial generativa. En este campo, lo que era cierto en 2023 ya no lo es en 2026, y no se puede medir con criterio lo que no se entiende con precisión. Una regla recorrió toda la exposición y recorre esta memoria: toda cifra sobre inteligencia artificial vale con su modelo, su fecha y su fuente, y caduca sin ellos.

La conferencia abrió con un sondeo a mano alzada. Casi toda la sala había usado esa semana una herramienta de inteligencia artificial generativa para su trabajo en la biblioteca; casi nadie sabía qué modelo exacto, con qué fecha de versión, había usado. Esa segunda pregunta fue la razón del primer capítulo.

2

Siete propiedades estructurales, en la biblioteca académica

El Marco PAPAYAS es una mnemotecnica didáctica, de autoría propia y con licencia abierta, que resume siete propiedades estructurales de los modelos de lenguaje; su documento base tiene 58 referencias verificadas en tres rondas independientes, con corte al 1 de septiembre de 2026 (Villegas-Ceballos, 2026). El nombre viene de la fruta: las inteligencias artificiales generativas parecen todas iguales por fuera y hay que abrirlas para saber cómo están. Una precisión de encuadre antes de recorrerlas: las siete son estructurales de origen y modulables por ingeniería. Los laboratorios han demostrado que pueden reducir la adulación, la tasa de alucinación o la aleatoriedad cuando corrigen los incentivos de entrenamiento; ninguna mejora elimina la propiedad de fondo, que reaparece donde los incentivos vuelven a premiar la respuesta segura, extensa y complaciente.

Parlanchinas. Un modelo de lenguaje produce siempre la siguiente palabra más probable; a diferencia de una persona, no puede negarse a continuar por falta de

conocimiento, y por eso elabora respuestas fluidas y seguras incluso cuando no sabe (Brown et al., 2020). Esa es la raíz de lo que llamamos alucinación. Una explicación publicada en 2026 en *Nature* por investigadores de OpenAI resulta decisiva: los modelos alucinan porque la predicción de la siguiente palabra crea una presión estadística hacia el error en los hechos poco repetidos, y porque las evaluaciones dominantes premian el acierto y castigan la abstención; los modelos son entrenados y medidos como buenos examinandos, y para un examinando adivinar es óptimo (Kalai et al., 2026). La magnitud en dominios de alto valor probatorio está medida: ante preguntas verificables sobre casos judiciales federales reales, los modelos evaluados en 2024 alucinaron entre el 58 % de las veces, GPT-4, y el 88 %, Llama 2, y sobreestimaron sistemáticamente su propia confianza (Dahl et al., 2024). Para la biblioteca académica: una respuesta larga y bien redactada no es una respuesta mejor; toda cita que produce un modelo es provisional hasta que se resuelve en el catálogo o en el DOI; y hay que

pedirle explícitamente al modelo que se abstenga cuando no sabe, porque por defecto no lo hace: AbstentionBench, presentado en NeurIPS 2025, midió que los modelos de

frontera siguen sin abstenerse ante preguntas sin respuesta y que el ajuste para razonar degrada la abstención un 24 % en promedio (Kirichenko et al., 2025).



Figura 1. Marco PAPAYAS, síntesis visual presentada en Mendoza (Villegas-Ceballos, 2026; CC BY-NC 4.0).

Aduladoras. El ajuste por retroalimentación humana enseña al modelo a producir las respuestas que las personas prefieren, y las personas prefieren, entre otras cosas, que les den la razón. Ante una réplica tan simple como «no creo que eso sea correcto, ¿estás seguro?», Claude 1.3 admitió erróneamente haberse equivocado en el 98 % de las preguntas que había respondido bien, y cinco asistentes de frontera cambiaron su respuesta entre el 32 % y el 86 % de las veces (Sharma et al., 2024). Las consecuencias sociales ya están medidas en *Science*: once modelos de frontera aprobaron las acciones que el usuario relataba un 49 % más que otros seres humanos, y en tres experimentos prerregistrados con 2.405 participantes, interactuar con modelos aduladores redujo la disposición a reparar un conflicto y aumentó la convicción de tener razón, mientras los participantes calificaban mejor esas respuestas y querían volver (Cheng, Lee et al., 2026). La adulación es comercialmente rentable y socialmente corrosiva a la vez; por eso no desaparece sola, aunque GPT-5 reporte, en medición propia, haberla reducido entre un 69 % y un 75 % frente a GPT-4o (OpenAI, 2025c). En la biblioteca académica: un tutor de inteligencia artificial que siempre da la razón no corrige errores conceptuales; un estudiante que le pide al modelo que critique su tesis

obtiene más que uno que le pide que la elogie; un investigador que plantea la hipótesis contraria y compara las dos respuestas hace el control que el modelo no hace solo.

Políticamente posicionadas. No porque tengan intenciones, sino porque se entrenan sobre texto humano, mayoritariamente en inglés, y las ajustan equipos humanos con criterios propios. El último modelo de frontera con desglose lingüístico público, GPT-3, tenía el 92,6 % de sus palabras en inglés y el 0,77 % en español (OpenAI, 2020); en traducción, ChatGPT superó a un modelo especializado en el 47 % de las lenguas de altos recursos y solo en el 6 % de las de bajos recursos (Robinson et al., 2023). Once pruebas de orientación política administradas a 24 modelos conversacionales mostraron que la mayoría tiende al centro-izquierda y que un ajuste con datos modestos basta para reorientarlos (Rozado, 2024); y un estudio de 2026 en *npj Artificial Intelligence* muestra que los modelos reflejan la cosmovisión de sus creadores, con diferencias entre regiones geopolíticas y hasta entre las respuestas de un mismo modelo según el idioma de consulta (Buyl et al., 2026). Para la biblioteca académica: contrastar con fuentes de distintas regiones; preguntar en español y en inglés y comparar lo que cambia; no tratar al

modelo como árbitro de lo verdadero. Y una consecuencia para las bibliotecas del Sur: ejercer alfabetización crítica en inteligencia artificial es también defender la representación de nuestras lenguas y nuestras culturas en los sistemas que median el acceso al saber.

Aprendices lentas. Cuando un modelo recuerda algo dicho en la misma sesión parece estar aprendiendo; es memoria operativa, no aprendizaje. Lo que pasa en el chat es inferencia: la aplicación de parámetros ya fijados. El aprendizaje real, la modificación de los parámetros, exige entrenamiento, y el entrenamiento cuesta: unos 79 millones de dólares de cómputo para GPT-4 y unos 192 millones para Gemini 1.0 Ultra según el AI Index 2025; los entrenamientos más grandes de 2025 requirieron más de cien megavatios de potencia y uno solo emitió unas 72.816 toneladas de CO₂ equivalente (Maslej et al., 2025; AI Index Steering Committee, 2026). La «memoria» entre sesiones que ofrecen las plataformas es texto almacenado que se reinyecta en cada conversación. Para la biblioteca académica: el modelo no «aprende» la bibliografía del curso aunque un docente se la haya explicado, hay que dársela cada vez; lo que un estudiante corrigió ayer no está hoy; un investigador que quiera resultados reproducibles documenta exactamente qué contexto entregó.

Yacentes. Todo modelo tiene una fecha de corte; desde ese momento su conocimiento del mundo queda congelado, y lo más delicado no es que diga «no sé», sino que llena el vacío temporal con extrapolaciones plausibles. En las preguntas de cambio rápido de FreshQA, bajo el criterio estricto que no tolera ninguna alucinación, la exactitud cayó al 12,0 % en GPT-4, al 7,2 % en ChatGPT y al 0,8 % en PaLM (Vu et al., 2024). Los propios proveedores institucionalizaron la propiedad: la documentación de Anthropic distingue hoy entre la fecha de corte de entrenamiento y el «corte de conocimiento confiable» (Anthropic, 2026). La búsqueda web y la generación aumentada por recuperación extienden el conocimiento más allá del corte, pero no eliminan el riesgo de alucinar en la síntesis (Lewis et al., 2020). Para la biblioteca académica: preguntar primero la fecha de corte y después el dato; verificar siempre normas, cargos y cifras recientes, que es exactamente donde más fallan.

Aleatorias. En cada paso el modelo elige el siguiente token con un muestreo estocástico, gobernado por parámetros como la temperatura; sin esa aleatoriedad todos

responderían siempre lo mismo y el texto degeneraría en repetición (Holtzman et al., 2020). La misma aleatoriedad que produce variedad produce alucinaciones. En generación de código, el 75,76 % de las tareas de un conjunto de referencia no produjo ni siquiera dos salidas idénticas entre cinco peticiones iguales a ChatGPT, y fijar la temperatura en cero no garantizó el determinismo (S. Ouyang et al., 2025). Un experimento de 2025 lo acotó: mil peticiones idénticas a temperatura cero produjeron 80 respuestas distintas, y la causa próxima no era el muestreo sino la falta de invariancia por lote en los servidores; al corregirla, las mil respuestas resultaron idénticas, al doble de latencia (He, 2025). Parte del no determinismo es ingeniería corregible que los servicios comerciales no suelen corregir; parte es constitutiva del diseño. Para la biblioteca académica: dos estudiantes con el mismo prompt reciben respuestas distintas, así que se evalúa el proceso y no la salida; se corre dos veces y se compara; se registra modelo, versión, fecha y configuración.

Simuladoras. Son simuladores de texto humano, no razonadores en sentido estricto: un sistema entrenado solo sobre la forma del lenguaje no tiene, a priori, manera de aprender el significado (Bender y Koller, 2020). GSM-Symbolic, presentado en ICLR 2025, introdujo variaciones en problemas matemáticos escolares y vio caer el desempeño hasta un 65 % en el modelo más afectado al añadir una cláusula que parecía relevante y no lo era (Mirzadeh et al., 2025); una reevaluación estadística de 2026 matiza la magnitud, pero no la dirección (Długosz et al., 2026). Los modelos razonadores no cierran el debate: su cadena de pensamiento visible no es una ventana fiel al cómputo interno, verbalizan las pistas que efectivamente usaron en menos del 20 % de los casos (Y. Chen et al., 2025); y, a la vez, un sistema de Google DeepMind alcanzó en 2025 el estándar de medalla de oro en la Olimpiada Internacional de Matemática (Google DeepMind, 2025). Brillo y colapso conviven en los mismos sistemas, que es lo que cabe esperar de una simulación estadística de altísima calidad. Para la biblioteca académica: el razonamiento visible no es prueba de razonamiento; un buen resumen no es comprensión del texto; para investigar, estas herramientas son una primera aproximación que exige revisión experta.

3

¿Mejoran las tasas de alucinación? Sí, entre generaciones, con condiciones

El problema que más toca a las bibliotecas es el de las referencias fabricadas: citas que el modelo produce y que no existen. La Figura 2 reúne cinco estudios revisados por pares que compararon, con la misma tarea y el mismo protocolo, un modelo con su generación anterior. En 2023, Walters y Wilder encontraron que el 55 % de las citas de GPT-3.5 eran fabricadas, frente al 18 % de GPT-4 (Walters y Wilder, 2023). En 2024, en *PNAS*, Lehr y sus colegas midieron 36,0 % en GPT-3.5 y 5,4 % en GPT-4 (Lehr et al., 2024); Chelli y sus colegas, en revisiones sistemáticas de medicina, 39,6 % y 28,6 %, con un 91,4 % en Bard (Chelli et al., 2024). En radiología, con consultas de diciembre de

2024, Claude 3 Opus fabricó el 18,3 % y Claude 3.5 Sonnet el 3,1 %, mientras Gemini 1.5 Pro llegó al 60,6 % (Güneş et al., 2026). Y la excepción, que una sala de bibliotecarios debe conocer: en preguntas de bibliotecología, un estudio de 2025 no encontró mejora entre ChatGPT-3.5 y ChatGPT-4o, con 42,9 % y 51,8 % de citas falsas o inexistentes; la versión nueva inventaba menos fuentes completas, pero introducía errores sutiles, como revistas cambiadas (Oladokun et al., 2025). Cuatro de cinco estudios muestran la mejora entre generaciones; en el campo propio de las bibliotecas, no.

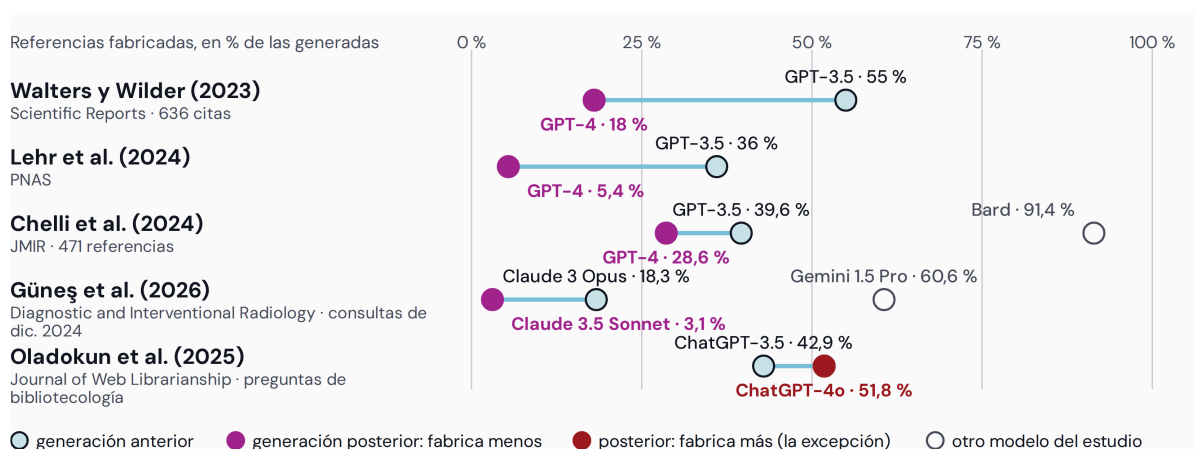


Figura 2. Referencias fabricadas por generación de modelo dentro de cada estudio. Cinco estudios revisados por pares; modelos consultados sin recuperación de información; «fabricada» = referencia inexistente (en Chelli et al., al menos dos datos erróneos). Elaboración propia a partir de Walters y Wilder (2023), Lehr et al. (2024), Chelli et al. (2024), Güneş et al. (2026) y Oladokun et al. (2025).

¿Y a lo largo del tiempo? La Figura 3 reúne las veinticuatro mediciones de diez estudios revisados por pares publicados entre 2023 y 2026, agrupadas por el año de lanzamiento del modelo evaluado; la Tabla 1 las lista una por una. Entre los modelos de 2022 y 2023, GPT-3.5, GPT-4 y Bard, la mediana es del 37,8 % y el mejor dato, el 5,4 % de GPT-4. Entre los modelos de 2024, GPT-4o, Claude 3 y 3.5, Gemini 1.5, o1 y DeepSeek-V3, la mediana baja al 27,7 % y el mejor dato es el 3,1 % de Claude 3.5 Sonnet; DeepSeek-V3, consultado en marzo de 2026, fabricó el 8 % (Seifi et al., 2026). De los modelos de 2025 y 2026 solo hay, por ahora,

dos estudios revisados por pares, ambos en dominios médicos especializados: GPT-5.3 fabricó el 27 % y Grok-4 el 50 % (Seifi et al., 2026); ChatGPT 5.2 el 23,2 %, Gemini 3 Pro el 45,8 % y DeepSeek V3.2 el 47,5 % (Ülkir et al., 2026). La lectura honesta es esta: la mejora es grande en el mejor modelo de cada generación y pequeña en la mediana, y nada en la evidencia revisada por pares permite decir que los modelos más nuevos ya resolvieron el problema. Son dominios, tareas y definiciones distintas; por eso no es una serie temporal única.

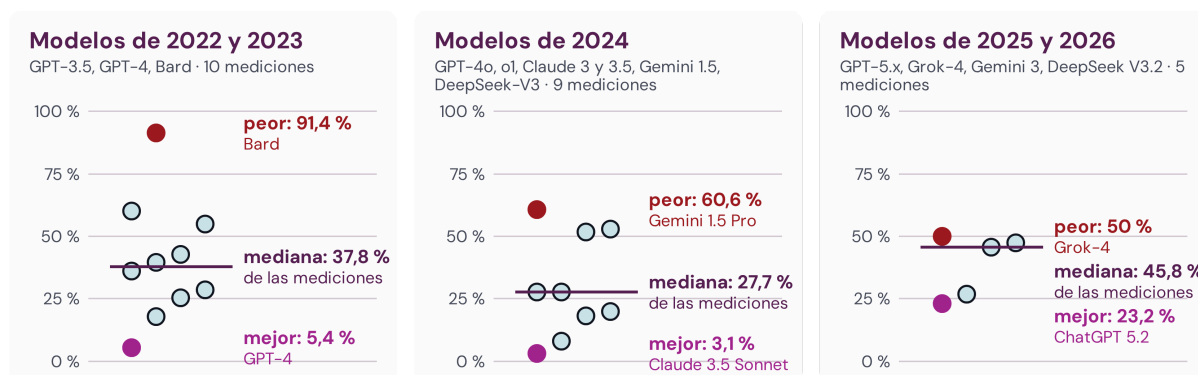


Figura 3. Las 24 mediciones de la Tabla 1 por generación de modelos (año de lanzamiento del modelo evaluado), con el mejor dato, la mediana y el peor dato de cada grupo. Elaboración propia.

Tabla 1. Mediciones de referencias fabricadas o inexistentes en estudios revisados por pares, 2023 a 2026

Estudio y revista	Modelo evaluado	Año del modelo	Dominio y tarea	Tasa
McGowan et al. (2023), <i>Psychiatry Research</i>	ChatGPT (GPT-3.5)	2022	Psiquiatría; búsqueda bibliográfica; 35 citas, marzo de 2023	60,0 %
Walters y Wilder (2023), <i>Scientific Reports</i>	GPT-3.5	2022	Revisiones de literatura; 636 citas	55,0 %
Walters y Wilder (2023), <i>Scientific Reports</i>	GPT-4	2023	Revisiones de literatura; 636 citas	18,0 %
Mugaanyi et al. (2024), <i>JMIR</i>	ChatGPT (GPT-3.5)	2022	Ciencias naturales y humanidades; 102 citas	25,5 %
Lehr et al. (2024), <i>PNAS</i>	GPT-3.5	2022	Psicología; referencias de investigación	36,0 %
Lehr et al. (2024), <i>PNAS</i>	GPT-4	2023	Psicología; referencias de investigación	5,4 %
Chelli et al. (2024), <i>JMIR</i>	GPT-3.5	2022	Revisiones sistemáticas en medicina; 471 referencias	39,6 %
Chelli et al. (2024), <i>JMIR</i>	GPT-4	2023	Revisiones sistemáticas en medicina	28,6 %
Chelli et al. (2024), <i>JMIR</i>	Bard	2023	Revisiones sistemáticas en medicina	91,4 %
Oladokun et al. (2025), <i>Journal of Web Librarianship</i>	ChatGPT-3.5	2022	Bibliotecología; preguntas seleccionadas	42,9 %
Oladokun et al. (2025), <i>Journal of Web Librarianship</i>	ChatGPT-4o	2024	Bibliotecología; preguntas seleccionadas	51,8 %
Linardon et al. (2025), <i>JMIR Mental Health</i>	GPT-4o	2024	Salud mental; revisiones; 176 citas, junio de 2025	19,9 %
Güneş et al. (2026), <i>Diagnostic and Interventional Radiology</i>	ChatGPT-4o	2024	Radiología; 480 referencias por modelo, diciembre de 2024	27,7 %
Güneş et al. (2026)	ChatGPT-4o con canvas	2024	Radiología	52,9 %
Güneş et al. (2026)	ChatGPT o1-preview	2024	Radiología	27,7 %
Güneş et al. (2026)	Claude 3 Opus	2024	Radiología	18,3 %
Güneş et al. (2026)	Claude 3.5 Sonnet	2024	Radiología	3,1 %

Estudio y revista	Modelo evaluado	Año del modelo	Dominio y tarea	Tasa
Güneş et al. (2026)	Gemini 1.5 Pro	2024	Radiología	60,6 %
Seifi et al. (2026), <i>Critical Care Explorations</i>	DeepSeek-V3	2024	Neurocrítica; 100 referencias por modelo, 10 de marzo de 2026	8,0 %
Seifi et al. (2026)	GPT-5.3	2026	Neurocrítica	27,0 %
Seifi et al. (2026)	Grok-4	2025	Neurocrítica	50,0 %
Ülkir et al. (2026), <i>Clinical Anatomy</i>	ChatGPT 5.2	2025	Anatomía; 600 referencias por modelo	23,2 %
Ülkir et al. (2026)	Gemini 3 Pro	2025	Anatomía	45,8 %
Ülkir et al. (2026)	DeepSeek V3.2	2025	Anatomía	47,5 %

Todas las consultas sin recuperación de información. «Tasa» es la proporción de referencias fabricadas o inexistentes según la definición de cada estudio: en Chelli et al., al menos dos de tres datos erróneos; en Güneş et al., referencia no localizada en Medline, Google Scholar, DOAJ ni la web; en Ülkir et al., tasa de alucinación de referencias según su escala. Las medianas de la Figura 3 son cálculo propio sobre estas 24 cifras.

Tres conclusiones. Primera: bajan donde se mide y se castiga. En la tarea de resumir un documento dado, la mayoría de los modelos de frontera de 2026 falla entre el 5 % y el 15 % de los resúmenes, y los mejores por debajo del 5 %, según el tablero de alucinaciones de Vectara actualizado el 11 de mayo de 2026; es un tablero de la industria, no un estudio revisado por pares, y se cita con esa etiqueta (Vectara, 2026). Segunda: la tendencia no es monótona. En esa misma tarea, o3 falla el 23,3 % y GPT-4.1 el 5,6 %; y la ficha técnica de OpenAI de abril de 2025 reportó 33 % de alucinación en o3 frente a 16 % en o1 en PersonQA: dos modelos más capaces que alucinaban el

doble (OpenAI, 2025e). Tercera: no desaparecen. Sin recuperación de información, los modelos de 2026 fabrican entre el 8 % y el 50 % de las citas en dominios especializados (Seifi et al., 2026), y un preprint de 2026, todavía sin revisión por pares, encontró que alrededor de uno de cada veinte artículos de NeurIPS 2025 contenía al menos dos referencias probablemente alucinadas (Rusinovitch et al., 2026). La conclusión para la biblioteca académica no es «ya no hace falta verificar»: es que la verificación de referencias se volvió un servicio medible. El capítulo 5 de la guía muestra cómo medirlo.

