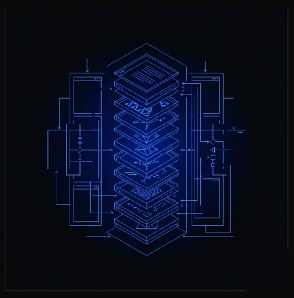
Evidencia de la que se puede responder

RESEARCH INTELLIGENCE PLATFORM



Un sistema operativo de evidencia que convierte literatura académica en decisiones institucionales trazables — una afirmación a la vez, anclada al estudio que la sostiene.

ESTADO	STACK	IA	VALIDACIÓN
Producción	Next.js 16 · React 19 · Firebase	AWS Bedrock · Claude	ESSA + 26 suites

01 EL PROBLEMA

CONCLUSIONES DE LAS QUE NADIE PODÍA RESPONDER

Una