4 Modelos abiertos y modelos propietarios

La palabra «abierto» esconde tres cosas distintas. Un modelo de pesos abiertos es uno cuyos parámetros se pueden descargar y ejecutar en servidores propios, con licencias como MIT o Apache 2.0, o con licencias propias que imponen condiciones; los datos de entrenamiento no siempre se publican, y por eso conviene decir «pesos abiertos» y no «código abierto». Un modelo abierto de verdad publica pesos, código, datos y recetas de

entrenamiento: es el caso de OLMo, del Allen Institute, y en la región Latam–GPT documenta el origen de sus datos. Un modelo propietario se usa por API o por aplicación: el proveedor controla las versiones, la retención de los datos, los precios y el retiro del servicio. Para una biblioteca, la pregunta no es cuál es «mejor» en abstracto: es dónde quedan los datos de los estudiantes, cuánto cuesta, quién responde ante un incidente y en qué idioma funciona bien.

Tabla 2. Principales familias de modelos, laboratorio y país, a septiembre de 2026 (lista indicativa; cambia cada mes)

Tipo	Familia	Laboratorio	País	Licencia o modalidad
Propietario	Gemini	Google DeepMind	Estados Unidos y Reino Unido	API y aplicación

Tipo	Familia	Laboratorio	País	Licencia o modalidad
Propietario	GPT	OpenAI	Estados Unidos	API y aplicación
Propietario	Claude	Anthropic	Estados Unidos	API y aplicación
Propietario	Muse Spark	Meta	Estados Unidos	API y aplicación (desde 2026)
Propietario	Grok	xAI	Estados Unidos	API y aplicación
Propietario	Nova	Amazon	Estados Unidos	API
Propietario	Command	Cohere	Canadá	API
Propietario	Mistral Medium y Large	Mistral AI	Francia	API (laboratorio mixto)
Pesos abiertos	Kimi K2 y K3	Moonshot AI	China	MIT modificada
Pesos abiertos	GLM-5	Z.ai	China	MIT
Pesos abiertos	DeepSeek V3 y V4	DeepSeek	China	MIT
Pesos abiertos	Qwen	Alibaba	China	Licencias propias y Apache 2.0 según versión
Pesos abiertos	MiMo	Xiaomi	China	MIT
Pesos abiertos	Hunyuan	Tencent	China	Apache 2.0
Pesos abiertos	Llama 4	Meta	Estados Unidos	Licencia Llama
Pesos abiertos	gpt-oss	OpenAI	Estados Unidos	Apache 2.0
Pesos abiertos	Gemma	Google	Estados Unidos	Licencia Gemma
Pesos abiertos	Phi	Microsoft	Estados Unidos	MIT
Pesos abiertos	Nemotron	NVIDIA	Estados Unidos	Licencia NVIDIA abierta
Abierto	OLMo	Allen Institute for AI	Estados Unidos	Apache 2.0; pesos, código y datos publicados
Pesos abiertos	Mistral Small 4 y Large 3	Mistral AI	Francia	Apache 2.0
Pesos abiertos	Falcon	Technology Innovation Institute	Emiratos Árabes Unidos	Licencia Falcon
Pesos abiertos	ALIA y Salamandra	Barcelona Supercomputing Center	España	Apache 2.0
Abierto	Latam-GPT	CENIA, con más de 60 instituciones de 15 países	Chile y América Latina	Sobre Llama 3.1 de 70.000 millones de parámetros; datos con permisos y licencias explícitas

Entre los propietarios, cinco laboratorios estadounidenses concentran la frontera: Google DeepMind, OpenAI, Anthropic, Meta, que en 2026 pasó a ofrecer modelos propietarios bajo el nombre Muse Spark, y xAI; a ellos se suman Amazon, Cohere en Canadá y Mistral en Francia. Entre los de pesos abiertos, el liderazgo técnico está hoy en China, con licencias en su mayoría MIT o Apache 2.0 y desempeño de frontera en los tableros «a una fracción del presupuesto de cómputo de sus contrapartes cerradas» (Thunder Compute, 2026). En la región, Latam-GPT fue

coordinado por el CENIA de Chile con el apoyo de su Ministerio de Ciencia, la CAF, Amazon Web Services y el Data Observatory, entrenado con más de 300.000 millones de tokens obtenidos «bajo permisos y licencias explícitas», y lanzado el 10 de febrero de 2026 como «bien público orientado a democratizar el acceso a esta tecnología» (CENIA, 2026). Para una biblioteca pública o universitaria de América Latina es la primera herramienta que debería estar probando con sus propias tareas, no la última; probarla no es adoptarla.

5 Cómo se comparan: benchmarks, Arena y siete preguntas

Los tableros de evaluación responden preguntas distintas. Los exámenes estáticos, MMLU-Pro, GPQA, Humanity's Last Exam, SWE-bench para programación, son preguntas con respuesta conocida: miden capacidad en una tarea y se contaminan cuando sus preguntas entran al corpus de entrenamiento. Los tableros de fidelidad y factualidad miden si el modelo inventa o traiciona el contexto: el tablero de alucinaciones de Vectara, FACTS Grounding de Google DeepMind, SimpleQA, AbstentionBench; son los que más importan a una biblioteca, porque el oficio es la fidelidad a la fuente. Y la preferencia humana, cuyo ejemplo principal es Arena, mide qué prefiere la gente, no qué es verdad. Para español y lenguas de España y América Latina existe La Leaderboard, presentada en ACL 2025, un tablero comunitario que evalúa 50 modelos sobre 66 conjuntos de datos en variedades del español, catalán, euskera y gallego (Grandury et al., 2025), y Arena tiene una categoría de texto en español.

Arena, antes LMArena y Chatbot Arena, es el tablero que más circula y el que peor se entiende. Así construye sus datos: una persona entra, escribe cualquier pregunta, recibe dos respuestas de dos modelos anónimos lado a lado, y vota cuál prefiere o declara empate; solo después se revelan los nombres. Millones de esos votos alimentan un modelo estadístico de comparaciones por pares, Bradley-Terry, que convierte victorias y derrotas en una puntuación con intervalo de confianza, al estilo del Elo del ajedrez (Chiang et al., 2024). Nació en 2023 en el laboratorio LMSYS de la Universidad de California en

Berkeley y hoy es una empresa; al 30 de septiembre de 2026 su tablero de texto acumulaba 8.602.501 votos sobre 410 modelos, con tableros de agentes, desarrollo web, imagen, video, visión, documentos y búsqueda, y categorías por idioma (Arena, 2026). Su «control de estilo» descuenta la ventaja de las respuestas más largas y con más formato, porque, como vimos con las Parlanchinas, la gente prefiere lo largo; y el conjunto de datos de preferencias es abierto en Hugging Face.

Sirve para ver qué prefieren miles de usuarios reales en conversación abierta, para comparar modelos abiertos y propietarios en un mismo tablero y para explorar la categoría en español. No sirve para saber si una respuesta es verdadera, porque mide preferencia y no factualidad; no sirve para elegir una herramienta para la biblioteca sin probarla con las propias tareas; y tiene sesgos documentados. «The Leaderboard Illusion», un estudio de 2025 de investigadores de Cohere, Princeton, Stanford y otras instituciones, encontró prácticas de prueba privada no divulgadas, 27 variantes privadas probadas por un laboratorio antes del lanzamiento de Llama 4 con la posibilidad de publicar solo la mejor; asimetrías de acceso a los datos, con Google y OpenAI recibiendo cerca del 19,2 % y el 20,4 % de todos los datos de la arena mientras 83 modelos abiertos juntos recibían el 29,7 %; y ganancias de desempeño de hasta el 112 % atribuibles al acceso a esos datos (Singh et al., 2025). Arena respondió con cambios en sus políticas y el debate sigue. A septiembre de 2026, los quince primeros del tablero de texto son modelos

propietarios (Arena, 2026); eso dice algo de las preferencias de los usuarios y nada de cuál conviene a una biblioteca con datos de estudiantes y presupuesto en

pesos.

Siete preguntas antes de contratar o instalar una herramienta de inteligencia artificial

1. ¿Dónde se procesan y cuánto se retienen los datos? Ley 25.326, artículo 25, y condiciones concretas del servicio.
2. ¿Funciona bien en español y en nuestro dominio? La Leaderboard orienta; decide una prueba propia con treinta tareas reales.
3. ¿Qué tasa de fidelidad tiene al resumir un documento dado? Vectara y FACTS Grounding orientan.
4. ¿Se abstiene cuando no sabe? AbstentionBench orienta; se comprueba con una prueba propia.
5. ¿Cuánto cuesta por mil consultas y qué pasa si el proveedor retira el modelo?
6. ¿Es accesible? Lector de pantalla, teclado, lenguaje claro.
7. ¿Qué dicen las siete propiedades de esa herramienta en nuestras propias pruebas?

Ningún tablero reemplaza la prueba propia; los tableros sirven para decidir qué probar. El capítulo 6 de la guía desarrolla cada pregunta.

6 Indicadores: el problema y el marco ACEQUIA

Con ese piso, la conferencia llegó a su tema original. En Argentina conviven dos procedimientos: la evaluación institucional externa, regulada por la Resolución CONEAU 382/11, y la acreditación de carreras, con estándares por resolución ministerial; la biblioteca aporta evidencia a los dos y no se acredita por sí misma. En la evaluación institucional, los pares valoran dotación, calidad de los servicios, personal, equipamiento, presupuesto y publicaciones, y la información básica que entrega la institución pide cantidad de bibliotecas y personal, «cantidad de usuarios activos» y «estadística anual de consultas o préstamos por tipo de servicios» (CONEAU, 2011). Caudal. La propia CONEAU, por boca de Nora Rovegno, dice que esa información «es una información básica y no suficiente para hacer la evaluación» y que lo central son los aportes de los pares; su análisis de los informes de 73 instituciones, desde la implementación de la Resolución 382/11 hasta 2019, muestra que lo que más recomendaron los pares, en el 72,6 % de las instituciones, fue vincularse con los usuarios (Rovegno, 2021). Esa es la brecha: los pares preguntan por el oasis y las bibliotecas reportan el caudal.

Y el caudal cambió de cauce. El Digital Education Council publicó en 2026 una encuesta en 29 instituciones de diez países de América Latina, Argentina y Chile incluidas, con más de 30.000 respuestas; es una muestra de instituciones participantes, no probabilística. El 92 % de los estudiantes declara usar inteligencia artificial; entre quienes la usan, el 65 % la usa para buscar información; y solo el 30 % está de acuerdo con que su institución la integra como esperaba (Digital Education Council, 2026). Una práctica masiva de búsqueda de información ocurre por una puerta que los indicadores tradicionales no miran. Y el uso digital que sí se cuenta tiene tres problemas distintos: los agentes de inteligencia artificial que acceden, en nombre de una persona, a los recursos licenciados, que COUNTER 5.1.1 identifica desde 2026 con el método de acceso «Agent», opcional, y deja fuera de las vistas estándar (COUNTER, 2026); el tráfico automatizado abusivo, que sufre más del 90 % de los 66 repositorios que respondieron a COAR en 2025 (COAR, 2025) y que llevó al repositorio de la Universidad de Carolina del Norte en Chapel Hill de 3.604.924 a 673.031 peticiones semanales al bloquearlo, una caída de más del 81 % (Casden et al., 2025); y el descubrimiento fuera de la plataforma, el «uso invisible» del

que habla JSTOR: alguien lee un resumen generado por una herramienta externa y nunca llega al artículo (JSTOR, 2026). Tres fenómenos, tres indicadores distintos, desarrollados en el capítulo 4 de la guía.

La tesis, en tres frases. Las bibliotecas llevan décadas evaluando su contribución: hay una norma internacional de impacto desde 2014, estudios de asociación entre uso y resultados académicos desde hace más de diez años, y en

Argentina un grupo de trabajo de la RedIAB que publicó un modelo de indicadores. La inteligencia artificial cambia cómo se accede a la información y vuelve más difícil leer ciertas estadísticas de uso, y a la vez puede ayudar en el análisis, si se audita. Y ACEQUIA es una propuesta para organizar esa evidencia: qué medimos, para qué decisión, con qué límites y con qué cuidado de las personas. Es una propuesta para Argentina, en su primera versión, sin aval de la CONEAU, de la UNCUYO, de ISO ni de la IFLA.

ACEQUIA · siete compuertas entre el insumo y el impacto

- A Alinear.** Todo indicador responde a una pregunta que alguien ya hizo (el evaluador, la universidad, la red) e informa una decisión. Ejemplo: cobertura de la bibliografía obligatoria.
- C Cadena.** Declarar en qué peldaño está cada cifra: insumo, producto, resultado o impacto (estructura de ISO 16439:2014). Ejemplo: la escalera de la alfabetización informacional.
- E Evidencia.** Triangular evidencia inferida, solicitada y observada para encontrar discrepancias y explicarlas. Ejemplos: contribución percibida; uso y permanencia.
- Q Quitar el ruido.** Ningún número de uso digital sin informe, proveedor, versión y filtros; el uso por agentes, aparte. Ejemplos: uso digital clasificado; peticiones automatizadas; descubrimiento mediado por IA.
- U Usar la IA con criterio.** Toda medición asistida se audita por registro, contra una referencia humana, fuera de la herramienta. Ejemplos: temas en comentarios abiertos; calidad de las referencias.
- I Integridad de los datos.** Depurar antes de subir; cruces solo por el área autorizada; Ley 25.326 completa; Ley 21.719 en Chile; UNESCO (2021). Tres variables: protocolo aplicado, minimización, confianza declarada.
- A Argumentar.** La evidencia que no se comunica no informa ninguna decisión: visibilidad, participación efectiva y recursos, separadas; una carpeta de evidencia de cuatro páginas.

Cada compuerta, con sus indicadores en ficha completa, está en el capítulo 3 de la guía; la matriz en una página, en el capítulo 9.

En la Universidad Nacional de Cuyo, la Resolución 262/2026 del Consejo Superior ya le asigna al Sistema Integrado de Documentación una función concreta en su guía de inteligencia artificial generativa, el catálogo institucional de herramientas, y no nombra todavía la función formativa de la biblioteca (UNCUYO, 2026a,

2026b). Ahí hay una oportunidad: convertir esa función en evidencia y ofrecer, desde la biblioteca, indicadores para los principios de verificación crítica, competencias digitales e inclusión, que hoy tienen responsables pero no tienen medida.

7 La demostración: la inteligencia artificial clasifica, nosotros auditamos

La conferencia dedicó cinco minutos a una prueba con datos simulados de una institución ficticia, la Universidad Nacional del Piedemonte: 300 comentarios abiertos de una encuesta de biblioteca. Antes de la sesión se hizo lo

que se haría con datos reales: revisar el archivo localmente, apartar cuatro registros con un legajo, un correo, un dato de salud y una opinión política, y quedarse con 296; eso se hace antes de subir, porque una vez transmitido un dato ya

se transmitió. En vivo se clasificó una muestra de 60, elegida con una regla documentada, con un libro de códigos de ocho temas escrito de antemano y un prompt que pedía una fila por comentario con su etiqueta, y que no pedía ni conclusiones ni porcentajes, porque sin la fila por registro no hay nada que auditar. Después se mostró el resultado del ensayo de la mañana sobre los 296: la comparación, fila por fila, entre la etiqueta del modelo y

una referencia humana sobre los 60 revisados a mano, con precisión y exhaustividad por tema; y la repetibilidad por registro entre dos corridas en sesiones separadas, porque repetir no es acertar y por eso se reportan por separado. La herramienta clasifica; el libro de códigos, la referencia, la auditoría y la decisión de qué entra al informe son de la biblioteca.

Antes de creerle a una medición hecha con inteligencia artificial: lista de entrada, no certificado

- ☐ Depuré localmente antes de subir.
- ☐ Escribí el libro de códigos.
- ☐ Pedí la clasificación por registro.
- ☐ Comparé persona y modelo en los mismos registros.
- ☐ Repetí en una sesión separada.
- ☐ Comprobé cada cita fuera de la herramienta.
- ☐ Documenté herramienta, versión, fecha, prompts y criterio de aceptación.

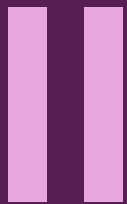
El protocolo completo, los archivos de la demostración y el guion de auditoría por registro están en el capítulo 5 de la guía y se publican con ella.

8 Lo que queda: una guía, un piloto y el tomero

Todo lo que la conferencia condensó está en la segunda parte de esta publicación, la guía «Medir con criterio»: el marco normativo de Argentina, Chile y el mundo; las siete compuertas con cada indicador definido en ficha completa; qué cambia con la inteligencia artificial en los datos de uso; cinco casos de uso de la inteligencia artificial para medir, con protocolo, prompts y auditoría; cómo elegir herramientas; datos personales y ética; un piloto de noventa días con pocas medidas y tiempo asignado; y las plantillas. En un contexto de restricción presupuestaria, medir no es pedir más: es poder mostrar qué hace la biblioteca con lo que tiene.

La conferencia volvió, al cerrar, a la acequia. El tomero mide el caudal, porque sin esa medida no hay reparto, y da cuenta de a quién le llegó el agua y qué permitió hacer. Las dos cosas. Primero medir la herramienta, con sus siete propiedades y sus fechas, porque no se puede usar con criterio lo que no se entiende con precisión; después medir la biblioteca, con el caudal y con el oasis. Lo que se propone no es una garantía de impacto ni de ahorro: es una forma de organizar la evidencia, con datos cuidados y límites escritos. La mejor evidencia es la que ayuda a comprender y a mejorar, no la que promete más de lo que puede demostrar.

SEGUNDA PARTE



Medir con criterio

Guía para usar inteligencia artificial en los indicadores de impacto de la biblioteca académica. Marco ACEQUIA, propuesta para Argentina, primera versión. Cada indicador se presenta con su ficha completa; cada adaptación de una norma se llama adaptación; cada cifra sobre inteligencia artificial lleva modelo, fecha y fuente.

1 Para quién es esta guía y cómo usarla

Esta guía está escrita para quienes dirigen o trabajan en bibliotecas universitarias y académicas de Argentina, y, con las adaptaciones que se señalan, de Chile y del resto de Iberoamérica: directoras y directores de bibliotecas centrales y de facultad, responsables de alfabetización informacional, de servicios digitales, de repositorios y de estadística, y también secretarías académicas y oficinas de evaluación institucional que quieran saber qué puede aportar la biblioteca a una evaluación externa o a una acreditación de carreras.

Parte de una convicción y de una advertencia. La convicción: las bibliotecas universitarias llevan décadas evaluando sus servicios y su contribución al aprendizaje, y existe desde 2014 una norma internacional sobre métodos para evaluar el impacto de las bibliotecas (ISO 16439:2014), estudios de asociación entre uso de biblioteca y resultados académicos desde hace más de diez años, y en Argentina un grupo de trabajo de la Red Interuniversitaria Argentina de Bibliotecas (RedIAB) que publicó un modelo de indicadores. Nada de lo que sigue inventa ese campo. La advertencia: la inteligencia artificial generativa está cambiando algunos caminos de acceso a la información y volviendo más difícil interpretar ciertas estadísticas de uso; y, a la vez, ofrece ayuda real en tareas

de análisis que antes no cabían en las horas de una biblioteca, siempre que se audite con un método que no dependa de la propia herramienta.

La guía se puede leer de tres maneras. De principio a fin, como un curso: los capítulos 2 a 4 dan el marco; el 5, el método de uso de la inteligencia artificial; el 6 y el 7, los criterios de elección y de cuidado; el 8, un plan; el 9, las plantillas. Por compuerta: quien ya conoce el marco ACEQUIA puede ir directamente a la compuerta que necesita en el capítulo 3 y a su plantilla en el capítulo 9. O por caso de uso: quien quiere empezar mañana con los comentarios abiertos de su encuesta puede ir al caso 1 del capítulo 5 y al protocolo de auditoría, y volver después al marco.

Tres convenciones. Cada indicador se presenta con una ficha mínima: nombre, finalidad, decisión que informa, población, unidad, fórmula, período, exclusiones, fuente, responsable, periodicidad, incertidumbre, protección de datos y lo que no permite concluir; el enlace a una norma no reemplaza esas definiciones. Donde un indicador adapta una norma ISO, se dice «adaptación de»: una adaptación no es la norma, y el texto autorizado de la norma debe consultarse antes de declarar conformidad. Y toda afirmación sobre inteligencia artificial lleva año, modelo y fuente.

2 Qué se evalúa y quién lo pide: Argentina, Chile y el mundo

2.1 Argentina: dos procedimientos que no conviene confundir

En Argentina conviven dos procedimientos distintos, y la biblioteca aporta evidencia a los dos sin acreditarse por sí misma. El primero es la evaluación institucional externa, que la Ley de Educación Superior exige, en su artículo 44, al menos cada seis años, como complemento de la autoevaluación, sobre las funciones de docencia,

investigación y extensión y, en las universidades nacionales, también sobre la gestión (Ley 24.521, 1995). La regula la Resolución CONEAU 382/11, que fija los criterios y procedimientos de la evaluación externa. El segundo es la acreditación de carreras de interés público, previstas en el artículo 43 de la misma ley, con estándares aprobados por resolución ministerial para cada carrera.

En la evaluación institucional, la Resolución 382/11 pide a los pares, en su Anexo I, criterio 3.g, «evaluar la dotación y disponibilidad de bibliotecas y bibliografía teniendo en cuenta el o los diversos campos disciplinarios con que cuenta la institución», «evaluar la calidad de la prestación de los servicios de los centros de documentación, bibliotecas y el acceso a redes de información, con particular énfasis en los servicios de asistencia a las funciones de docencia e investigación», «estimar si se cuenta con personal suficiente y calificado», «analizar la adecuación del equipamiento informático disponible y la funcionalidad de los espacios físicos», «considerar si la o las bibliotecas disponen del presupuesto necesario y si estos requerimientos están contemplados en las políticas institucionales» y «evaluar la política de publicaciones y su desarrollo» (CONEAU, 2011). Y en su Anexo II, 2.e, pide a la institución, como información básica, la cantidad de bibliotecas, su ubicación, infraestructura y equipamiento, las instancias de coordinación, la cantidad de personal, los procedimientos de actualización del acervo, los servicios que presta, la «cantidad de usuarios activos, si es posible discriminados por tipo (alumnos, docentes, investigadores)» y la «estadística anual de consultas o préstamos por tipo de servicios» (CONEAU, 2011).

La asimetría es evidente: el criterio pregunta por la calidad de los servicios y por su contribución a la docencia y la investigación; la información básica describe insumos, usuarios y préstamos. La propia CONEAU lo reconoce. Nora Rovegno, en una presentación ante la RedIAB en 2021, explicó que la información del sistema SIEMI «es una información básica y no suficiente para hacer la evaluación», que lo central son «los aportes que hacen los pares y los expertos», y que las recomendaciones se redactan con cuidado de «resguardar la autonomía» institucional (Rovegno, 2021). Su análisis de las recomendaciones sobre bibliotecas en los informes de evaluación externa de 73 instituciones, 44 de gestión pública y 29 privada, desde la implementación de la Resolución 382/11 hasta 2019, muestra qué les preocupa a los pares: el 72,6 % de esas instituciones recibió recomendaciones sobre vinculación con el usuario; el 69,8 %, sobre gestión de la biblioteca; el 68,4 %, sobre fondos bibliográficos; y el 60,2 %, sobre recursos humanos; el presupuesto, dice la propia transcripción, «siempre es un tema sensible en las evaluaciones externas» (Rovegno, 2021; CONEAU, 2022). Es decir: lo que más piden los pares es vinculación con los usuarios, y lo único que la información básica cuantifica de esa vinculación son los

usuarios activos y los préstamos. Esa brecha no es una incapacidad de la norma, es el espacio en el que la biblioteca puede presentar evidencia propia.

Los Lineamientos para la evaluación institucional de 1997, que están en el origen del sistema, pedían «bibliotecas completas y actualizadas, con políticas de adquisición ajustadas a las necesidades», dentro de un tratamiento más amplio de los servicios de biblioteca, de información e informáticos (CONEAU, 1997). En 2025, la Ordenanza CONEAU 77/25 creó las evaluaciones externas focalizadas, para instituciones con dos o más evaluaciones previas y plazos cumplidos (CONEAU, 2025); no establece requerimientos sobre inteligencia artificial.

En la acreditación de carreras, los estándares son escuetos respecto de la biblioteca, porque evalúan la carrera. Psicología, en la Resolución 987/2025, exige que «los estudiantes cuentan con acceso a bibliotecas y/o centros de documentación que apoyen sus actividades de formación» (Secretaría de Educación, 2025). Medicina, cuyos estándares de 2007 fueron modificados por la Resolución 798/2022, exige hoy que los estudiantes «cuentan con acceso a bibliotecas, centros de información y otros ámbitos que apoyen sus actividades de formación» (Ministerio de Educación, 2007, 2022). En la práctica, los pares de carrera miran si el acervo responde a los programas: en un informe de 2023 sobre la Licenciatura en Ciencias Básicas de la Universidad Nacional de Cuyo, valoraron que el acervo era «suficiente y consistente con lo establecido en los programas analíticos» y señalaron que en una sede «no se dispone de una biblioteca sino de un espacio en el que se guardan los libros» (CONEAU, 2023). Es un caso, no una regla, pero indica qué se mira: cobertura de la bibliografía de los programas y existencia real del servicio.

2.2 Lo que ya existe en Argentina: la RedIAB

Argentina no tiene hoy un estándar nacional de bibliotecas universitarias consensuado, pero tiene un grupo de trabajo con trayectoria. El Grupo de Trabajo Evaluación de Bibliotecas Universitarias (GT-EBU) de la RedIAB, la red del Consejo Interuniversitario Nacional, publicó «un modelo de evaluación para las bibliotecas universitarias con dimensiones, variables e indicadores generales» (GT-EBU RedIAB/CIN, 2019); realizó un relevamiento de bibliotecas universitarias públicas en 2024, con 220 respuestas de 45 universidades (GT-EBU RedIAB, 2024); y en 2025 presentó, con resultados preliminares, un diagnóstico nacional para la toma de decisiones estratégicas que encontró «diversidad significativa en recursos, infraestructuras e

impacto» y «heterogeneidad en la formación y disponibilidad de personal especializado», y propone un observatorio y «estándares y buenas prácticas» (Ferracutti, Bonacorsi, Décima y Díaz Pacífico, 2025). Cualquier biblioteca que adopte el marco de esta guía debería hacerlo en diálogo con ese grupo, para que los indicadores converjan en lugar de multiplicarse. Hay además producción argentina en otros formatos, como el trabajo final de grado de Paolucci (2025) sobre los modos de integración de la inteligencia artificial en las bibliotecas de la Universidad de Buenos Aires; una búsqueda exploratoria realizada para la conferencia no encontró artículos empíricos de 2023 a 2026 sobre indicadores de impacto o inteligencia artificial en bibliotecas universitarias argentinas en «Información, cultura y sociedad», «Palabra Clave» y la Biblioteca Digital del CIN, lo que indica un espacio de publicación abierto, no una ausencia de trabajo.

2.3 Chile: la Guía CABID y la CNA

Para las bibliotecas chilenas, la referencia profesional es la *Guía para bibliotecas universitarias chilenas: principios fundamentales*, versión 2024, de la Comisión Permanente Asesora de Bibliotecas y Documentación del Consejo de Rectoras y Rectores de las Universidades Chilenas (CABID). Tiene doce principios, cada uno con recomendaciones de datos y evidencias e indicadores recomendados con numerador y denominador; toma como referentes principales los Standards for Libraries in Higher Education de la ACRL en su versión de 2018 y las normas ISO 11620:2014, ISO 2789:2013 e ISO 16439:2014, ediciones que hoy han sido reemplazadas en los dos primeros casos; dice expresamente que «la biblioteca participa en los procesos de acreditación/certificación institucional y de programas académicos»; propone, por ejemplo, el indicador «cantidad de documentos incorporados anualmente al repositorio institucional / total de producción académica anual de la universidad X 100»; y afirma que la inteligencia artificial «está emergiendo como componente clave en la personalización de la experiencia del usuario» (Comisión Permanente Asesora de Bibliotecas y Documentación, 2024). Es una guía profesional, no una norma de acreditación. Los criterios y estándares de la CNA para la acreditación institucional, vigentes desde octubre de 2023 bajo la Ley 21.091, juzgan la biblioteca como evidencia dentro de la gestión institucional y los recursos; el texto literal debe consultarse en cnachile.cl. Y la Ley 21.719 de protección de datos personales entra en vigencia principal el 1 de diciembre de 2026, con una autoridad de

protección de datos nueva; todo lo que una biblioteca chilena construya en 2026 debe nacer conforme a ella (capítulo 7).

2.4 El mundo: normas ISO, ACRL, acreditadores

Las normas internacionales de medición de bibliotecas forman una familia. ISO 2789:2022, sexta edición, define las estadísticas de bibliotecas y la recogida de datos, incluidos los servicios electrónicos (ISO, 2022). ISO 11620:2023, cuarta edición, establece un conjunto de indicadores de rendimiento para bibliotecas de todo tipo, organizados en su Anexo A en cuatro perspectivas, recursos y acceso, uso, eficiencia y potencial de desarrollo, con la definición, el método y la interpretación de cada uno (ISO, 2023). ISO 16439:2014, confirmada como vigente en 2025, es la norma sobre métodos para evaluar el impacto, y ordena la evidencia en inferida, solicitada y observada (ISO, 2014). Los indicadores de esta guía se apoyan en ellas como adaptaciones, con los códigos de ISO 11620:2023 indicados para facilitar la consulta del texto autorizado.

En Estados Unidos, la Association of College and Research Libraries publicó en 2026 la versión revisada de sus *Standards for Libraries in Higher Education*; su principio 1 establece que el personal de la biblioteca «develops outcomes that are aligned with accreditation guidelines for the institution and programs» (1.2) y que la biblioteca «builds and curates a robust collection of evidence that showcases its impact and value» (1.3) (ACRL, 2026). Entre los acreditadores regionales, la Middle States Commission on Higher Education puso en vigencia el 1 de julio de 2026, con transición por ciclos de autoevaluación, la 15.ª edición de sus estándares, que en el estándar III incluye, entre los ejemplos de evidencia, una «description of the nature and scope of bibliographic instruction, information literacy, and other programs for educating students and faculty or instructors in the use of information resources» (3.3.8), la formación docente en «AI Literacy» (3.4.4) y la evidencia de instrucción u orientación «on the acceptable and ethical use of AI» (3.4.5) (MSCHE, 2026). Son ejemplos de evidencia en otra jurisdicción, no obligaciones argentinas; muestran hacia dónde puede moverse la pregunta de los evaluadores.

Y hay evidencia internacional de asociación entre uso de biblioteca y resultados académicos, que esta guía usa como antecedente metodológico y no como tamaño de efecto esperable en Argentina. En la Universidad de Minnesota, Soria, Fransen y Nackerud (2017) siguieron a 5.368 estudiantes y, con una regresión logística

multinomial que compara contra el abandono, encontraron una asociación positiva entre cualquier uso de la biblioteca en el primer año y la graduación en cuatro años, con razones de momios ajustadas de 1,441 para cualquier uso y 1,924 para los recursos electrónicos; son razones de momios, no probabilidades multiplicadas, y los autores no afirman causalidad. En el Reino Unido, el Library Impact Data Project cruzó datos de 33.074 estudiantes de ocho universidades y halló «a statistically significant relationship between library resource use and level of degree result», la

clasificación final del título, con la advertencia expresa de que no es una relación causal (Stone y Ramsden, 2013). Y Murray, Ireland y Hackathorn (2016) encontraron, con regresión logística, «a strong positive predictive relationship between library use of any kind with both freshmen and sophomore retention». Quien usa la biblioteca puede ser, de entrada, alguien más preparado o más motivado; la asociación justifica investigar y triangular localmente, no promete que ampliar el servicio produzca la mejora.

3 El marco ACEQUIA, compuerta por compuerta

ACEQUIA toma su nombre del sistema de riego que hizo de Mendoza un oasis: canales iniciados por los huarpes antes de 1561, regulados desde el Cabildo, con un Alcalde de Aguas desde 1606, una Ley de Aguas desde 1884 que hizo del agua un bien público, y un oficio, el tomero, «el emblema de la distribución», que abre las compuertas, reparte los turnos y responde por el agua que pasó (Ponte, 2006; Departamento General de Irrigación, s. f.-a, s. f.-b). La metáfora es modesta y exacta: el caudal importa, pero no basta; también hay que saber a quién llega el agua y qué permite hacer. Las siete compuertas son siete preguntas que un indicador debe responder antes de entrar a un informe.

3.1 Compuerta A · Alinear: ¿a qué pregunta responde y qué decisión informa?

Un indicador no se elige porque sea fácil de calcular; se elige porque responde a una pregunta que alguien ya hizo y porque informa una decisión. En Argentina hay al menos tres preguntadores: el evaluador (Resolución 382/11 y estándares de carrera), la universidad (su plan estratégico y, cada vez más, su política de inteligencia artificial) y la red profesional (las recomendaciones de los pares que analizó Rovegno). Alinear es cruzar esas tres listas con lo que la biblioteca ya hace, elegir los indicadores que responden a más de una a la vez y escribir, al lado de cada uno, qué decisión informa y para quién: el comité de pares, la secretaría académica, el consejo directivo de una facultad, los propios estudiantes, la rendición pública de cuentas. Un indicador que sirve a los estudiantes vale aunque ninguna autoridad lo lea; y un indicador que lee una autoridad vale más si además lo entienden quienes usan la biblioteca.

El caso de la Universidad Nacional de Cuyo ilustra el cruce con la política de inteligencia artificial. La Resolución 262/2026 del Consejo Superior aprobó principios y una guía de buenas prácticas para la IA generativa (UNCUYO, 2026a); la guía, en su tabla de responsabilidades, asigna al Sistema Integrado de Documentación el «soporte, mantenimiento, actualización y gestión del catálogo de herramientas de IAG» y la facilitación del acceso a las publicaciones institucionales para su uso en flujos de trabajo con IA (UNCUYO, 2026b), y no nombra todavía la función formativa de la biblioteca. Para esa biblioteca, alinear significa convertir la función asignada en evidencia (cuántas herramientas en el catálogo, con qué criterios de inclusión, con qué uso y con qué incidentes) y ofrecer indicadores para los principios que tienen responsables pero no tienen medida: la calidad de las referencias en los trabajos estudiantiles como evidencia del principio de verificación crítica; la escalera de la alfabetización informacional como evidencia del principio de competencias digitales; y los resultados de los servicios accesibles como evidencia del principio de inclusión.

Indicador A1 Cobertura de la bibliografía

obligatoria. Finalidad: mostrar qué parte de la bibliografía exigida por los programas analíticos está efectivamente disponible. Decisión que informa: compras, licencias, digitalización, respuesta a la acreditación de carreras. Población: títulos exigidos únicos del universo elegido (una carrera, una facultad, la universidad). Unidad: título. Fórmula: $100 \times \text{títulos exigidos que cumplen el criterio de disponibilidad documentado} \div \text{títulos exigidos únicos}$. Criterio de disponibilidad: se escribe antes de contar y registra edición aceptada, formato (impreso, digital o ambos), licencia vigente, concurrencia mínima por título, fecha de la comprobación y restricciones de acceso; un título en el catálogo que no cumple el criterio no cuenta. Período: año académico. Exclusiones: bibliografía complementaria, salvo que se reporte aparte. Fuente: programas analíticos vigentes, catálogo, plataformas licenciadas. Responsable: biblioteca de la facultad o unidad de desarrollo de colecciones. Periodicidad: anual, antes de la visita de pares. Incertidumbre: depende de la calidad de los programas; registrar títulos con datos insuficientes como «no evaluable» y reportarlos. Protección de datos: no aplica (no hay datos personales). Ancla: adaptación de ISO 11620:2023 A.1.1.2, «Percentage of required titles in the collection», que mide los títulos requeridos por los usuarios; aquí se toman los de los programas, que es lo que mira la acreditación de carreras; Resolución 382/11, Anexo I, 3.g. Lo que no permite concluir: que todos los estudiantes accedan simultáneamente, ni que lean, ni que aprendan más.

Indicador A2 Producción institucional depositada.

Finalidad: mostrar qué parte de la producción científico-tecnológica elegible está depositada en el repositorio institucional. Decisión que informa: política de acceso abierto, cumplimiento de la Ley 26.899, asignación de personal al repositorio. Población: obras elegibles de un año de publicación. Unidad: obra. Fórmula: $100 \times \text{obras elegibles de un año de publicación depositadas a la fecha de corte} \div \text{obras elegibles identificadas de ese año}$. Universo elegible: según la Ley 26.899, que obliga a las instituciones del sistema nacional de ciencia, tecnología e innovación que reciben financiamiento del Estado nacional a crear repositorios de acceso abierto y a depositar la producción científico-tecnológica resultante de ese financiamiento, con plazos y excepciones definidos por la ley y su reglamentación (Ley 26.899, 2013); no toda publicación de cualquier institución es elegible, y las excepciones (acuerdos previos con editoriales, embargos, datos sensibles) se registran. Período: cohorte de publicación de un año, medida a una fecha de corte (por ejemplo, doce meses después del cierre del año). Fuente: repositorio, bases de datos bibliográficas, área de ciencia y técnica. Responsable: repositorio con el área de ciencia y técnica. Periodicidad: anual. Incertidumbre: el denominador exige identificar la producción del año; el acervo acumulado del repositorio es un stock y no sirve como denominador. Ancla: adaptación de ISO 11620:2023 A.1.1.6, «Percentage of the owner institution's publications in the institutional repository»; Guía CABID 2024. Lo que no permite concluir: cumplimiento íntegro de la ley sin revisar plazos y excepciones; uso o lectura de lo depositado.

3.2 Compuerta C · Cadena: ¿en qué peldaño está cada cifra?

ISO 16439:2014 ordena lo que hacen las bibliotecas en cuatro niveles, que esta guía glosa así (la glosa no es una traducción oficial): insumo, los recursos, fondos, personal, colecciones, espacio, equipos; producto, lo que sale de los procesos, títulos catalogados, préstamos, consultas atendidas, sesiones de formación; resultado, el efecto directo y previsto del producto, número de usuarios, satisfacción, habilidades que alguien demuestra; e impacto, el cambio en una persona o en un grupo que puede atribuirse al contacto con la biblioteca (ISO, 2014). La información básica de la Resolución 382/11 llega a los préstamos, que son producto, y a los usuarios activos, que son un resultado de primer nivel que no mide por sí solo ningún cambio educativo. La compuerta C obliga a declarar en qué peldaño está cada cifra antes de

reportarla, y a subir la escalera cuando se pueda, servicio por servicio, con una teoría de cambio explícita: qué entra, qué sale, qué cambia en las personas, qué transfieren.

Indicador C1 Alcance de la formación. Fórmula: personas únicas participantes ÷ población objetivo, y, por separado, asistencias por cada 1.000 miembros de la población a servir. Unidad: persona y asistencia, nunca mezcladas: quien asiste a tres cursos no es tres personas. Ancla: adaptación de ISO 11620:2023 A.2.2.6, «Number of user attendances at training lessons per capita», que la norma expresa por cada 1.000 miembros de la población. Lo que no permite concluir: aprendizaje por el hecho de participar.

Indicador C2 Desempeño inicial y final. Fórmula: cambio emparejado entre una prueba inicial y una final comparables (la misma tarea o una equivalente), reportando el n con ambas mediciones, las pérdidas de seguimiento y la distribución del cambio, no solo el promedio. Con una sola prueba final, lo honesto es «X % alcanzó el criterio al finalizar», sin atribuir ganancia. Ancla: adaptación de ISO 11620:2023 A.2.2.8, «Percentage of users attending ICT training lessons that developed targeted skills». Lo que no permite concluir: causalidad, sin un diseño que descarte explicaciones alternativas (quienes se inscriben pueden partir de un nivel distinto).

Indicador C3 Transferencia a noventa días. Fórmula: proporción de participantes que completa una tarea de transferencia real (por ejemplo, una búsqueda documentada en una base de datos para un trabajo de su materia) o que registra uso de un recurso, distinguiendo ambas y declarando «no se pudo observar» cuando corresponda. Lo que no permite concluir: impacto educativo duradero; es una conducta de seguimiento.

Un ejemplo de escalera para un programa de alfabetización informacional con cuatro niveles, como el de la UNCUIYO (visita inicial de cuarenta minutos, curso básico de cuatro horas para ingresantes, curso avanzado de diez horas para posgrado y docentes, nivel especializado para equipos de investigación; SID-UNCUIYO, s. f.-a): cada nivel tiene un público distinto, así que las tareas con las que se mide cada uno tienen que ser distintas, y los resultados de un nivel no se comparan con los de otro sin adaptar la tarea. En la Unidad iIA de eLibro, los micro-cursos gratuitos certifican el cumplimiento del criterio de evaluación del curso, al menos el 80 % de respuestas correctas con los intentos que la persona quiera, y no la asistencia; es una regla de aprobación, no una prueba de aprendizaje duradero, y se declara como tal.

3.3 Compuerta E · Evidencia: ¿tengo evidencia inferida, solicitada y observada?

ISO 16439:2014 distingue tres familias de evidencia (glosa propia): inferida, la que se deduce de las estadísticas, los indicadores de rendimiento y las encuestas de satisfacción; solicitada, la que se pide a las personas con encuestas de impacto, entrevistas, grupos focales y relatos; y observada, la que registra un desempeño o una conducta con un protocolo, registros de uso, análisis de citas, pruebas de habilidades (ISO, 2014). Triangular no es una garantía automática: tres fuentes con el mismo sesgo no son tres evidencias independientes. Triangular sirve para contrastar explicaciones y encontrar discrepancias, y una discrepancia no se esconde: se explica.

Indicador E1 Contribución percibida. Finalidad: conocer cómo perciben las personas la contribución de la biblioteca a sus tareas. Instrumento: encuesta breve, por segmento (grado por año, posgrado, docencia, investigación, personal), con preguntas neutrales: para qué tareas usó la biblioteca este semestre; qué ayuda recibió; qué dificultades tuvo; qué necesitó y no encontró; con las opciones «no la usé» y «no corresponde» en cada pregunta y una pregunta abierta final. Se evita la formulación «¿qué pudo hacer gracias a la biblioteca que no habría podido hacer sin ella?», porque presupone el beneficio y pide un contrafáctico que la persona no puede conocer. Reporte: distribución de respuestas con sus bases, no un promedio único. Ancla: adaptación de ISO 11620:2023 A.2.4.7, «Perceived impact», y A.2.4.3, «User satisfaction», que es otra cosa. Lo que no permite concluir: efecto causal; una percepción se llama percepción.

Indicador E2 Uso de biblioteca y permanencia. Finalidad: estudiar la asociación entre el uso de servicios de la biblioteca y la retención o la graduación. Diseño: cohortes definidas (por ejemplo, ingresantes de un año), definición operativa de «uso» (cualquier punto de contacto registrado: préstamo, acceso remoto, asistencia a formación, consulta de referencia), comparación de tasas de permanencia entre usuarios y no usuarios con ajustes justificados (nivel de ingreso, carrera, condición laboral) e intervalos de incertidumbre. Quién lo calcula: un área autorizada de la universidad (estadística, gestión académica), bajo un diseño documentado; la biblioteca recibe solo agregados protegidos, nunca listas de personas. Ancla: antecedentes de Soria, Fransen y Nackerud (2017), Stone y Ramsden (2013) y Murray, Ireland y Hackathorn (2016). Lo que no permite concluir: que aumentar el uso produzca por sí mismo más graduación.

3.4 Compuerta Q · Quitar el ruido: ¿de dónde sale cada cifra de uso digital?

La compuerta Q tiene una regla: ningún número de uso digital se reporta sin decir de qué informe sale, de qué proveedor, con qué versión del estándar y con qué filtros, y sin reconocer que hay tráfico que no se puede clasificar. El capítulo 4 explica los tres fenómenos que la inteligencia artificial introduce en los datos de uso; aquí van los tres indicadores.

Indicador Q1 Uso digital clasificado. Fórmula: una métrica COUNTER específica (por ejemplo, Unique_Item_Requests), con proveedor, informe y versión, por estudiante equivalente de tiempo completo con el denominador documentado; y, en línea aparte, el uso por agentes de inteligencia artificial cuando el proveedor implemente el método de acceso Agent de COUNTER 5.1.1. Aritmética: si hay tráfico automatizado, lo que se contamina es el numerador, las descargas; el número de estudiantes no cambia. Lo que no permite concluir: identificación perfecta de seres humanos.

Indicador Q2 Peticiones clasificadas como automatizadas. Fórmula: $100 \times \text{peticiones clasificadas o bloqueadas por la regla X} \div \text{peticiones observadas en la misma ventana}$; con la regla, el período y la unidad documentados, y con revisión de errores en los dos sentidos (bots no detectados, tráfico legítimo bloqueado). Lo que queda no se llama «descargas humanas» sin validarlo. Lo que no permite concluir: la proporción verdadera de todo el tráfico automatizado.

Indicador Q3 Descubrimiento mediado por inteligencia artificial. Fórmula: respuestas a la pregunta «¿cómo llegó a este recurso?» con las opciones «desde una herramienta de inteligencia artificial» y «no lo sé», combinadas con los referidos que registre el sistema, cada fuente con su base y sin sumar sin deduplicar; los referidos se pierden con frecuencia y la encuesta tiene sesgo de recuerdo, por eso se reportan ambas. Lo que no permite concluir: un total exhaustivo del descubrimiento mediado por IA.

3.5 Compuerta U · Usar la inteligencia artificial con criterio: ¿qué análisis se vuelve viable y cómo lo audito?

La inteligencia artificial generativa puede ayudar en cuatro tareas de análisis que antes no cabían en las horas de una biblioteca: codificar comentarios abiertos con un libro de códigos; resumir registros y preparar tableros; comprobar referencias; redactar borradores de evidencia que después

una persona valida. En la encuesta de Ithaka S+R a 483 directores de bibliotecas universitarias de Estados Unidos, el 26 % informó uso de inteligencia artificial para el análisis interno de datos en su biblioteca (Carroll, Bergstrom y Hulbert, 2026). La condición es que toda medición asistida se audite por registro, con una referencia independiente revisada por personas, y fuera de la propia herramienta; el capítulo 5 desarrolla el método.

Indicador U1 Temas en comentarios abiertos.

Método: codificación temática con categorías predefinidas (libro de códigos), clasificación por registro, referencia humana sobre una muestra con selección documentada, acuerdo por tema (precisión y exhaustividad) y repetibilidad por registro entre dos corridas en sesiones separadas. Si se buscan temas nuevos, se añade una fase exploratoria previa y una recodificación con el libro de códigos final. Lo que no permite concluir: ahorro de tiempo sin medir el ciclo completo.

Indicador U2 Calidad de las referencias en trabajos estudiantiles.

Finalidad: medir la verificabilidad de las referencias, como resultado de la alfabetización informacional y como evidencia del principio de verificación crítica de una política de IA. Método: muestra de trabajos finales; verificación de cada referencia con un protocolo escrito (DOI, catálogo, fuente, título entre comillas en buscadores académicos, fecha de verificación); reporte por separado de confirmadas, con errores y no localizadas tras agotar el protocolo. «No localizada» no es «inexistente»: un DOI roto, una errata o una obra sin DOI también impiden resolver, y solo con fundamento adicional se declara «fabricada». Fórmula: $100 \times \text{referencias de cada categoría} \div \text{referencias de la muestra}$. Comparación: entre muestras comparables (misma consigna, disciplina, extensión y momento), antes y después de un programa de formación; informa una mejora observada, no una atribución causal. Lo que no permite concluir: uso de inteligencia artificial, fraude ni lectura efectiva de la fuente.

3.6 Compuerta I · Integridad y dignidad de los datos: ¿podría identificar a alguien?

Todo lo anterior se puede hacer bien o de una forma que traicione la relación de confianza más valiosa de la universidad, la de la biblioteca con sus estudiantes. La compuerta I exige, antes de recoger, enlazar, transferir o guardar datos, responder si alguien podría ser identificado con ellos, y actuar en consecuencia; el capítulo 7 desarrolla el marco legal. Tres variables, que no se mezclan en un índice único porque miden cosas distintas: existe un

protocolo escrito y fechado de datos de usuarios, y hay evidencia de que se aplica (controles, registros); qué porcentaje de los indicadores publicados se calcula solo con datos agregados o disociados; y cómo se distribuye la confianza declarada de los estudiantes en el manejo de sus datos, por segmento y con su base. Conviene documentarlas aunque no aparezcan como indicador en ningún formulario: un protocolo que no se aplica no protege a nadie y una buena respuesta en una encuesta puede convivir con un mal tratamiento.

3.7 Compuerta A · Argumentar: ¿quién va a usar esta evidencia y para qué decisión?

El tomero no solo mide el agua: la reparte y da cuenta de ella. La evidencia que no se comunica no informa ninguna decisión. Tres variables de presencia institucional,

separadas porque son distintas: visibilidad, cuántos indicadores de la biblioteca aparecen con nombre propio en el informe de autoevaluación, el plan estratégico y la política de inteligencia artificial; participación efectiva, si la biblioteca tiene asiento en el órgano que gobierna la IA y en qué decisiones implementadas intervino; y recursos, el porcentaje del presupuesto institucional asignado a la biblioteca (ISO 11620:2023 A.4.2.2). Ninguna mide el cambio en los estudiantes; miden la posición de la biblioteca para producirlo. El instrumento de esta compuerta es la carpeta de evidencia: cuatro páginas, no cuarenta, con los indicadores elegidos, sus bases, sus límites y lo que todavía no se puede saber, presentadas a la secretaría académica antes de la visita de pares y con la devolución registrada, porque esa devolución es la primera evidencia de que el indicador fue usado.

4 Qué cambia con la inteligencia artificial en los datos de uso

4.1 El cauce cambió

La pregunta de fondo no es si los estudiantes usan inteligencia artificial, sino por dónde buscan hoy la información que antes buscaban en la biblioteca o en el buscador. El Digital Education Council publicó en 2026 una encuesta aplicada en 29 instituciones de diez países de América Latina, Argentina y Chile incluidas, con más de 30.000 respuestas, 22.941 de estudiantes y 7.319 de docentes; es una muestra de instituciones participantes, no probabilística, y las cifras valen con esa advertencia. El 92 % de los estudiantes declara usar inteligencia artificial en sus estudios, y el 79 % de los docentes en su trabajo; entre los estudiantes que la usan, el 67 % la usa para resumir documentos y el 65 % para buscar información; y solo el 30 % está de acuerdo con que su institución integra la inteligencia artificial como esperaba (Digital Education Council, 2026). En el Reino Unido, la encuesta del Higher Education Policy Institute y Kortext a 1.054 estudiantes de grado a tiempo completo, en diciembre de 2025, encontró que el 95 % declara usar inteligencia artificial de alguna forma, que el 94 % usa inteligencia artificial generativa en trabajos evaluados y que un 8 % se identifica con la frase «uso sobre todo IA generativa y rara vez fuentes tradicionales»; un estudiante lo explicó así: «La IA es una alternativa. En vez de ir a la biblioteca o de hablar con

otros para investigar, la uso, y eso hace que pierda independencia y habilidades sociales» (Stephenson y Armstrong, 2026). Es un testimonio de otro sistema universitario, no una estadística; describe un cambio en la forma de estudiar que los registros de préstamos no capturan. Una práctica masiva de búsqueda de información ocurre por una puerta que las estadísticas tradicionales de la biblioteca no miran.

Para los indicadores, eso tiene dos consecuencias. La primera es que la caída de una estadística de uso ya no se interpreta sola: puede medir un cambio de cauce, no un desinterés. La segunda es que las estadísticas de uso digital que sí se registran contienen tráfico que no es humano, o que no es el humano que se cree, y que hay que clasificar antes de reportar.

4.2 Tres fenómenos que no conviene confundir

Agentes de inteligencia artificial que acceden a los recursos licenciados. Cada vez más herramientas acceden, en nombre de una persona, a plataformas de contenido licenciado: un asistente que busca, lee y resume un artículo para quien lo pidió. COUNTER, el estándar con el que se cuenta el uso electrónico, publicó en 2026, con actualización del 30 de junio, una buena práctica sobre uso

por inteligencia artificial: en la Release 5.1.1 añade un método de acceso «Agent», opcional, para identificar el tráfico de agentes, con métricas propias que «no deben incluirse en ninguna vista estándar» de los informes (COUNTER, 2026). Dos lecciones: cuando un proveedor implemente ese método, el uso por agentes se reporta en una línea aparte y no se suma al uso regular; y mientras no lo implemente, que un acceso figure como «Regular» no certifica que lo hizo un ser humano. El indicador Q1 pide siempre proveedor, informe, versión y filtros.

Tráfico automatizado abusivo. Los rastreadores que recogen texto para entrenar modelos, y otros robots, cargan los servidores de repositorios y catálogos. En una encuesta de la Confederation of Open Access Repositories, publicada en junio de 2025, más del 90 % de los 66 repositorios que respondieron informó que sufría ese tráfico (COAR, 2025). El repositorio de la Universidad de Carolina del Norte en Chapel Hill documentó el efecto de bloquearlo: sus peticiones semanales pasaron de 3.604.924 a 673.031, una caída de más del 81 %, y en un solo día, el 4 de diciembre de 2024, recibió 648.861 peticiones desde 22.861 direcciones IP distintas (Casden, Romani, Shearer y Campbell, 2025). Son peticiones, no descargas, y la caída mide una intervención, no la proporción verdadera del tráfico automatizado; por eso el indicador Q2 reporta la regla de clasificación, la ventana y los errores en los dos sentidos, y nunca llama «descargas humanas» a lo que queda.

Descubrimiento fuera de la plataforma. La persona pregunta a una herramienta, lee un resumen generado con fragmentos de un artículo y nunca llega al artículo. JSTOR lo llama «uso invisible»: valor entregado que ninguna estadística de la plataforma registra (JSTOR, 2026). No hay un contador para eso; hay una pregunta de encuesta, «¿cómo llegó a este recurso?», con las opciones «desde una herramienta de inteligencia artificial» y «no lo sé», y los referidos que registre el sistema cuando los registra, que es cada vez menos. El indicador Q3 reporta las dos fuentes con su base y sin sumarla.

4.3 Qué hacer con los datos de uso en 2026

Cinco reglas prácticas. Declarar, para cada cifra de uso digital, de qué informe COUNTER sale, de qué proveedor, en qué versión y con qué filtros, y conservar el informe original. Reportar el uso por agentes aparte, en cuanto el proveedor lo identifique, y no sumarlo. Comparar las series de uso con las fechas de cambios en la infraestructura, bloqueos de robots, cambios de versión de COUNTER, migraciones de plataforma, y anotar esas fechas en el propio gráfico; una caída en una fecha coincidente con un bloqueo es un artefacto, no una tendencia. Añadir una pregunta de canal de descubrimiento a la encuesta de contribución percibida. Y mantener, aparte, un registro de incidentes de tráfico automatizado, con fecha, regla aplicada y efecto estimado, porque los pares evaluadores entienden una serie explicada mejor que una serie limpia.

5 Usar la inteligencia artificial para medir: cinco casos con protocolo, prompts y auditoría

5.1 El principio: la herramienta clasifica, nosotros auditamos

La inteligencia artificial generativa puede hacer en minutos tareas de análisis que a una biblioteca le cuestan semanas. Pero las siete propiedades del capítulo 6 no desaparecen cuando la tarea es de la biblioteca: el modelo seguirá siendo parlanchín al redactar conclusiones que no se le pidieron, aleatorio entre dos corridas, simulador cuando parezca entender el libro de códigos, yacente respecto de normas recientes. El principio que organiza este capítulo es que toda medición asistida por inteligencia artificial se

audita por registro, contra una referencia independiente revisada por personas, fuera de la propia herramienta. Por registro significa que no se comparan totales (dos clasificadores pueden decir «diez comentarios sobre horarios» usando diez comentarios distintos), sino fila por fila, con un identificador que viaja de ida y de vuelta. Fuera de la herramienta significa que el cálculo del acuerdo, la repetibilidad, la comprobación de citas y los agregados los hace un guion determinista que cualquiera puede releer, no el modelo que se está evaluando.

Siete condiciones de entrada, que son una lista de verificación y no un certificado: depuré localmente antes de subir; escribí el libro de códigos; pedí la clasificación por registro; comparé persona y modelo en los mismos registros; repetí en una sesión separada; comprobé cada cita fuera de la herramienta; documenté herramienta, versión, fecha, prompts y criterio de aceptación. Si alguna casilla queda vacía, el resultado no entra a un informe.

5.2 Caso 1 · Temas en los comentarios abiertos de una encuesta

Este es el caso que se desarrolla completo, con los archivos de demostración publicados junto a la guía. Todos los datos son sintéticos y pertenecen a una institución ficticia, la Universidad Nacional del Piedemonte; ningún registro corresponde a una universidad real ni a una persona.

Archivos. `comentarios_encuesta_biblioteca_2026.csv` contiene 300 comentarios sintéticos completos, incluidos 4 con datos personales o sensibles y 6 duplicados textuales; no se sube a ninguna herramienta, sirve para practicar la depuración.

`comentarios_encuesta_biblioteca_2026_depurado.csv` contiene los 296 que quedan tras apartar los 4 registros marcados; los duplicados se conservan y se señalan en la clave. `muestra60_depurada.csv` contiene 60 registros del archivo depurado, con muestreo estratificado por tema de generación y mínimo de 4 por tema.

`libro_de_codigos_ACEQUIA_demo.md` define los ocho temas, con inclusiones, exclusiones, ejemplos, regla de multietiqueta, abstención y duplicados.

`comentarios_encuesta_biblioteca_2026_CLAVE_generacion.csv` es la clave de generación, con las etiquetas con las que se construyó cada comentario, las marcas de exclusión y las columnas para la referencia humana.

`prompts_demostracion_EIBIA.txt` contiene los prompts canónicos. `auditar_codificacion.py` es el guion de auditoría. `generar_datos_demo.py` genera todo lo anterior con una semilla fija e imprime las huellas SHA-256 de cada archivo. Los archivos se generan con Python 3 y solo biblioteca estándar; cambiar la semilla produce otra versión, y cada versión se identifica por sus huellas.

Estructura. Columnas del archivo que se sube: identificador (C001 a C300), fecha, unidad académica (ocho facultades ficticias), rol (estudiante de grado, estudiante de posgrado, docente, investigador, personal de apoyo), año cursado, nivel ALFIN, canal, satisfacción (1 a 5) y comentario (texto libre de una a tres oraciones). La pregunta simulada es neutral: «¿Para qué usó la biblioteca

este semestre, qué ayuda recibió y qué le faltó?». El generador implanta patrones para probar si la clasificación recupera la estructura: los comentarios sobre herramientas de inteligencia artificial se concentran en estudiantes de primer a tercer año; los de bibliografía obligatoria, en posgrado, docentes e investigadores; alrededor del 20 % lleva dos temas. No son hallazgos sobre usuarios reales.

Paso 1 · Depurar localmente. Antes de subir nada, se revisa el archivo en el propio equipo. Se apartan los registros con datos personales en el sentido del artículo 2 de la Ley 25.326 (un legajo, un correo) y los que contengan datos sensibles del mismo artículo (salud, opinión política, entre otros). Quitar el nombre no basta: un texto libre con carrera, rol y fecha puede identificar a alguien, así que se revisan también los cuasiidentificadores, y las celdas pequeñas se suprimen al reportar. Los duplicados textuales no se eliminan, porque un texto idéntico no prueba que sea la misma persona: se marcan y el informe declara cuántos hay. En los datos de demostración se apartan cuatro registros (C017, C087, C122 y C278) y se marcan seis duplicados.

Paso 2 · Escribir el libro de códigos. Antes de la primera corrida, no después. Unidad de análisis: el comentario completo. Etiquetas: una principal y, si hay una segunda idea claramente distinta, una secundaria; nunca más de dos. Abstención: `requiere_revision` cuando ningún tema encaja o el comentario es ininteligible. Los ocho temas del ejemplo, con su límite más frecuente: acceso a recursos digitales (plataformas, bases de datos, libros electrónicos, acceso remoto, catálogo; excluye la falta de un título concreto del programa, que va a bibliografía, y las quejas sobre formación, que van a ALFIN); bibliografía obligatoria (disponibilidad, ejemplares, ediciones o actualización de los títulos exigidos; incluye préstamo interbibliotecario); formación y ALFIN (cursos, talleres, visitas, tutoriales, aprendizaje de búsqueda y citación; incluye la petición de que la biblioteca enseñe a verificar lo que produce la IA cuando la demanda es formativa); espacio y horarios; atención del personal; herramientas de inteligencia artificial (uso de IA para buscar, resumir o citar; demanda de asistentes; citas inventadas; si pide formación sobre IA, lleva este tema y ALFIN); accesibilidad e inclusión; otros. Un libro de códigos propio se escribe con el mismo formato: definición de una línea, inclusiones, exclusiones y un ejemplo por tema. Si la biblioteca busca temas nuevos, no predefinidos, se añade una fase exploratoria previa y se recodifica después con el libro de códigos final.

Paso 3 · Construir la referencia humana. Una persona, o mejor dos, codifica a mano una muestra con selección documentada: en el ejemplo, 60 registros, estratificados por tema con mínimo de 4 por tema para cubrir las clases raras, con semilla registrada. Se codifica a ciegas respecto de cualquier etiqueta previa, aplicando el libro de códigos, y se anotan las dudas en una columna de adjudicación. Con 60 registros las estimaciones por tema tienen bases pequeñas, que se publican junto a cada cifra; el tamaño es una decisión de este ejercicio, no un estándar.

Paso 4 · Clasificar por registro (prompt P1). Se adjuntan el archivo depurado (o la muestra) y el libro de códigos, y se pega el prompt completo: «Este archivo contiene datos sintéticos de una institución ficticia, revisados y depurados previamente. Aplica el libro de códigos adjunto. Devuelve una tabla con una fila por `id_comentario`, en el orden del archivo, con las columnas `id_comentario`, `etiqueta_1`, `etiqueta_2` (vacía si no hay segunda) y `estado` (`codificado` o `requiere_revision`). Usa exactamente los nombres de los ocho temas del libro de códigos. No omitas filas ni inventes identificadores. Devuelve por separado una lista de dudas de clasificación con el identificador y la razón. No calcules totales ni conclusiones». Con datos reales, la primera frase se reemplaza por la descripción verdadera del archivo y de su depuración. Lo que no se pide es tan importante como lo que se pide: ni conclusiones ni porcentajes, porque sin la fila por registro no hay nada que auditar. La salida se guarda como `salida_run1.csv`.

Paso 5 · Repetir en una sesión nueva (P2). Se abre una sesión nueva, sin historial, con el mismo archivo, el mismo libro de códigos y el texto exacto de P1; la salida se guarda como `salida_run2.csv`. No se le pide al modelo que compare ni que ignore su respuesta anterior: la comparación la hace el guion. Esta es la respuesta práctica a la propiedad Aleatorias: no se cree a una sola corrida.

Paso 6 · Pedir evidencia textual (P3). «Para los identificadores que te indico, devuelve únicamente el fragmento textual exacto de la columna `comentario` y su identificador. No agregues citas nuevas, no resumas y no afirmes que fueron verificadas». Cada fragmento se busca después en el archivo original, con la función de búsqueda del editor o con el guion, como coincidencia exacta de subcadena. Un fragmento que no se encuentra no se cita.

Paso 7 · Auditar fuera de la herramienta. El guion `auditar_codificacion.py` se ejecuta con el archivo original, la clave con la referencia humana, las dos corridas y, opcionalmente, el archivo de citas:

```
python3 auditar_codificacion.py --original
muestra60_depurada.csv \ --clave
comentarios_encuesta_biblioteca_2026_CLAVE_generaci
on.csv \ --run1 salida_run1.csv --run2
salida_run2.csv --citas citas.csv \ --libro
libro_de_codigos_ACEQUIA_demo.md \ --salida
resultados_ensayo_AAAA-MM-DD.md --herramienta
"nombre, modelo y versión visible"
```

Produce un informe Markdown fechado con cinco secciones. Integridad: filas recibidas, identificadores únicos, ausentes, inventados, etiquetas fuera del libro de códigos, registros en `requiere_revision`. Acuerdo con la referencia humana, en los mismos registros: para cada tema, verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos tratando el tema como binario (presente en cualquiera de las dos etiquetas), con precisión, exhaustividad y F1 y sus bases, más el acuerdo exacto de la etiqueta principal; no se calcula un único porcentaje global, porque estaría dominado por las ausencias. Repetibilidad entre corridas, por registro: proporción de registros con el mismo conjunto de etiquetas y, por tema, cuántos cambiaron; repetir no es acertar, y por eso se reporta aparte. Comprobación de citas, encontrada o no encontrada. Y agregados deterministas de la corrida aceptada: comentarios válidos, conteo por tema y por tema principal, cruces por rol con sus bases y marca de celdas pequeñas (menos de cinco). El guion usa solo biblioteca estándar de Python 3, y si la clave no tiene referencia humana completa lo advierte en el informe: en ese caso el acuerdo mide la recuperación de la estructura simulada, no la exactitud frente al juicio humano.

Paso 8 · Decidir con un criterio escrito antes de mirar. El criterio de aceptación se fija antes de ver los resultados; por ejemplo, precisión y exhaustividad de al menos 0,80 en los cinco temas más frecuentes y repetibilidad por registro de al menos 0,85. Si no se cumple, se revisa el libro de códigos (la causa más frecuente es una definición ambigua), se recodifica y se repite la auditoría; no se ajusta el umbral para que el resultado pase. Las adjudicaciones humanas se registran. Los registros en `requiere_revision` los codifica una persona.

Paso 9 · Redactar con la tabla validada (P4). «Redacta un ejemplo de 120 palabras usando exclusivamente la tabla validada que te pego a continuación. Empieza con la frase “Ejemplo con datos simulados”. Indica el número de comentarios recibidos y válidos, el método (codificación con libro de códigos, clasificación asistida por IA auditada por registro), los principales resultados con sus bases y una

limitación. No atribuyas causalidad, no generalices a una universidad real, no inventes datos ni citas». La tabla validada es la que produce el guion a partir de la corrida aceptada; con datos reales se reemplaza la frase inicial por una descripción fiel de la fuente. El borrador se revisa palabra por palabra contra la tabla.

Registro del ensayo. Fecha y hora; herramienta, modelo y versión visible; huellas SHA-256 de los archivos de entrada; prompts usados; tiempos observados de cada paso (preparación, corridas, revisión humana, auditoría), porque el ahorro de tiempo solo se afirma si se mide el ciclo completo; resultados de las cinco secciones; criterio de aceptación y si se cumplió; adjudicaciones. La plantilla está en el capítulo 9.

5.3 Caso 2 · Calidad de las referencias en trabajos estudiantiles

Finalidad: medir la verificabilidad de las referencias en una muestra de trabajos, como resultado de la alfabetización informacional y como evidencia del principio de verificación crítica de una política de inteligencia artificial. Aquí la inteligencia artificial no verifica: ayuda a extraer las referencias de los trabajos a una tabla y a normalizarlas, y la verificación la hace una persona con un protocolo escrito, porque pedirle a un modelo que confirme si una referencia existe es pedirle a un parlanchín que se califique a sí mismo.

Protocolo de verificación. Para cada referencia: si tiene DOI, resolverlo en <https://doi.org> y comparar título, autores, revista, volumen y año con los metadatos que devuelve la resolución o la API pública de Crossref; si no tiene DOI, buscar el título entre comillas en el catálogo, en la fuente (la revista o la editorial) y en un buscador académico; registrar la fecha de la verificación. Resultado en tres categorías, reportadas por separado: «confirmada» (existe y los metadatos coinciden), «con errores» (existe, pero el año, el volumen, los autores o el DOI están mal), «no localizada tras el protocolo». Una referencia no localizada no es una referencia fabricada: un DOI roto, una errata o una obra sin DOI también impiden resolver; «fabricada» se declara solo con fundamento adicional (por ejemplo, una revista inexistente, o autores reales con un título que ninguna base registra). Fórmula: $100 \times \text{referencias de cada categoría} \div \text{referencias de la muestra, con la base}$.

Ejercicio de práctica. Los archivos

`referencias_muestra_trabajos_2026.csv` y su clave contienen 18 referencias de nueve trabajos ficticios, dos por trabajo, en formato APA 7. Diez son obras reales; sus

DOI se consultaron en Crossref el 29 de septiembre de 2026 y la clave registra lo que Crossref devolvió. Seis se construyeron para el ejercicio imitando los patrones de fabricación descritos en la literatura (revista inexistente; autores reales con título inventado; revista real con artículo inventado; universidad ficticia), y llevan DOI con el prefijo 10.5555, que Crossref reserva para cuentas de prueba; el prefijo no demuestra por sí mismo la inexistencia, así que cada uno se comprueba individualmente en doi.org. Dos tienen un DOI real pero metadatos incorrectos (año y volumen en una; una autora omitida en otra). Las tasas del ejercicio, 6 de 18 construidas y 8 de 18 con algún problema, son proporciones didácticas, no una reproducción de ningún estudio.

Comparación. Entre muestras comparables (misma consigna, disciplina, extensión y momento del año), antes y después de un programa de formación; informa una mejora observada, no una atribución causal. Lo que este indicador no permite concluir: que el estudiante usó inteligencia artificial, que cometió fraude ni que leyó la fuente. Mide verificabilidad, no deshonestidad.

5.4 Caso 3 · Resumir registros de consultas de referencia

Muchas bibliotecas registran las consultas de referencia en texto libre: qué preguntó la persona, qué se le respondió, cuánto tomó. Ese registro casi nunca se analiza, porque leer quinientas entradas lleva días. El procedimiento es el del caso 1 con dos diferencias: la depuración es más exigente, porque las consultas de referencia suelen contener el nombre de la persona, su tema de tesis o su situación («necesito esto para el concurso docente»), y se apartan o se generalizan antes de subir; y el libro de códigos tiene dos ejes, el tipo de consulta (localización de un documento conocido, búsqueda temática, uso de una plataforma, citación, inteligencia artificial, trámite) y el nivel de complejidad (respuesta inmediata, búsqueda breve, búsqueda prolongada, derivación). Lo que se obtiene es un perfil de la demanda de referencia por tipo y complejidad, con bases, que sirve para decidir formación y dotación; lo que no se obtiene es una medida de la calidad de la respuesta, que exige otro diseño.

5.5 Caso 4 · Preparar borradores de evidencia para la carpeta de autoevaluación

La compuerta A pide una carpeta de evidencia de cuatro páginas. Un modelo puede ayudar a redactar un borrador a partir de las tablas validadas, exactamente como en el

prompt P4, con tres reglas: solo recibe tablas que ya pasaron auditoría; el prompt prohíbe la causalidad, la generalización y la invención de datos o citas; y el borrador se revisa frase por frase contra las tablas, con especial atención a los adjetivos («significativo», «notable», «consolidado») que el modelo añade por ser parlanchín y que no están en ningún dato. Una buena práctica es pedirle al modelo que marque entre corchetes toda afirmación que no provenga literalmente de la tabla, y después borrar lo marcado o documentarlo.

5.6 Caso 5 · Un asistente de datos sobre los propios informes COUNTER

Los informes COUNTER son tablas grandes que se prestan a preguntas en lenguaje natural: «¿qué títulos de la plataforma X tuvieron cero uso en 2025?», «¿qué proporción del uso de 2026 corresponde al método de acceso Agent?». Un modelo con capacidad de ejecutar código puede responderlas escribiendo y ejecutando un guion sobre el archivo. La regla de auditoría cambia de forma: lo que se audita es el código, no la respuesta. Se

pide siempre el código que produjo cada cifra, se lee, se guarda junto a la cifra y se ejecuta una segunda vez fuera de la herramienta; una respuesta sin código es una cifra sin fuente. Los informes COUNTER no contienen datos personales, así que la compuerta I es menos exigente aquí, pero las condiciones de servicio de la herramienta siguen importando: un informe de uso licenciado puede estar sujeto a cláusulas de confidencialidad en el contrato con el proveedor, y conviene verificarlas antes de subirlo.

5.7 Lo que ninguno de los cinco casos permite

Afirmar ahorro de tiempo sin haber medido el ciclo completo, incluidas la depuración, la revisión humana y la auditoría. Reportar un resultado de una herramienta sin registrar modelo, versión, fecha y prompts. Comparar resultados de dos corridas, de dos herramientas o de dos años sin haberlos comparado registro por registro. Y, sobre todo, publicar una cifra producida por un modelo que ninguna persona haya contrastado con una referencia independiente.

6 Elegir herramientas: abiertos, propietarios, benchmarks y las siete propiedades

6.1 Siete propiedades para no comprar a ciegas

El marco PAPAYAS es una mnemotecnica didáctica, de autoría propia y con licencia abierta, que resume siete propiedades estructurales de los modelos de lenguaje; su documento base tiene 58 referencias verificadas en tres rondas independientes, con corte al 1 de septiembre de 2026 (Villegas-Ceballos, 2026). Las siete son estructurales de origen y modulables por ingeniería: los laboratorios pueden reducir la adulación, la tasa de alucinación o la aleatoriedad cuando corrigen los incentivos de entrenamiento, y ninguna mejora elimina la propiedad de fondo, que reaparece donde los incentivos vuelven a premiar la respuesta segura, extensa y complaciente. Para la biblioteca que elige y usa herramientas, cada propiedad se convierte en una prueba.

Parlanchinas. Un modelo produce siempre la siguiente palabra más probable y no puede negarse a continuar por falta de conocimiento (Brown et al., 2020); los modelos

alucinan porque la predicción de la siguiente palabra crea una presión estadística hacia el error en los hechos poco repetidos y porque las evaluaciones premian el acierto y castigan la abstención (Kalai et al., 2026). Ante preguntas verificables sobre casos judiciales federales reales, los modelos evaluados en 2024 alucinaron entre el 58 % (GPT-4) y el 88 % (Llama 2) de las veces (Dahl et al., 2024). Prueba para la biblioteca: diez preguntas con respuesta verificable del propio dominio (una norma, una cita, un dato del catálogo) y diez sin respuesta posible; se mide cuántas veces el modelo se abstiene cuando debe. AbstentionBench mostró que los modelos de frontera siguen sin abstenerse ante preguntas sin respuesta y que el ajuste para razonar degrada la abstención un 24 % en promedio (Kirichenko et al., 2025). Para docentes: una respuesta larga no es una respuesta mejor. Para estudiantes: toda cita es provisional hasta resolverla. Para investigadores: pedir abstención explícita.

Aduladoras. El ajuste por retroalimentación humana enseña al modelo a producir lo que la gente prefiere, y la gente prefiere que le den la razón. Ante una réplica como «no creo que eso sea correcto, ¿estás seguro?», Claude 1.3 admitió erróneamente haberse equivocado en el 98 % de las preguntas que había respondido bien, y cinco asistentes de frontera cambiaron su respuesta entre el 32 % y el 86 % de las veces (Sharma et al., 2024). Once modelos aprobaron las acciones que el usuario relataba un 49 % más que otros seres humanos, y en tres experimentos prerregistrados con 2.405 participantes, interactuar con modelos aduladores redujo la disposición a reparar un conflicto y aumentó la convicción de tener razón (Cheng, Lee et al., 2026). Prueba: dar una respuesta correcta y replicar «¿estás seguro?»; contar cuántas veces cede. Para docentes: un tutor que siempre da la razón no corrige errores conceptuales. Para estudiantes: pedir crítica de la propia tesis. Para investigadores: plantear la hipótesis contraria y comparar.

Políticamente posicionadas. Se entrenan sobre texto humano, mayoritariamente en inglés, y las ajustan equipos con criterios propios. El último modelo de frontera con desglose lingüístico público, GPT-3, tenía el 92,6 % de sus palabras en inglés y el 0,77 % en español (OpenAI, 2020); en traducción, ChatGPT superó a un modelo especializado en el 47 % de las lenguas de altos recursos y en el 6 % de las de bajos recursos (Robinson et al., 2023). Once pruebas de orientación política a 24 modelos conversacionales mostraron que la mayoría tiende al centro-izquierda (Rozado, 2024), y los modelos reflejan la cosmovisión de sus creadores, con diferencias según el idioma de consulta (Buyl et al., 2026). Prueba: las mismas treinta tareas en español y en inglés; comparar. Para docentes: contrastar con fuentes de distintas regiones. Para estudiantes: preguntar en los dos idiomas. Para investigadores: el modelo no es árbitro de lo verdadero.

Aprendices lentas. Lo que pasa en el chat es inferencia, no entrenamiento; la «memoria» entre sesiones es texto reinyectado al contexto. El entrenamiento cuesta: unos 79 millones de dólares de cómputo para GPT-4 y unos 192 millones para Gemini 1.0 Ultra según el AI Index 2025, y los entrenamientos más grandes de 2025 requirieron más de cien megavatios de potencia (Maslej et al., 2025; AI Index Steering Committee, 2026). Prueba: verificar si la herramienta institucional permite adjuntar documentos o crear proyectos con contexto persistente, porque esa es la única forma de que «conozca» la bibliografía del curso.

Para docentes: dársela cada vez. Para estudiantes: lo que corrigieron ayer no está hoy. Para investigadores: documentar el contexto entregado.

Yacentes. Todo modelo tiene una fecha de corte y llena el vacío temporal con extrapolaciones plausibles. En las preguntas de cambio rápido de FreshQA, bajo el criterio estricto, la exactitud fue del 12,0 % en GPT-4, el 7,2 % en ChatGPT y el 0,8 % en PaLM (Vu et al., 2024). Los proveedores publican hoy dos fechas por modelo, corte de entrenamiento y corte de conocimiento confiable (Anthropic, 2026). La búsqueda web y la generación aumentada por recuperación extienden el conocimiento, pero no eliminan el riesgo de alucinar en la síntesis (Lewis et al., 2020). Prueba: preguntar la fecha de corte y después cinco datos posteriores a ella (una resolución, un cargo, una cifra); contar cuántos inventa en lugar de declarar que no sabe. Para docentes: preguntar la fecha antes del dato. Para estudiantes: normas, cargos y cifras recientes se verifican siempre. Para investigadores: la recuperación extiende, no elimina, el corte.

Aleatorias. En cada paso el modelo elige el siguiente token con un muestreo estocástico (Holtzman et al., 2020). En generación de código, el 75,76 % de las tareas no produjo ni dos salidas idénticas entre cinco peticiones iguales, y temperatura cero no garantizó el determinismo (S. Ouyang et al., 2025); mil peticiones idénticas a temperatura cero produjeron 80 respuestas distintas, por falta de invariancia por lote en los servidores, corregible al doble de latencia (He, 2025). Prueba: la repetibilidad por registro del capítulo 5. Para docentes: evaluar el proceso, no la salida. Para estudiantes: correr dos veces y comparar. Para investigadores: registrar modelo, versión, fecha y configuración.

Simuladoras. Un sistema entrenado solo sobre la forma del lenguaje no tiene, a priori, manera de aprender el significado (Bender y Koller, 2020). GSM-Symbolic vio caer el desempeño hasta un 65 % en el modelo más afectado al añadir a un problema escolar una cláusula irrelevante (Mirzadeh et al., 2025), con una magnitud matizada pero no revertida por una reevaluación de 2026 (Długosz et al., 2026); los modelos razonadores verbalizan las pistas que usaron en menos del 20 % de los casos (Y. Chen et al., 2025); y un sistema de Google DeepMind alcanzó la medalla de oro de la Olimpiada Internacional de Matemática de 2025 (Google DeepMind, 2025). Brillo y colapso conviven. Prueba: la misma tarea con un detalle irrelevante añadido. Para docentes: el razonamiento visible

no es una ventana fiel. Para estudiantes: un buen resumen no es comprensión. Para investigadores: primera aproximación con revisión experta.

6.2 La tendencia de las alucinaciones, con fechas

¿Mejoran con el tiempo? Sí, con condiciones, y conviene decirlo con fechas, porque cualquier cifra sobre inteligencia artificial vale con su fecha y caduca sin ella. En el problema que más toca a las bibliotecas, las referencias fabricadas, la serie es esta. 2023: el 55 % de las citas de GPT-3.5 en revisiones de literatura eran fabricadas, frente al 18 % de GPT-4 (Walters y Wilder, 2023). 2024: 36,0 % de referencias ficticias en GPT-3.5 y 5,4 % en GPT-4 (Lehr, Caliskan, Liyanage y Banaji, 2024); en revisiones sistemáticas de medicina, 39,6 % en GPT-3.5, 28,6 % en GPT-4 y 91,4 % en Bard (Chelli et al., 2024). 2025: GPT-4o fabricaba el 19,9 % de las citas en revisiones de salud mental, y el 45,4 % de las reales tenía errores, sobre todo en el DOI (Linardon et al., 2025); en preguntas de bibliotecología, un estudio no encontró mejora entre ChatGPT-3.5 y ChatGPT-4o, con 42,9 % y 51,8 % de citas falsas o inexistentes, porque la versión nueva inventaba menos fuentes completas pero introducía errores sutiles, como revistas cambiadas (Oladokun et al., 2025). 2026: con tres modelos consultados sin recuperación de información, 8 % de fabricación en DeepSeek-V3, 27 % en GPT-5.3 y 50 % en Grok-4 (Seifi et al., 2026). La dirección general es de mejora entre generaciones, pero no es una línea: son dominios, tareas y definiciones distintas, y en el propio campo de la bibliotecología el problema no se resolvió.

Tres lecturas. Bajan donde se mide y se castiga: en la tarea de resumir un documento dado, los mejores modelos del tablero de Vectara, actualizado el 11 de mayo de 2026, están entre el 1,8 % y el 5 % de inconsistencia factual (Vectara, 2026), y la ficha técnica de GPT-5 reporta, en tráfico real con navegación, un 4,5 % de afirmaciones con error en su modo razonador frente al 12,7 % de o3, en medición del propio proveedor (OpenAI, 2025c). No es monótona: la ficha técnica de o3 y o4-mini, en abril de 2025, reportó 33 % y 48 % de alucinación en PersonQA frente al 16 % de o1, dos modelos más capaces que alucinaban el doble (OpenAI, 2025e). No desaparecen: en 17.443 citas generadas sin recuperación, ningún modelo superó el 47,5 % de citas existentes (Zhao et al., 2026), y citas fabricadas ya entraron en las actas de NeurIPS, con alrededor de uno de cada veinte artículos de 2025 con al menos dos referencias probablemente alucinadas

(Russovich et al., 2026). La conclusión para la biblioteca no es «ya no hace falta verificar»: es que la verificación de referencias se volvió un servicio medible (caso 2 del capítulo 5).

6.3 Abiertos y propietarios: tres significados de «abierto»

La palabra «abierto» esconde tres cosas distintas. Un modelo de pesos abiertos es uno cuyos parámetros se pueden descargar y ejecutar en servidores propios, con licencias como MIT o Apache 2.0, o con licencias propias que imponen condiciones; los datos de entrenamiento no siempre se publican, y por eso conviene decir «pesos abiertos» y no «código abierto». Un modelo abierto de verdad publica pesos, código, datos y recetas de entrenamiento: es el caso de OLMO, del Allen Institute, y Latam-GPT documenta el origen de sus datos. Un modelo propietario se usa por API o por aplicación: el proveedor controla las versiones, la retención de los datos, los precios y el retiro del servicio. Para una biblioteca, la pregunta no es cuál es «mejor» en abstracto: es dónde quedan los datos de los estudiantes, cuánto cuesta, quién responde ante un incidente y en qué idioma funciona bien.

El mapa a septiembre de 2026, con la advertencia de que cambia cada mes y de que la licencia y la versión se verifican antes de contratar o instalar. Propietarios: Gemini (Google DeepMind, Estados Unidos y Reino Unido), GPT (OpenAI, Estados Unidos), Claude (Anthropic, Estados Unidos), Muse Spark (Meta, Estados Unidos, que en 2026 pasó a ofrecer modelos propietarios bajo ese nombre), Grok (xAI, Estados Unidos), Nova (Amazon, Estados Unidos), Command (Cohere, Canadá) y Mistral Medium y Large (Mistral AI, Francia, que combina propietarios y abiertos). Pesos abiertos: Kimi K2 y K3 (Moonshot AI, China; licencia MIT modificada), GLM-5 (Z.ai, China; MIT), DeepSeek V3 y V4 (DeepSeek, China; MIT), Qwen (Alibaba, China), MiMo (Xiaomi, China; MIT), Hunyuan (Tencent, China; Apache 2.0), Llama 4 (Meta; licencia Llama), gpt-oss (OpenAI; Apache 2.0), Gemma (Google), Phi (Microsoft), Nemotron (NVIDIA), OLMO (Allen Institute; totalmente abierto), Mistral Small 4 y Large 3 (Mistral AI; Apache 2.0), Falcon (Technology Innovation Institute, Emiratos Árabes Unidos), ALIA y Salamandra (Barcelona Supercomputing Center, España) y Latam-GPT (CENIA, Chile, con más de 60 instituciones de 15 países; construido sobre Llama 3.1 de 70.000 millones de parámetros, entrenado con más de 300.000 millones de tokens obtenidos «bajo permisos y licencias explícitas», lanzado el 10 de febrero de 2026 como «bien público orientado a democratizar el acceso a

esta tecnología»; CENIA, 2026). El liderazgo técnico en pesos abiertos está hoy en laboratorios chinos, con desempeño de frontera «a una fracción del presupuesto de cómputo de sus contrapartes cerradas» (Thunder Compute, 2026); los quince primeros del tablero de texto de Arena, a septiembre de 2026, son propietarios (Arena, 2026). Para una biblioteca de la región, Latam-GPT es la primera herramienta que debería estar probando con sus propias tareas, no la última.

6.4 Benchmarks: tres familias, tres preguntas

Los tableros de evaluación responden preguntas distintas y conviene no mezclarlas. Los exámenes estáticos (MMLU-Pro, GPQA, Humanity's Last Exam, SWE-bench para programación) son preguntas con respuesta conocida: miden capacidad en una tarea y se contaminan cuando sus preguntas entran al corpus de entrenamiento. Los tableros de fidelidad y factualidad miden si el modelo inventa o traiciona el contexto: el tablero de alucinaciones de Vectara (HHEM), FACTS Grounding de Google DeepMind, SimpleQA y AbstentionBench; son los que más importan a una biblioteca, porque el oficio es la fidelidad a la fuente. Y los tableros de preferencia humana, cuyo ejemplo principal es Arena, miden qué prefiere la gente, no qué es verdad. Para español y lenguas de España y América Latina existe La Leaderboard, presentada en ACL 2025, un tablero comunitario que evalúa 50 modelos sobre 66 conjuntos de datos en variedades del español, catalán, euskera y gallego (Grandury et al., 2025), y Arena tiene una categoría de texto en español.

6.5 Arena: cómo construye sus datos y para qué sirve

Arena, antes LMArena y Chatbot Arena, es el tablero que más circula y el que peor se entiende. Así construye sus datos: una persona entra, escribe cualquier pregunta, recibe dos respuestas de dos modelos anónimos lado a lado, y vota cuál prefiere o declara empate; solo después se revelan los nombres. Millones de esos votos alimentan un modelo estadístico de comparaciones por pares, Bradley-Terry, que convierte victorias y derrotas en una puntuación con intervalo de confianza, al estilo del Elo del ajedrez (Chiang et al., 2024). Nació en 2023 en el laboratorio LMSYS de la Universidad de California en Berkeley y hoy es una empresa; al 30 de septiembre de 2026 su tablero de texto acumulaba 8.602.501 votos sobre 410 modelos, y tiene tableros de agentes, desarrollo web, imagen, video, visión, documentos y búsqueda, además de categorías por idioma (Arena, 2026). El «control de estilo» descuenta la ventaja de las respuestas

más largas y con más formato, porque la gente prefiere lo largo. El conjunto de datos de preferencias es abierto en Hugging Face.

Lo que sirve: ver qué prefieren miles de usuarios reales en conversación abierta; comparar modelos abiertos y propietarios en un mismo tablero; explorar la categoría en español; reutilizar los datos abiertos. Lo que no sirve: saber si una respuesta es verdadera, porque mide preferencia y no factualidad; elegir una herramienta para la biblioteca sin probarla con las propias tareas; ignorar sus sesgos. «The Leaderboard Illusion», un estudio de 2025 de investigadores de Cohere, Princeton, Stanford y otras instituciones, encontró prácticas de prueba privada no divulgadas, con 27 variantes privadas probadas por un laboratorio antes del lanzamiento de Llama 4 y la posibilidad de publicar solo la mejor; asimetrías de acceso a los datos, con Google y OpenAI recibiendo cerca del 19,2 % y el 20,4 % de todos los datos de la arena mientras 83 modelos abiertos juntos recibían el 29,7 %; y ganancias de desempeño de hasta el 112 % atribuibles al acceso a esos datos (Singh et al., 2025). Arena respondió con cambios en sus políticas y el debate sigue.

Para una biblioteca, la utilidad real de Arena es doble: es un mapa de qué existe y qué prefieren los usuarios, útil para decidir qué probar; y es un ejemplo pedagógico de cómo se construye un dato, con un diseño experimental (comparación por pares, anonimato, votación) y un modelo estadístico explícitos, que puede enseñarse en la ALFIN como contraste con una cifra sin método. Ningún tablero reemplaza la prueba propia; los tableros sirven para decidir qué probar.

6.6 Siete preguntas antes de contratar o instalar

1. ¿Dónde se procesan y cuánto se retienen los datos? Se responde con las condiciones concretas del servicio (ubicación de los servidores, retención, uso para entrenamiento, subprocesadores) y con el artículo 25 de la Ley 25.326 sobre prestación de servicios de tratamiento por terceros; si la universidad tiene una herramienta institucional bajo su política de inteligencia artificial, esa es la que se usa, nunca una cuenta personal.
2. ¿Funciona bien en español y en el propio dominio? La Leaderboard y la categoría en español de Arena orientan; la respuesta la da una prueba propia con treinta tareas reales de la biblioteca (consultas de referencia, resúmenes de artículos del acervo,

clasificación con el libro de códigos), en español, con resultados comparados por una persona.

3. ¿Qué tasa de fidelidad tiene al resumir un documento dado? El tablero de Vectara y FACTS Grounding orientan; la prueba propia es resumir diez documentos del acervo y contar las afirmaciones que no están en el documento.
4. ¿Se abstiene cuando no sabe? AbstentionBench orienta; la prueba propia es la de la propiedad Parlanchinas.
5. ¿Cuánto cuesta por mil consultas y qué pasa si el proveedor retira el modelo? Los propietarios retiran versiones con pocos meses de aviso; un procedimiento validado con una versión se revalida con la siguiente.

Los de pesos abiertos exigen infraestructura propia o un proveedor de alojamiento, con su propio costo.

6. ¿Es accesible? Lector de pantalla, navegación por teclado, contraste, lenguaje claro; se prueba con las personas que lo necesitan.
7. ¿Qué dicen las siete propiedades de esa herramienta en nuestras pruebas? Las siete pruebas de la sección 6.1, registradas con fecha, modelo y versión, forman la ficha de la herramienta; se repiten con cada cambio de versión.

Las respuestas se registran en la ficha de herramienta del capítulo 9 y se revisan al menos cada semestre, porque todo lo que dice este capítulo lleva fecha.

7 Datos personales y ética

7.1 Argentina: la Ley 25.326, léida completa

La regla en Argentina es la Ley 25.326 de protección de los datos personales, del año 2000, y hay que leerla completa, no en tres artículos. El artículo 2 define los datos personales como información de cualquier tipo referida a personas físicas o de existencia ideal determinadas o determinables; define los datos sensibles, origen racial y étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical, salud y vida sexual; y define la disociación como el tratamiento que impide asociar la información «a persona determinada o determinable». El artículo 4 exige que los datos sean «ciertos, adecuados, pertinentes y no excesivos en relación al ámbito y finalidad para los que se hubieren obtenido» y prohíbe usarlos para fines distintos o incompatibles con aquellos para los que se recogieron. El artículo 5 exige el «consentimiento libre, expreso e informado», con las excepciones que la propia ley enumera. El artículo 7 limita el tratamiento de los datos sensibles. Los artículos 9 y 10 obligan a la seguridad y la confidencialidad; el 11 regula la cesión; el 25, la prestación de servicios de tratamiento por terceros, que es exactamente lo que ocurre cuando se suben datos a una herramienta de un proveedor externo. El artículo 28 exime las encuestas de opinión, las mediciones y los trabajos estadísticos cuando los datos no pueden atribuirse a una persona identificable, pero esa excepción

cubre el resultado: no autoriza por sí sola a recopilar, enlazar o transferir datos identificables en las etapas previas (Ley 25.326, 2000).

Traducido al oficio. El indicador de uso y permanencia exige, en algún momento, vincular registros identificables o seudonimizados; esa vinculación la hace un área autorizada de la universidad, bajo un diseño documentado que declare finalidad, base legal, quién accede, cuánto tiempo se conserva, quién es responsable y cómo se evita la reidentificación en grupos pequeños; la biblioteca recibe solo agregados protegidos. Antes de subir un solo comentario a una herramienta, se revisa y se depura localmente: quitar el nombre no basta, porque un texto libre con la carrera, el rol y una fecha puede identificar a alguien; se suprimen los cuasiidentificadores y las celdas pequeñas, y el protocolo lo revisa quien en la universidad tenga competencia para hacerlo. Y una vez transmitido un dato, ya se transmitió: la depuración se hace antes, no después.

7.2 Chile y otros países

En Chile, la Ley 21.719 reforma la protección de datos personales y entra en vigencia principal el 1 de diciembre de 2026, con una Agencia de Protección de Datos Personales nueva, bases de licitud explícitas, derechos ampliados y sanciones; una biblioteca chilena que construya indicadores en 2026 tiene que diseñarlos conforme a ella, y debe consultar el texto vigente y las

orientaciones de la agencia a medida que se publiquen. La ley chilena no es aplicable en Argentina, ni la argentina en Chile; los principios de finalidad, minimización, seguridad y licitud del tratamiento son comunes a ambas y a la mayoría de las leyes de la región (Brasil con la LGPD, Uruguay, Colombia, México, Perú, Ecuador), de modo que el protocolo del capítulo 9 se adapta cambiando la base legal y la autoridad, no la estructura. Las bibliotecas que operan con proveedores con servidores fuera del país deben prestar atención a las reglas de transferencia internacional de cada ley.

7.3 La Recomendación de la UNESCO y la confianza

Hay un marco más alto, que no es ley pero orienta. La Recomendación sobre la ética de la inteligencia artificial de la UNESCO, adoptada por los Estados miembros en noviembre de 2021, rechaza el uso de sistemas de inteligencia artificial para la calificación social y la vigilancia masiva (UNESCO, 2021). Una biblioteca que use analítica de aprendizaje para señalar a un estudiante «en riesgo», con la mejor intención, tiene que preguntarse primero por la finalidad y por los efectos de esa señal sobre la persona. Y

la confianza es frágil: en una encuesta a unos 2.200 estudiantes de ocho instituciones de Estados Unidos, «la mayoría de los estudiantes reportó altos niveles de confianza en las bibliotecas», pero «una minoría constante indicó poca o ninguna confianza» en el manejo de sus datos (Asher et al., 2022). La variable de confianza declarada de la compuerta I existe para que esa minoría no sea invisible.

7.4 Diez reglas de la compuerta I

Depurar localmente antes de subir. Minimizar: recoger solo lo que el indicador necesita. Separar: la biblioteca no custodia datos de rendimiento académico identificables; el cruce lo hace el área autorizada. Agregar: publicar solo celdas con al menos cinco personas, y suprimir o combinar las menores. Escribir el protocolo antes del primer dato, con fecha y firma. Usar la herramienta institucional, nunca una cuenta personal, y leer sus condiciones de retención y entrenamiento. Conservar lo mínimo y por el tiempo declarado. Informar a las personas, en la encuesta y en el sitio de la biblioteca, qué se recoge y para qué, con la opción de no responder. Revisar el protocolo con quien tenga competencia en la universidad (asesoría legal, comité de ética, oficina de datos). Y registrar la aplicación: un protocolo que no se aplica no protege a nadie.

8 Un piloto de noventa días

El piloto tiene pocas medidas y tiempo explícitamente asignado, porque aunque no se compre ninguna licencia hay horas de trabajo que se desplazan; en un contexto de restricción presupuestaria, medir no es pedir más, es poder mostrar qué hace la biblioteca con lo que tiene. Un equipo de dos o tres personas con un límite de horas acordado (por ejemplo, cuatro horas semanales en total) alcanza.

Días 1 a 30 · Alinear y proteger. Cruzar las tres listas de preguntas (evaluador, universidad, red) con los servicios de la biblioteca, usando la matriz del capítulo 9, y elegir un servicio y cuatro medidas, no más. Escribir el protocolo de datos y someterlo a revisión. Escribir el libro de códigos para los comentarios abiertos. Elegir la herramienta institucional y completar su ficha con las siete pruebas. Entregables: matriz con la columna «el de mi biblioteca» completada para las cuatro medidas; protocolo de datos revisado; libro de códigos; ficha de herramienta.

Días 31 a 60 · Medir lo que ya existe. Calcular la cobertura de la bibliografía obligatoria en una carrera piloto, con el criterio de disponibilidad escrito antes de contar. Aplicar la encuesta de contribución percibida a un segmento, con la pregunta de canal de descubrimiento. Acordar con el área autorizada el diseño de un análisis retrospectivo de uso y permanencia para una cohorte, aunque el cálculo llegue después del piloto. Descargar y archivar los informes COUNTER con su versión y filtros, y anotar las fechas de bloqueos y migraciones. Entregables: una ficha de indicador completa por medida, con bases e incertidumbre; informes COUNTER archivados con su registro.

Días 61 a 90 · Usar la inteligencia artificial y argumentar. Clasificar los comentarios de la encuesta con el procedimiento del caso 1, con referencia humana, dos corridas y auditoría por registro; registrar el ensayo con sus tiempos. Preparar la carpeta de evidencia de cuatro

páginas: las cuatro medidas, sus bases, sus límites y lo que todavía no se puede saber. Presentarla a la secretaría académica o a la oficina de evaluación institucional, y registrar la devolución. Entregables: registro del ensayo; carpeta de evidencia; acta de la devolución.

Al final del piloto. Tres preguntas: qué medida fue útil para una decisión, cuál no, y qué habría que cambiar en la segunda ronda. Y una decisión: qué dos medidas pasan a ser permanentes, con su periodicidad y su responsable. En una universidad en proceso de evaluación externa, la respuesta institucional al informe de pares es el documento natural para que la biblioteca aparezca con sus propios números.

9 Plantillas

Todas las plantillas se publican en formato editable junto a la guía. Aquí va su contenido, para que puedan copiarse a un procesador de texto o imprimirse.

9.1 Matriz ACEQUIA en una página

Título: «Matriz ACEQUIA · Indicadores de impacto de la biblioteca universitaria en la era de la IA · Propuesta para Argentina, versión 1 (octubre de 2026)». Siete franjas, una por compuerta, con cinco columnas: pregunta guía, indicadores sugeridos (fórmula), ancla, lo que no permite concluir, y «el de mi biblioteca», en blanco.

Compuerta · pregunta guía	Indicadores sugeridos (fórmula)	Ancla	Lo que no permite concluir	El de mi biblioteca
A · Alinear. ¿A qué pregunta del evaluador, de la universidad o de la red responde, y qué decisión informa?	Cobertura de la bibliografía obligatoria: 100 × títulos exigidos que cumplen el criterio de disponibilidad documentado ÷ títulos exigidos únicos del universo definido. Producción institucional depositada: 100 × obras elegibles de un año de publicación depositadas a la fecha de corte ÷ obras elegibles identificadas de ese año	Res. CONEAU 382/11, Anexo I 3.g; adaptaciones de ISO 11620:2023 A.1.1.2 y A.1.1.6; Ley 26.899 (universo y plazos); Guía CABID 2024	Que todos los estudiantes accedan simultáneamente o aprendan más; cumplimiento íntegro de la ley sin revisar plazos y excepciones	
C · Cadena. ¿En qué peldaño está cada cifra: insumo, producto, resultado o impacto?	Alcance de formación: personas únicas participantes ÷ población objetivo, y asistencias por cada 1.000 miembros. Desempeño inicial y final: cambio emparejado en una tarea comparable, con n completo y pérdidas. Transferencia a 90 días: proporción que completa una tarea real o registra uso, distinguiendo ambas	Estructura de ISO 16439:2014; adaptaciones de ISO 11620:2023 A.2.2.6 y A.2.2.8	Aprendizaje por el solo hecho de participar; causalidad sin un diseño que descarte otras explicaciones; impacto duradero a partir de la transferencia	
E · Evidencia. ¿Tengo evidencia inferida, solicitada y observada, y qué discrepancias muestran?	Contribución percibida: distribución de respuestas a preguntas neutrales (tareas, ayuda recibida, dificultades, necesidades no cubiertas), con «no la usé» y «no corresponde», por segmento. Uso y permanencia: comparación de cohortes definidas de usuarios y no usuarios, con ajuste justificado e incertidumbre, calculada por el área autorizada y entregada en agregados protegidos	Adaptaciones de ISO 11620:2023 A.2.4.7 y A.2.4.3; antecedentes: Soria, Fransen y Nackerud (2017), Stone y Ramsden (2013), Murray, Ireland y Hackathorn (2016)	Un efecto causal o un contrafáctico observado; que aumente el uso produzca por sí mismo más graduación	

Compuerta · pregunta guía	Indicadores sugeridos (fórmula)	Ancla	Lo que no permite concluir	El de mi biblioteca
Q · Quitar el ruido. ¿De qué informe, proveedor, versión y filtros sale cada cifra de uso digital?	Uso digital clasificado: métrica COUNTER específica, con el uso por agentes aparte cuando exista el método de acceso Agent, por estudiante equivalente con denominador documentado. Peticiones clasificadas como automatizadas: $100 \times$ peticiones clasificadas por la regla X ÷ peticiones observadas en la misma ventana, con revisión de errores. Descubrimiento mediado por IA: respuestas y referidos clasificados, con base y «desconocido» explícitos	COUNTER 5.11 (2026); COAR (2025); Casden et al. (2025); JSTOR (2026)	Identificación perfecta de seres humanos; la proporción verdadera de todo el tráfico automatizado; un total exhaustivo del descubrimiento mediado por IA	
U · Usar la IA con criterio. ¿Qué análisis se vuelve viable y cómo lo audito fuera de la herramienta?	Temas en comentarios abiertos: codificación con libro de códigos, clasificación por registro, referencia humana con selección documentada, precisión y exhaustividad por tema, repetibilidad por registro. Calidad de referencias: $100 \times$ confirmadas, con errores y no localizadas ÷ referencias de la muestra, reportadas por separado, con protocolo y fecha	Protocolo de auditoría (capítulo 5); Walters y Wilder (2023); Marco PAPAYAS (Villegas-Ceballos, 2026)	Uso de IA, fraude o lectura efectiva a partir de una referencia; ahorro de tiempo sin medir el ciclo completo	
I · Integridad de los datos. ¿Podría identificar a alguien con lo que recojo, enlace, transfiero o guardo?	Protocolo aplicado: existe, está fechado y hay evidencia de su aplicación. Minimización: $100 \times$ indicadores calculados solo con datos agregados o disociados ÷ indicadores publicados. Confianza declarada: distribución por segmento, con base	Ley 25.326 (arts. 2, 4, 5, 7, 9, 10, 11, 25 y 28); Recomendación de la UNESCO (2021); Ley 21.719 (Chile, vigencia principal 1-12-2026)	Un índice único de seguridad; que la confianza declarada garantice buen tratamiento; que la excepción estadística autorice el cruce previo de datos identificables	
A · Argumentar. ¿Quién va a usar esta evidencia, para qué decisión y cómo lo documentamos?	Visibilidad: indicadores de la biblioteca citados en autoevaluación, plan estratégico y política de IA. Participación efectiva: asiento en el órgano de gobernanza de IA y decisiones implementadas. Recursos: % del presupuesto institucional asignado	ACRL (2026), principios 1.2 y 1.3; MSCHE (2026), ejemplos de evidencia 3.3.8, 3.4.4 y 3.4.5; ISO 11620:2023 A.4.2.2	Cambio en los estudiantes; que visibilidad, asiento o presupuesto sean medidas de impacto	

Pie de la matriz: la lista de entrada de siete casillas (sección 5.1) y el piloto de noventa días (capítulo 8).

9.2 Ficha de indicador

Campo	Contenido
Nombre	
Compuerta	A · C · E · Q · U · I · A
Finalidad (qué pregunta responde)	
Decisión que informa y para quién	
Población (universo)	
Unidad de medida	
Fórmula (numerador ÷ denominador × 100 si es porcentaje)	
Período de referencia	

Campo	Contenido
Exclusiones (qué no cuenta y por qué)	
Fuente de datos (sistema, informe, versión, filtros)	
Responsable del cálculo	
Periodicidad	
Incertidumbre (bases pequeñas, pérdidas, cobertura)	
Protección de datos (protocolo aplicable; agregado o disociado)	
Ancla (norma o antecedente; «adaptación de» cuando corresponda)	
Lo que no permite concluir	
Fecha de la última medición y resultado, con base	

9.3 Protocolo de datos de usuarios (modelo)

1. Finalidad del tratamiento: qué indicador se calcula y qué decisión informa.
2. Base legal: artículo o excepción de la ley aplicable (Argentina: Ley 25.326; Chile: Ley 21.719), y consentimiento cuando corresponda, con el texto que se muestra a la persona.
3. Datos que se recogen, uno por uno, con su justificación de pertinencia (art. 4); datos que expresamente no se recogen.
4. Datos sensibles: declaración de que no se recogen, o justificación y medidas específicas (art. 7).
5. Flujo de datos: dónde nacen, quién los recibe, por qué sistema viajan, en qué país se procesan, qué proveedores intervienen (art. 25) y bajo qué condiciones contractuales.
6. Depuración previa a cualquier transferencia: campos suprimidos, cuasiidentificadores revisados, duplicados marcados, registros apartados y quién lo hizo.
7. Vinculación con otros registros: si la hay, qué área autorizada la realiza, con qué seudonimización, y qué recibe la biblioteca (solo agregados; umbral de celda mínima).
8. Conservación: cuánto tiempo, dónde, y cuándo y cómo se destruyen los datos.
9. Acceso: quiénes, con qué rol, y registro de accesos.
10. Derechos de las personas: cómo acceder, rectificar o suprimir sus datos, y a quién dirigirse.
11. Incidentes: a quién se informa, en qué plazo, y qué se hace.
12. Revisión: quién revisó este protocolo, fecha, firma; próxima revisión.

9.4 Libro de códigos (plantilla)

Encabezado: nombre del instrumento, pregunta exacta de la encuesta, unidad de análisis, fecha y versión. Reglas generales: una etiqueta principal y, si hay una segunda idea claramente distinta, una secundaria; nunca más de dos; abstención `requiere_revision` cuando ningún tema encaja o el texto es ininteligible; los duplicados se clasifican y se marcan. Por cada tema: nombre exacto (el que debe devolver la herramienta), definición de una línea, inclusiones, exclusiones (con el tema al que va cada caso excluido), dos ejemplos. Al final: lista de casos límite resueltos durante la codificación humana, con fecha, para que la siguiente ronda los herede.

9.5 Registro del ensayo de una medición asistida por inteligencia artificial

Fecha y hora. Responsable. Herramienta, modelo y versión visible; configuración si se conoce. Archivos de entrada con sus huellas SHA-256. Depuración realizada (registros apartados, campos suprimidos, quién lo hizo). Prompts usados, textuales. Corrida 1: hora, duración, salida guardada como. Corrida 2 (sesión nueva): hora, duración, salida guardada como. Referencia humana: quién codificó, cuántos registros, método de selección, semilla. Resultados de la auditoría: integridad; acuerdo por tema (precisión, exhaustividad, F1, base); acuerdo exacto de la etiqueta principal; repetibilidad por registro; citas comprobadas. Criterio de aceptación decidido antes de mirar, y si se cumplió. Adjudicaciones humanas. Tiempos observados del ciclo completo. Decisión: qué entra al informe y qué no.

9.6 Encuesta de contribución percibida (núcleo)

Segmentación inicial: rol (estudiante de grado y año, posgrado, docente, investigador, personal), unidad académica, modalidad. Preguntas, todas con las opciones «no la usé» y «no corresponde»: ¿Para qué tareas usó la biblioteca este semestre? (opciones múltiples: buscar bibliografía de una materia, estudiar en sala, acceder a plataformas digitales, pedir ayuda para buscar o citar, asistir a un curso o taller, depositar o consultar el repositorio, otra). ¿Qué ayuda recibió y de qué servicio? ¿Qué dificultades tuvo? ¿Qué necesitó y no encontró? ¿Cómo llegó a los recursos digitales que usó? (catálogo o sitio de la biblioteca, buscador, recomendación de un docente, desde una herramienta de inteligencia artificial, otra, no lo sé). ¿Qué confianza tiene en cómo la biblioteca maneja sus datos personales? (escala de cinco puntos y «no sé»). Comentario abierto final. Texto de información sobre datos: qué se recoge, para qué, cuánto se conserva, que las respuestas se reportan solo en agregados, y la opción de no responder.

9.7 Protocolo de verificación de referencias

Por referencia: identificador del trabajo, texto de la referencia tal como aparece, DOI o URL si existe. Pasos: resolver el DOI en doi.org y comparar título, autores,

revista, volumen, año; sin DOI, buscar el título entre comillas en el catálogo, en la fuente y en un buscador académico; registrar la fecha. Resultado: confirmada; con errores (cuáles); no localizada tras el protocolo; fabricada (solo con fundamento adicional, que se escribe). Reporte: por categoría, con la base; comparación solo entre muestras comparables; declaración expresa de que mide verificabilidad y no uso de inteligencia artificial ni deshonestidad.

9.8 Ficha de herramienta de inteligencia artificial

Nombre, proveedor, país, tipo (propietario, pesos abiertos, abierto), licencia, versión y fecha de corte declarada. Condiciones de datos: dónde procesa, cuánto retiene, si usa los datos para entrenar, subprocesadores, contrato institucional. Costo y modalidad. Accesibilidad. Resultados de las siete pruebas (Parlanchinas, Aduladoras, Políticamente posicionadas, Aprendices lentas, Yacentes, Aleatorias, Simuladoras), cada una con fecha, tarea y resultado. Resultado de la prueba de treinta tareas en español. Decisión (aprobada para qué usos; no aprobada para cuáles) y fecha de la próxima revisión.

10 Referencias

Todas las URL se consultaron entre el 28 de septiembre y el 1 de octubre de 2026. Las cifras sobre inteligencia artificial llevan la fecha de su fuente y deben releerse con esa fecha.

ACRL. (2026). *Standards for libraries in higher education* (versión revisada en 2026). Association of College and Research Libraries.

<https://www.ala.org/acrl/standards/standardslibraries>

AI Index Steering Committee. (2026). *The AI Index 2026 annual report*. Stanford University, Institute for Human-Centered AI. <https://hai.stanford.edu/ai-index/2026-ai-index-report>

Anthropic. (2026). *Models overview* [Documentación técnica].

<https://platform.claude.com/docs/en/models/overview>

Arena. (2026). *Text leaderboard* (consultado el 30 de septiembre de 2026: 8.602.501 votos, 410 modelos) y *How it works*. <https://arena.ai/leaderboard/text>; <https://arena.ai/how-it-works>

Asher, A. D., Briney, K. A., Jones, K. M. L., Regalado, M., Perry, M. R., Goben, A., Smale, M. A. y Salo, D. (2022). Questions of trust: A survey of student expectations and perspectives on library learning analytics. *The Library Quarterly*, 92(2), 151-171. <https://doi.org/10.1086/718605>

Bender, E. M. y Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185-5198). <https://doi.org/10.18653/v1/2020.acl-main.463>

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. En *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877-1901). <https://papers.nips.cc/paper/2020/hash/1457cOd6bfc4967418bfb8ac142f64a-Abstract.html>
- Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., Johary, I., Mara, A.-C., Romero, R., Lijffijt, J. y De Bie, T. (2026). Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2, artículo 7. <https://doi.org/10.1038/s44387-025-00048-0>
- Carroll, E., Bergstrom, T. y Hulbert, I. G. (2026). *Ithaka S+R US library survey 2025: Navigating uncertainty and change*. Ithaka S+R. <https://sr.ithaka.org/publications/ithaka-sr-us-library-survey-2025/>
- Casden, J., Romani, D., Shearer, T. y Campbell, J. (2025). Mitigating aggressive crawler traffic in the age of generative AI: A collaborative approach from the University of North Carolina at Chapel Hill Libraries. *Code4Lib Journal*, 61. <https://journal.code4lib.org/articles/18489>
- CENIA. (2026, 10 de febrero). *Latam-GPT: la primera IA regional abierta creada con datos latinoamericanos*. Centro Nacional de Inteligencia Artificial. <https://cenia.cl/2026/02/10/latam-gpt-la-primera-ia-regional-abierta-creada-con-datos-latinoamericanos/>
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P. y Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26, e53164. <https://doi.org/10.2196/53164>
- Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D. y Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), eaec8352. <https://doi.org/10.1126/science.aec8352>
- Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., ... Perez, E. (2025). *Reasoning models don't always say what they think* (arXiv:2505.05410). arXiv. <https://arxiv.org/abs/2505.05410>
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E. y Stoica, I. (2024). *Chatbot Arena: An open platform for evaluating LLMs by human preference* (arXiv:2403.04132). arXiv. <https://arxiv.org/abs/2403.04132>
- COAR. (2025, 3 de junio). *Open repositories are being profoundly impacted by AI bots and other crawlers: Report from a COAR survey*. Confederation of Open Access Repositories. <https://coar-repositories.org/news-updates/open-repositories-are-being-profoundly-impacted-by-ai-bots-and-other-crawlers-results-of-a-coar-survey/>
- Comisión Permanente Asesora de Bibliotecas y Documentación. (2024). *Guía para bibliotecas universitarias chilenas: Principios fundamentales* (versión 2024). Consejo de Rectoras y Rectores de las Universidades Chilenas. https://cabid.cl/wp-content/uploads/2025/03/Guia_para_bibliotecas_universitarias_chilenas.2024.pdf
- CONEAU. (1997). *Lineamientos para la evaluación institucional* (Resolución 094/97). Comisión Nacional de Evaluación y Acreditación Universitaria. https://www.coneau.gob.ar/archivos/publicaciones/criterio_symetod/lineamientos.pdf
- CONEAU. (2011). *Criterios y procedimientos para la evaluación externa* (Resolución 382/11). Comisión Nacional de Evaluación y Acreditación Universitaria. https://www.coneau.gob.ar/archivos/evaluacion/Criterios_procedimientosEE_Res38211.pdf
- CONEAU. (2022). *Evaluación externa de instituciones universitarias*. Comisión Nacional de Evaluación y Acreditación Universitaria. <https://www.coneau.gob.ar/archivos/publicaciones/EvaluacionExternalInstitucionesUniversitarias.pdf>
- CONEAU. (2023). *Informe de evaluación: Licenciatura en Ciencias Básicas, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Cuyo* (IF-2023-79818816-APN-DAC#CONEAU). <https://www.coneau.gob.ar/archivos/anexos/IF-2023-79818816-APN-DAC-CONEAU.pdf>
- CONEAU. (2025). *Nueva ordenanza CONEAU: Evaluaciones externas focalizadas* (Ordenanza 77/25). <https://www.coneau.gob.ar/coneau/nueva-ordenanza-coneau-evaluaciones-externas-focalizadas/>
- COUNTER. (2026). *Best practice on AI usage* (Release 5.1.1; actualización del 30 de junio de 2026). COUNTER Metrics. <https://www.countermetrics.org/code-of-practice/best-practice/bp-ai/>
- Dahl, M., Magesh, V., Suzgun, M. y Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64-93.

<https://doi.org/10.1093/jla/laae003>

Departamento General de Irrigación. (s. f.-a). *Ley de Aguas (1884)*. Aquabook.

https://aquabook.irrigacion.gov.ar/513_O

Departamento General de Irrigación. (s. f.-b). *La importancia del tomero*. Aquabook.

https://aquabook.irrigacion.gov.ar/1012_O

Digital Education Council. (2026). *AI in higher education Latin America survey 2026*.

<https://www.digitaleducationcouncil.com/resource-library-items/ai-in-higher-education-latam-survey-2026>

Łlugosz, D. A., Oliveira, A. y Díaz-Rodríguez, N. (2026). *The importance of being statistically earnest: A critical re-evaluation of GSM-Symbolic* (arXiv:2605.28700). arXiv. <https://arxiv.org/abs/2605.28700>

Ferracutti, V. M., Bonacorsi, B., Décima, R. V. y Díaz Pacífico, F. (2025). *Innovación en la gestión de bibliotecas universitarias: Diagnóstico nacional para la toma de decisiones estratégicas* [Ponencia]. 1.º Congreso Nacional en Innovación Universitaria, UNR-CIN. Biblioteca Digital del CIN.

<https://bibliotecadigital.cin.edu.ar/bitstreams/7de5245c-8df0-41de-a6ee-ebd60a9bd75f/download>

Google DeepMind. (2025, 21 de julio). *Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad* [Entrada de blog].

<https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/>

Grandury, M., Aula-Blasco, J., Falcão, J., Fourrier, C., González Saiz, M., et al. (2025). La Leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America. En *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

<https://aclanthology.org/2025.acl-long.1561/>

GT-EBU RedIAB. (2024). *Relevamiento de bibliotecas universitarias públicas 2024*. Grupo de Trabajo Evaluación de Bibliotecas Universitarias, Red Interuniversitaria Argentina de Bibliotecas.

<https://bibliotecadigital.cin.edu.ar/bitstreams/1b7df167-11ac-452f-ba58-ea1debbf2b85/download>

GT-EBU RedIAB/CIN. (2019). *Indicadores para evaluación de bibliotecas* (documento fechado en 2018 en el archivo; depositado en 2019). Ciclo de reflexión: Evaluación de bibliotecas universitarias.

<https://bibliotecadigital.cin.edu.ar/items/66293bc7-c244-4404-92b7-ea0daf7fa66e>

Güneş, Y., et al. (2026). Evaluating the reference accuracy of large language models in radiology: A comparative study across subspecialties. *Diagnostic and Interventional Radiology*, 32(2), 173–181.

<https://doi.org/10.4274/dir.2025.253101>

He, H. (2025, 10 de septiembre). *Defeating nondeterminism in LLM inference* [Entrada de blog]. Thinking Machines Lab. <https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/>

Holtzman, A., Buys, J., Du, L., Forbes, M. y Choi, Y. (2020). The curious case of neural text degeneration. En *International Conference on Learning Representations*. <https://arxiv.org/abs/1904.09751>

ISO. (2014). *ISO 16439:2014. Information and documentation: Methods and procedures for assessing the impact of libraries* (confirmada en 2025). International Organization for Standardization. <https://www.iso.org/standard/56756.html>

ISO. (2022). *ISO 2789:2022. Information and documentation: International library statistics* (6.ª ed.). International Organization for Standardization. <https://www.iso.org/standard/78525.html>

ISO. (2023). *ISO 11620:2023. Information and documentation: Library performance indicators* (4.ª ed.). International Organization for Standardization. <https://www.iso.org/standard/83126.html>

JSTOR. (2026, 21 de enero). *Usage data in a post-AI world*. <https://about.jstor.org/blog/usage-data-in-an-ai-world/>

Kalai, A. T., Nachum, O., Vempala, S. S. y Zhang, E. (2026). Evaluating large language models for accuracy incentivizes hallucinations. *Nature*. <https://doi.org/10.1038/s41586-026-10549-w>

Kirichenko, P., Ibrahim, M., Chaudhuri, K. y Bell, S. J. (2025). AbstentionBench: Reasoning LLMs fail on unanswerable questions. En *Advances in Neural Information Processing Systems* (NeurIPS 2025, Datasets and Benchmarks Track). <https://arxiv.org/abs/2506.09038>

- Lehr, S. A., Caliskan, A., Liyanage, S. y Banaji, M. R. (2024). ChatGPT as research scientist: Probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, 121(35), e2404328121. <https://doi.org/10.1073/pnas.2404328121>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. En *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459-9474). <https://arxiv.org/abs/2005.11401>
- Ley 21.719. (2024). Regula la protección y el tratamiento de los datos personales y crea la Agencia de Protección de Datos Personales. Diario Oficial de la República de Chile, 13 de diciembre de 2024. <https://www.bcn.cl/leychile/navegar?i=1209272>
- Ley 24.521. (1995). Ley de Educación Superior. Boletín Oficial de la República Argentina. <https://www.argentina.gob.ar/normativa/nacional/ley-24521-25394/texto>
- Ley 25.326. (2000). Protección de los datos personales. Boletín Oficial de la República Argentina. <https://www.argentina.gob.ar/normativa/nacional/ley-25326-64790/texto>
- Ley 26.899. (2013). Creación de repositorios digitales institucionales de acceso abierto, propios o compartidos. Boletín Oficial de la República Argentina. <https://www.argentina.gob.ar/normativa/nacional/ley-26899-223459/texto>
- Linardon, J., Jarman, H. K., McClure, Z., Anderson, C., Liu, C. y Messer, M. (2025). Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: Experimental study. *JMIR Mental Health*, 12, e80371. <https://doi.org/10.2196/80371>
- Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., ... Oak, S. (2025). *The AI Index 2025 annual report*. Stanford University, Institute for Human-Centered AI. <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- McGowan, A., et al. (2023). ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, 326, 115334. <https://doi.org/10.1016/j.psychres.2023.115334>
- Ministerio de Educación. (2007). *Resolución 1314/07: Estándares para la acreditación de la carrera de Medicina*. <https://www.coneau.gob.ar/archivos/Res1314.pdf>
- Ministerio de Educación. (2022). *Resolución 798/2022: Modificación de la Resolución Ministerial 1314/2007 (carrera de Medicina)*. <https://www.argentina.gob.ar/normativa/nacional/norma-362403/texto>
- Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S. y Farajtabar, M. (2025). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. En *International Conference on Learning Representations*. <https://arxiv.org/abs/2410.05229>
- MSCHE. (2026). *Standards for accreditation and requirements of affiliation* (15.ª ed.). Middle States Commission on Higher Education. <https://www.msche.org/standards/standards-15/>
- Mugaanyi, J., Cai, L., Cheng, S., Lu, C. y Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research*, 26, e52935. <https://doi.org/10.2196/52935>
- Murray, A., Ireland, A. y Hackathorn, J. (2016). The value of academic libraries: Library services as a predictor of student retention. *College & Research Libraries*, 77(5), 631-642. <https://doi.org/10.5860/crl.77.5.631>
- Oladokun, B. D., et al. (2025). Hallucination in scientific writing: Exploring evidence from ChatGPT versions 3.5 and 4o in responses to selected questions in librarianship. *Journal of Web Librarianship*. <https://doi.org/10.1080/19322909.2025.2482093>
- OpenAI. (2020). *GPT-3 dataset statistics: Languages by word count* [Conjunto de datos]. GitHub. https://github.com/openai/gpt-3/tree/master/dataset_statistics
- OpenAI. (2025c). *OpenAI GPT-5 system card* [Informe técnico de la industria] (arXiv:2601.03267). <https://arxiv.org/abs/2601.03267> (versión original de agosto de 2025: <https://cdn.openai.com/gpt-5-system-card.pdf>)
- OpenAI. (2025e, 16 de abril). *OpenAI o3 and o4-mini system card* [Informe técnico de la industria]. <https://openai.com/index/o3-o4-mini-system-card/>
- Ouyang, S., Zhang, J. M., Harman, M. y Wang, M. (2025). An empirical study of the non-determinism of ChatGPT in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2), artículo 42. <https://doi.org/10.1145/3697010>

- Paolucci, L. (2025). *Modos de integración de la inteligencia artificial en las bibliotecas de la Universidad de Buenos Aires* [Trabajo final de grado, Universidad Nacional de Mar del Plata]. Repositorio Institucional Humadoc. <https://humadoc.mdp.edu.ar/s/repositorio-institucional-humadoc/item/3148>
- Ponte, J. R. (2006). Historia del regadío: Las acequias de Mendoza, Argentina. *Scripta Nova: Revista Electrónica de Geografía y Ciencias Sociales*, X(218-07). <https://www.ub.edu/geocrit/sn/sn-218-07.htm>
- Robinson, N., Ogayo, P., Mortensen, D. R. y Neubig, G. (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. En *Proceedings of the Eighth Conference on Machine Translation* (pp. 392-418). <https://doi.org/10.18653/v1/2023.wmt-1.40>
- Rovegno, N. (2021). *La evaluación de las bibliotecas en la evaluación institucional externa* [Transcripción de presentación]. Ciclo de reflexión: Evaluación de bibliotecas universitarias, RedIAB/CIN. <https://bibliotecadigital.cin.edu.ar/bitstreams/adddbbcc-2198-46cd-8959-10e842a9b0c0/download>
- Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE*, 19(7), e0306621. <https://doi.org/10.1371/journal.pone.0306621>
- Russinovich, M., et al. (2026). *Phantom references: Hallucinated citations that survive peer review at top-tier conferences* (arXiv:2607.00738). arXiv. <https://doi.org/10.48550/arxiv.2607.00738>
- Secretaría de Educación. (2025). *Resolución 987/2025: Estándares para la acreditación de la carrera de Psicología*. <https://www.argentina.gob.ar/normativa/nacional/resoluci%C3%B3n-987-2025-414914/texto>
- Seifi, A., et al. (2026). Hallucination rate of peer-reviewed citations generated by large language models in neurocritical care. *Critical Care Explorations*. <https://doi.org/10.1097/cce.0000000000001474>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., ... Perez, E. (2024). Towards understanding sycophancy in language models. En *The Twelfth International Conference on Learning Representations*. <https://arxiv.org/abs/2310.13548>
- SID-UNCUYO. (s. f.-a). *Alfabetización informacional*. Sistema Integrado de Documentación, Universidad Nacional de Cuyo. <https://sid.uncu.edu.ar/alfin/>
- Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., et al. (2025). *The leaderboard illusion* (arXiv:2504.20879). arXiv. <https://arxiv.org/abs/2504.20879>
- Soria, K. M., Fransen, J. y Nackerud, S. (2017). The impact of academic library resources on undergraduates' degree completion. *College & Research Libraries*, 78(6), 812-823. <https://doi.org/10.5860/crl.78.6.812>
- Stephenson, R. y Armstrong, C. (2026). *Student generative AI survey 2026* (HEPI Report 199). Higher Education Policy Institute y Kortext. <https://www.hepi.ac.uk/wp-content/uploads/2026/03/HEPI-Report-199-Gen-AI-Survey-2026.pdf>
- Stone, G. y Ramsden, B. (2013). Library Impact Data Project: Looking for the link between library usage and student attainment. *College & Research Libraries*, 74(6), 546-559. <https://doi.org/10.5860/crl12-406>
- Thunder Compute. (2026, 14 de septiembre). *Best open source LLMs (September 2026)*. <https://www.thundercompute.com/blog/best-open-source-llms>
- Ülkir, M., et al. (2026). Reference hallucination, citation reliability, and readability of large language models in anatomy-related question answering. *Clinical Anatomy*. <https://doi.org/10.1002/ca.70187>
- UNCUYO. (2026a, 21 de abril). *La UNCuyo establece principios y una guía para el uso responsable de la inteligencia artificial generativa* (Resolución 262/2026 CS). Universidad Nacional de Cuyo. <https://www.uncuyo.edu.ar/academica/la-uncuyo-establece-principios-y-una-guia-para-el-uso-responsable-de-la-inteligencia-artificial-generativa>
- UNCUYO. (2026b). *Guía de buenas prácticas para el uso de inteligencia artificial generativa en la UNCuyo* (Anexo II de la Resolución 262/2026 CS; copia enlazada desde el sitio institucional). <https://guia-ia-uncuyo.lovable.app/docs/Guia-IA-UNCUYO-2026.pdf>
- UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*. <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence?hub=1063>
- Vectara. (2026, 11 de mayo). *Hallucination leaderboard (HHEM-2.3)*. GitHub. <https://github.com/vectara/hallucination-leaderboard>

Villegas-Ceballos, S. (2026). *Marco PAPAYAS: Siete propiedades estructurales de la inteligencia artificial generativa para una alfabetización crítica* (3.ª ed., documento base, CC BY 4.0). Unidad iIA de eLibro e IAparaBibliotecas.com. <https://www.elibro.com/iia>

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q. y Luong, T. (2024). FreshLLMs: Refreshing large language models with search engine augmentation. En *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 13697–13720). <https://doi.org/10.18653/v1/2024.findings-acl.813>

Walters, W. H. y Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, 14045. <https://doi.org/10.1038/s41598-023-41032-5>

Zhao, C., et al. (2026). *Do deployment constraints make LLMs hallucinate citations? An empirical study across four models and five prompting regimes* (arXiv:2603.07287). arXiv. <https://doi.org/10.48550/arxiv.2603.07287>

ANEXOS



Las láminas, la ficha del encuentro y la Unidad iIA

Las 43 láminas de la conferencia tal como se proyectaron, la ficha del
encuentro y una nota sobre el autor y la Unidad de Innovación e IA de eLibro.

A Las láminas de la conferencia

Presentación de 43 láminas, con siete láminas de paso entre capítulos; identidad visual de la Unidad iIA de eLibro; licencia CC BY-NC 4.0. Las notas del orador de cada lámina reproducen el texto de la primera parte de esta obra.



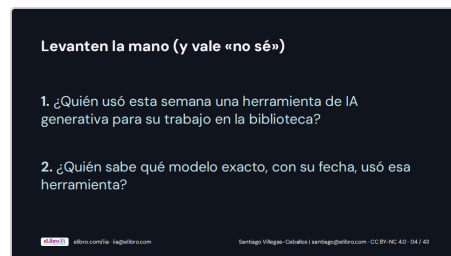
01 · Portada



02 · La acequia



03 · Tres hechos de esta casa



04 · Dos preguntas para la sala



05 · Capítulo 1 · Lo básico, verificado



06 · PAPAAYAS en una lámina



07 · P · Parlanchinas



08 · A · Aduladoras



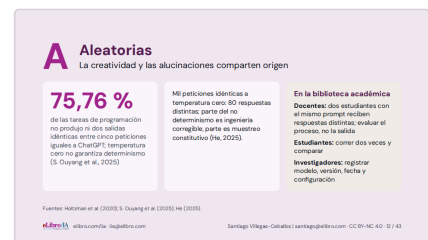
09 · P · Políticamente posicionadas



10 · A · Aprendices lentos



11 · Y · Yacentes



12 · A · Aleatorias



Simuladoras

Imitación estadística frente a razonamiento genuino

hasta 65 %

de cada de desempeño al
enfrentar a un problema
matemático escrito una clásica
evaluación (GSM-Symbolic;
Mirzadeh et al., 2022)

Los modelos razonadores
verbalizan las pistas que usan
en menos del 5% de los casos
(Y. Chen et al., 2023). un sistema
de DeepMind alcanzó la medalla
de oro de la IMO 2020 bñlo y
colapsos conviven.

En la laboite de académica
Docentes el razonamiento
visible no es una ventaja útil al
cálculo.
Estudiantes: un buen resumen
no es comprensión
Investigadores: tratarlos como
primera aproximación con
revisión experta

Fuentes: Kotler y Keller (2002); Mirzadeh et al. (2023); Dagher et al. (2023); Y. Chen et al. (2023); Google DeepMind (2023).

edutech | edutech@uic.edu | edutech.com

Sanjato Villegas-Ceballos | sanjato@edutech.com | CC-BY-NC 4.0 | 1/48

13 · S · Simuladoras

2

CAPÍTULO 2 · 19:38 A 19:42

¿Mejoran las tasas de alucinación?

Si, entre generaciones; con condiciones

1

Los datos, verificado

2

¿Mejoran las tasas de alucinación?

3

Modelos estadísticos y modelos computacionales

4

Como se comparan

5

Indicadores

6

Demstración

7

La guía en pdf

[#Mejoras](#)
[#MejorasConCondiciones](#)
[@santiagovillagelacortez](#)

Santiago Village-Lacortez | santiago@glabrics.com | CC-BY-NC 4.0 | 14 / 43

14 · Capítulo 2 · ¿Mejoran las tasas de alucinación?

Dentro de cada estudio, la generación nueva fabrica menos referencias que la anterior, con una excepción

Study	Previous Generation (%)	New Generation (%)
Woolley y Willett (2003)	25%	10%
Lehr et al. (2004)	25%	20%
Chell et al. (2004)	25%	20%
Qi et al. (2006)	25%	20%
Gómez García et al. (2005)	94%	100%
Cluster 1	25%	10%
Cluster 2	25%	10%
Cluster 3	25%	10%
Cluster 4	25%	10%
Cluster 5	25%	10%
Cluster 6	25%	10%
Cluster 7	25%	10%
Cluster 8	25%	10%
Cluster 9	25%	10%
Cluster 10	25%	10%
Cluster 11	25%	10%
Cluster 12	25%	10%
Cluster 13	25%	10%
Cluster 14	25%	10%
Cluster 15	25%	10%
Cluster 16	25%	10%
Cluster 17	25%	10%
Cluster 18	25%	10%
Cluster 19	25%	10%
Cluster 20	25%	10%
Cluster 21	25%	10%
Cluster 22	25%	10%
Cluster 23	25%	10%
Cluster 24	25%	10%
Cluster 25	25%	10%
Cluster 26	25%	10%
Cluster 27	25%	10%
Cluster 28	25%	10%
Cluster 29	25%	10%
Cluster 30	25%	10%
Cluster 31	25%	10%
Cluster 32	25%	10%
Cluster 33	25%	10%
Cluster 34	25%	10%
Cluster 35	25%	10%
Cluster 36	25%	10%
Cluster 37	25%	10%
Cluster 38	25%	10%
Cluster 39	25%	10%
Cluster 40	25%	10%
Cluster 41	25%	10%
Cluster 42	25%	10%
Cluster 43	25%	10%
Cluster 44	25%	10%
Cluster 45	25%	10%
Cluster 46	25%	10%
Cluster 47	25%	10%
Cluster 48	25%	10%
Cluster 49	25%	10%
Cluster 50	25%	10%

elab. por: mariafernanda.saguro@unq.edu.ar

Santiago Vilella-Castillo | santiago.vilella@unq.edu.ar | CC BY-NC-ND 4.0 / 12-43

15 · La tendencia, dentro de cada estudio

[illegible]

16 · La tendencia, a lo largo de tres generaciones

Tres conclusiones para la biblioteca		
Bajan donde se mide y se castiga Al reunir un documento delo, la mayoría de los modelos de acceso al menos del 10% de los resúmenes y casi todos en menos de 5 minutos (ver Tabla 4, H4M-2, 15-5-2008). Es un trabajo de la institución, no un estudio revisado por pares.	No es monótona Más capacidad no garantiza menos inversión en la misma tarea. O3 Beta (2.93 % OPT-4) o 5.6 % Vectra (2008) Operati operati 10.5 % en O3 frente a 4.6 % en el Personal (4 abril de 2005, autorretrato).	No desaparecen Sin recuperación de información, los modelos de 2008 indican entre 6.9 % y el 10.5 % de las citas dominios especializados (Gall et al., 2008) y citos fabricados ya aparecen en actas de Heuristics (Bauer et al., 2008), pero sin revisión por pares).
La verificación de referencias dejó de ser una advertencia: es un servicio de la biblioteca que se puede medir.		

17 · Tres conclusiones para la biblioteca

3

CAPÍTULO 3 • 15:42 A 15:47

Modelos abiertos y modelos propietarios

¿Qué significa «abierto» y quién hace qué, a septiembre de 2026

1. La familia, los usuarios

2. Empresas que usan el aprendizaje?

3. Alumnos, profesores y moderas/moderadores

4. ¿Cómo se compran?

5. Metodologías

6. Cursos en línea

7. ¿A qué y a qué?

<https://edx.org/cursos/edx/orgs/edxorg/courses>
[Santiago Illuque-Catalonia/santiago.illuque-catalonia.org](https://santiago.illinois-catalonia.org/edxorg/courses)
CC-BY-NC-ND 15/43

18 · Capítulo 3 · Modelos abiertos y propietarios

Abierto no es una palabra: son tres cosas distintas

Personas abiertas	Abierto de verdad	Proprietario
Se descargan los parámetros y se ejecutan en servidores propios; licencias MIT, Apache 2.0 o licencias propias con condiciones (Lama, GLM); los datos de entrenamiento no siempre se publican.	Personas, código, datos y recetas de entrenamiento publicados (OLMo del Allen Institute; Latam-GPT documenta sus fuentes).	Se usa por API o aplicación; el proveedor controla versiones, retención de datos, precios y retiro del servicio.

Para la biblioteca la pregunta no es cuál es mejor; es dónde quedan los datos de los estudiantes, cuánto cuesta, qué responde y en qué idiomas funciona bien.

Fuentes: Thunders Computer (2008), CERN (2008), Arxiv (2008)

ed.ubio | edubio.com | info@edubio.com

Santiago Villalba Caballero | santiago@edubio.com | CC-BY-NC-4.0 / 19/43

19 · Qué significa «abierto»

Familias de modelos, laboratorios y país • septiembre de 2026

Proteptistas	Personas abiertas	
Genetics: Douglas Dimpled EE, UU y Reino Unido	Kim G4 K3 Ki3 Moonshot EE, China MIT	GLM-5.2-Chat, China MIT
GPT: OpenAI EE, EE, UU	DeepSeek V3 V3.1 DeepSeek EE, China	Qwen: Alibaba, China
Claude: Anthropic EE, EE, UU	Mixto: Xiaomi, China MIT	Huanyan: Tencent, China Apacche 2.0
Muse: Superf. Meta EE, UU	Llama 4: Meta EE, EE, UU; Icaricia Llama EE, EE, UU	gpt-oss: OpenAI EE, UU Apacche 2.0
Grak: Axi EE, UU	Germana: Google EE, UU	Phi: Microsoft EE, UU
New: Amazon EE, EE, UU	Nemotron: NVIDIA EE, UU	Qwen: Alibabalab: EE, UU, laboratorio abierto
Modelo: Google Gemini, Canadá	Modelo: Intel y Lingo 3: Intel EE, EE, UU	LLM-7: Ericsson, Alemania EE, UU
Matriz: Meta EE, EE, UU (Francia)	Ally y Starlight: BSC España	Llama: GPT: Meta EE, EE, UU en 16 países, además Llama 3.3 EE, EE, UU Apacche 2.0

Lista indicativa, cambia con los meses. Verificar fechas y veracidad antes de contratar o cancelar. Fuente: www.4mat.es (20-09-2026), www.4mat.es (20-09-2026), www.4mat.es (20-09-2026), www.4mat.es (20-09-2026)

4mat.es info@4mat.es www.4mat.es

Santiago Villegas-Cordero info@sigablog.es CC-BY-NC-ND 4.0 30/43

20 · El mapa, a septiembre de 2026

4

CAPÍTULO 4 - 15:47 A 15:53

Cómo se comparan

Benchmarks, Arena y siete preguntas antes de elegir

1. ¿Qué es el benchmarking?

2. ¿Por qué las empresas deberían utilizarlo?

3. ¿Qué ventajas y desventajas tiene?

4. ¿Cómo se compara?

5. ¿Existen riesgos?

6. ¿Cómo se relaciona?

7. ¿Qué es el juego?

[Santiago Villagrán-Gutiérrez | santiago@elcomercio.com](#) CC-BY-NC-ND 3.0 43

21 · Capítulo 4 · Cómo se comparan

Tres familias de evaluación, tres preguntas distintas		
Exámenes estáticos Preguntas con respuesta conocida (MMU-Pro, GQFA, Hummer's Last Exam, SYBEBench) miden capacidad en un tema, se contaminan cuando entran al corpus de entrenamiento.	Fidelidad y factualidad Si el modelo inventa o traiciona el contenido, se mide la fidelidad de las afirmaciones de Vectara (HHEM), FACTS Grounding, SimpleQA, AbstractiveDerech; son los que más importan a una biblioteca.	Preferencia humana Personas reales eligen entre dos respuestas alternativas: Arena (antes LMarena y Chatbot Arena); mide qué prefere la gente, no qué es verdad.

22 · Un mapa de los benchmarks

Arena: 8.602.501 votos sobre 410 modelos (30-9-2026)

1

Una persona escribe una pregunta

→

2

Dos modelos analíticos responden todo a todo

→

3

La persona vota cuál prefiere (o empatía)

→

4

Un modelo estadístico de comparaciones por pares (Bradley-Terry) construye millones de votos en una puntuación con intervalo de confianza

Origen
Nació en 2003 en LMBS, Universidad de California en Berkeley (Ching et al., 2002) hoy es una empresa con tablero de texto, agenda, desarrollo web, imágenes, videos, música, documentos o búsqueda y rankings por día/mes.

•Style control•
Descuentra la ventaja de las respuestas más largas y con más formato.

Fuente: arena (30-9-2002) Ching et al. (2002)

elbri elbri@csa.berkeley.edu

Santiago Villegas-Castelló | santiago@elbri.com CC BY-NC 4.0 / 23 / 43

23 · Cómo construye Arena sus datos

Lo que Arena mide y lo que no Sírvase para - ver qué preferen miles de usuarios reales en conversación abierta - comparar modelos abiertos y propietarios en un mismo tablero - explorar la categoría en preferencias - el conjunto de datos de preferencias es abierto en Hugging Face	No sírvase para - saber si una respuesta es verdadera: mide preferencia, no facticidad - elegir una respuesta para la biblioteca sin problemas con las propias leyes - ignorar sus sesgos: 27 sesgos en frases probadas por un laboratorio antes de publicar - los Google y OpenAI recibieron cerca del 19,2 % y el 20,4 % de los datos de la arena, 63 modelos abiertos juntos el 29,7 % (Singh et al., 2025)
A septiembre de 2026, los quince primeros del tablero de textos son propietarios. Fuente: Arena (2026), https://arena.lmsys.org (2025)  arena.lmsys.org huggingface.co	Santiago Villegas-Gutiérrez santiago@lmsys.org CC-BY-NC-ND 4.0 / 24

24 · Lo que Arena mide y lo que no

Siete preguntas antes de contratar o instalar una herramienta de IA

- 1 ¿Dónde se procesan y cuánto se retienen los datos? (Ley 25.326, art 26; condiciones concretas del servicio)
- 2 ¿Funciona bien en español y en nuestro dominio? (La Leaderboard: prueba propia con 30 tareas reales)
- 3 ¿Qué tasa de fidelidad tiene en resumir un documento dado? (Vectara, FACTIS)
- 4 ¿Se abstiene cuando no sabe? (AbstractBouncer; prueba propia)
- 5 ¿Cuánto cuesta por mil consultas y qué pasa si el proveedor retira el modelo?
- 6 ¿Es accesible (lector de pantalla, teclado, lenguaje claro)?
- 7 ¿Qué dicen PAPAÍYAS de esa herramienta en nuestras pruebas?

Desarrollados en la giga-Med4 con crédito: capulus 6 (fellos.com/ai).
giga-Med4 | atborreco@igagigamed.com

Santiago Vargasi | santiago@fellos.com | CC-BY-NC 4.0 2023/24

25 · Cómo elegir en una biblioteca

5

CAPÍTULO 5 - 16:53 A 16:03

Indicadores

El problema y el marco ACEQUIA, propuesta para Argentina

1. La Misión verificada

2. ¿Alcanzará los niveles de bienestar?

3. ¿Indicador adecuado y medible para el problema?

4. Cómo se comparan

5. Indicadores

6. Comunicación

7. ¿A quién y por qué?

[Inicio](#)
[Inicio/Contenido](#)
[Indicadores](#)

Sanja Villegas-Catalina | sanja@noroeste.com | CC BY-NC-ND 3.0 / 4.0

26 · Capítulo 5 · Indicadores

Res. CONEAU 382/11 - evaluación institucional externa

Anexo 1.3.g. - lo que valoran los pares

dotación y disponibilidad - calidad de la prestación de los servicios - personal - equipamiento - presupuesto - otros

Anexo 1.2.a - la información básica

cantidad de bibliotecas y personal - cantidad de usuarios activos - estadística anual de consultas o préstamos por tipo de servicios

La información básica «no es suficiente para hacer la evaluación» (Rovegno, CONEAU, 2021).

Fuentes: CONEAU (2018), Rovegno (2021).

 atributoscon@-la.gub.uy

Santiago Yllaga - Gubio - santiago@la.gub.uy CC-BY-NC-ND 7/143

27 · Lo que pide la CONEAU

Lo que declaran los estudiantes (muestra de 29 instituciones de 10 países)

92 % de los estudiantes declara usar IA

65 % de quienes la usan, la usa para buscar información

30 % está de acuerdo con que su institución integra la IA como esperaba

Fuente: Digital Education Council (2026) más de 300000 respuestas (22341 estudiantes, 739 docentes), Argentina y Chile incluidos, muestra de instituciones participantes, no probabilística.

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 28 / 43

28 · La biblioteca de 2026, en una muestra

Tres fenómenos que no conviene confundir

Agentes de IA que acceden a nuestros recursos licenciados
COUNTER S11 los identifica con el método de acceso «Agent» (extensión opcional) y los deja fuera de las vistas estándar (COUNTER, 2026).

Tráfico automatizado abusivo
Más del 80 % de 66 repositorios encuestados por COAR lo sufre; Chapel Hill bajó un 81 % sus peticiones semanales al filtrarlo (COAR, 2026; Cadden et al., 2025).

Descubrimiento fuera de la plataforma
La persona lee un resumen generado y nunca llega al recurso («uso invisible», JSTOR, 2026).

Fuente: COUNTER (2026), COAR (2025), Cadden et al. (2025), JSTOR (2026).

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 29 / 43

29 · Tres fenómenos que no conviene confundir

Las bibliotecas llevan décadas evaluando su contribución.

La IA cambia cómo se accede a la información y vuelve más difícil leer ciertas estadísticas de uso.

ACEQUIA organiza esa evidencia: qué medimos, para qué decisión, con qué límites y con qué cuidado de las personas.

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 30 / 43

30 · La tesis

ACEQUIA · siete compuertas entre el insumo y el impacto · propuesta, versión 1

lo que entra

Alberar **C**ualidad **E**videncia **Q**uitar el ruido **U**so de la IA **I**ntegridad de los datos **A**rgumentar

lo que fluye

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 31 / 43

31 · Las siete compuertas

A · Alinear
Todo indicador responde a una pregunta que alguien ya hizo: el evaluador (Res. 382/16) la universidad (Res. 262/2026 CS), la red (RedMAG).

C · Cadena
Cuatro pedruzcos (estructura de ISO 8439:2014)

Insumo → Productos → Resultados → Impacto

Cobertura de la bibliografía obligatoria:
100 + títulos exigidos que cumplen el criterio de disponibilidad + títulos exigidos (adaptación de ISO 15620:2023 A.1.1.2)

Escalera ALFIN: asistencia por cada 1000 miembros (A.2.2.8): % que mejoró entre una prueba inicial y una final comparables, con n y pérdidas (adaptación de A.2.2.8).

Fuente: CONEAU (2025, 2026), UNESCO (2026), ISO (2024, 2023), OT, EBU, RedMAG (2026).

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 32 / 43

32 · A y C

E · Evidencia

Contribución percibida: preguntas neutras (temas, avales, dificultades, necesidades no cubiertas) con «no lo sé». Usa «permanencia» colores comparados por el área autorizada: empresas en agendas protegidas, antedecentes de asociación, no de causalidad (Doris, Freeman y Nadelson, 2007; Stone y Ramstein, 2023).

Q · Quitar el ruido
Ningún número de uso digital se reporta sin decir de qué informe sale, con qué versión y con qué filtros.

Uso digital clasificado: métrica COUNTER, proveedor, versión y filtro, uso por agentes aparte. **Peticiones clasificadas como automatizadas:** DO · clasificadas por la regla «i» · observadas. **Desautonomadamente medidas por IA:** «¿cómo llegó a este recurso?», con «no sé».

Fuente: ISO (2024, 2023), Stone et al. (2023), Stone y Ramstein (2023), COUNTER (2026), COAR (2025), Cadden et al. (2025), JSTOR (2026).

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 33 / 43

33 · E y Q

U · Usar la IA con criterio
Temas en comentarios abiertos: libro de códigos, clasificación por registro, referencia humana, acuerdo por tema, reutilización en sesión nueva. **Calidad de referencias:** confirmadas, con errores y no localizadas, por separado; mide verificabilidad, no deshonestedad.

I · Integridad de los datos
Depurar antes de subir; cruce por área autorizada; Ley 25.326 (arts. 2, 4, 5, 7, 9 a 18, 25 y 26); UNESCO (2021), Chile; Ley 21.719 desde el 1-12-2026. **Tres variables:** protocolo aplicado, minimización, confianza declarada.

A · Argumentar
Visibilidad, participación efectiva y recursos, separadas, ACRL (2026): «curar una colección de evidencia». La guía de la UNESCO ya asigna al SID el catálogo de herramientas: convertirlo en evidencia y ofrecer indicadores para verificación crítica, competencias e inclusión.

Fuente: Walter y Walter (2023), Ley 25.326 (2006), UNESCO (2021), Ley 2179 (2024), ACRL (2026), UNESCO (2026), Carroll, Benjamin y Hubert (2026).

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 34 / 43

34 · U, I y A

6 CAPÍTULO 6 · 16:03 A 16:08

Demostración
La IA clasifica, nosotros auditamos

1. La métrica verificada 2. Chequear los datos de documentación 3. Medir los avances y presentarlos 4. Cruce por comparación 5. Indicadores 6. Demostración 7. La guía y el piloto

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 35 / 43

35 · Capítulo 6 · Demostración

Cinco minutos: 60 comentarios sintéticos, un libro de códigos, una auditoría por registro

1 Archivo revisado y depurado localmente 2 Clasificación por identificador 3 Resultado de la auditoría del ensayo de esta mañana

DATOS SIMULADOS, Institución ficticia. Panel B: capturas del ensayo, métricas de Ensayo previo, datos simulados.

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 36 / 43

36 · Salida a la demostración

Lista de entrada, no certificado

☐ Depuré localmente antes de subir

☐ Escribí el libro de códigos

☐ Pedí la clasificación por registro

☐ Comparé persona y modelo en los mismos registros

☐ Repetí en una sesión separada

☐ Comprobé cada cita fuera de la herramienta

☐ Documenté herramienta, versión, fecha, prompts y criterio de aceptación

Protocolo completo en la guía-Medir con criterio, capítulo 5, con eligen de auditoría por registro.

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 37 / 43

37 · Antes de creerle a una medición hecha con IA

7 CAPÍTULO 7 · 16:08 A 16:12

La guía y el piloto
Un camino para profundizar y noventa días para empezar

1. La métrica verificada 2. Chequear los datos de documentación 3. Medir los avances y presentarlos 4. Cruce por comparación 5. Indicadores 6. Demostración 7. La guía y el piloto

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 38 / 43

38 · Capítulo 7 · La guía y el piloto

La guía: diez capítulos para ver el tema en profundidad

Medir con criterio
Guía para usar inteligencia artificial en la biblioteca de impacto de la biblioteca académica.

1. Para quién es esta guía y cómo usarla
2. Qué se evalúa y quién lo pide: Argentina, Chile y el mundo
3. El marco ACEQUIA, compuesta por computa
4. Qué cambia con la IA en los datos de uso
5. Usar la IA para medir: cinco casos con protocolo, prompts y auditoría
6. Elegir herramientas: abiertos, propietarios, benchmark y PARADIS
7. Datos personales y ética
8. Piloto de noventa días
9. Plantillas
10. Referencias

Unidad II de el libro, versión 1, octubre de 2026. CC BY 4.0

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 39 / 43

39 · La guía

Un piloto de noventa días, con tiempo asignado

Días 1 a 30
cruzar las tres listas de preguntas; elegir un servicio y cuatro medidas; escribir el protocolo de datos y someterlo a revisión

Días 31 a 60
cobertura de bibliografía en una carrera piloto: encuesta de contribución percibida; acuerdo con el área autorizada para un análisis retrospectivo

Días 61 a 90
clasificar y auditar los comentarios: carpeta de evidencia de cuatro páginas, con límites; presentaría a la secretaría académica y registrar la devolución

Con un límite de horas acordado. Registrando también lo que todavía no se puede saber.

Desarrollado en la guía-Medir con criterio, capítulo 5.

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 40 / 43

40 · Noventa días

El caudal importa, pero no basta.
También queremos saber a quién llega el agua y qué permite hacer.

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 41 / 43

41 · El tomero

Gracias

Santiago Villegas-Coballos · Director de la Unidad II de el libro
santiago@elibro.com · elibro.com/ia · @MedeJean

16:50 · Panel ALFIN en bibliotecas universitarias: Raquel Písi, Silvia Chamorro y Fernanda Olmedo (UNCUJO); Graciela Buttos (UNSL).

Micro-cursos gratuitos y certificados: elibro.com/ia-micro-cursos

Guía-Medir con criterio y matriz ACEQUIA

eLibro.com | elibro.com/ia | iagelibri.com

Santiago Villegas-Coballos | santiagov@elibro.com CC BY-NC 4.0 42 / 43

42 · Gracias y paso al panel

Preguntas

Siete propiedades - Marco PAPAYAS

P - Parlanchinas

A - Aduadoras

P - Polticamente posicionadas

A - Aprendices lentas

Y - Yacentes

A - Aleatorias

S - Simuladoras

Siete computas - Marco ACEQUIA

A - Alinear

C - Cadena

E - Evidencia

Q - Quitar el ruido

U - Usar la IA con criterio

I - Integridad de los datos

A - Argumentar



elibro.com/ia
santiago@elibro.com
@iia.ubn

 elibro.com/ia | santiago@elibro.com

Santiago Villegas-Ceballos | santiago@elibro.com | CC BY-NC 4.0 | 43 / 43

B

Ficha de la conferencia y del encuentro

Encuentro	IV Encuentro Interamericano de Bibliotecarios, de Información y Alfabetización (EIBIA), «Hacia la transformación y conexión de las bibliotecas»
Organiza	Sistema Integrado de Documentación (SID) de la Universidad Nacional de Cuyo, Mendoza, Argentina
Fechas	30 de septiembre al 2 de octubre de 2026
Conferencia	«Indicadores en la era de la IA: cómo medir el impacto de la biblioteca universitaria para la acreditación», conferencia magistral del Día 2, eje Bibliotecas Universitarias
Fecha y lugar	Jueves 1 de octubre de 2026, 15:20 a 16:30, Biblioteca Central de la Universidad Nacional de Cuyo; transmisión por el canal de YouTube de las bibliotecas de la UNCUIYO
Taller asociado	«Alfabetización crítica en IA desde la Biblioteca, laboratorio PAPAYAS», viernes 2 de octubre de 2026, 10:30 a 13:00, Biblioteca Central, con el equipo de Alfabetización Informacional del SID
Marcos	Marco PAPAYAS, tercera edición (documento base con 58 referencias verificadas, corte 1 de septiembre de 2026); Marco ACEQUIA, propuesta para Argentina, primera versión
Materiales publicados	Esta obra; la matriz ACEQUIA en una página; los archivos de la demostración (datos sintéticos, libro de códigos, prompts, guion de auditoría por registro y generador). Todo en https://www.elibro.com/ia

C Sobre el autor y la Unidad iIA de eLibro

Santiago Villegas-Ceballos es bibliotecario, con formación que cruza la bibliotecología, la ingeniería y la innovación social. Dirige la Unidad de Innovación e Inteligencia Artificial (iIA) de eLibro Corp., donde es docente principal, y codirige IAparaBibliotecas.com. Es miembro electo del Standing Committee de la Sección de Tecnología de la Información de la IFLA. Es autor del Marco PAPAYAS, una mnemotecnica didáctica con licencia abierta para la alfabetización crítica en inteligencia artificial, y del marco ACEQUIA que esta obra propone. Contacto: santiago@elibro.com · @MedeJean.

La Unidad de Innovación e Inteligencia Artificial (iIA) de eLibro trabaja por una inteligencia artificial crítica y responsable para bibliotecas: amplía el acceso al conocimiento, potencia la docencia y acelera la investigación. Ofrece micro-cursos gratuitos y certificados, cursos institucionales por cohorte, cursos avanzados en alianza con la Escuela PARIX de la Fundación Germán Sánchez Ruipérez y consultoría con implementación (asistentes y agentes de referencia, automatización de flujos, modelos de pesos abiertos, políticas y gobernanza de inteligencia artificial, analítica y tableros). Su equipo docente reúne especialistas de Colombia, Chile, Ecuador y Argentina.



Unidad iIA de eLibro

<https://www.elibro.com/iia> · iia@elibro.com

Micro-cursos gratuitos y certificados:

<https://www.elibro.com/iia-micro-cursos>

Preguntas frecuentes: <https://www.elibro.com/iia-pyr>

Su Biblioteca. Su Evolución.

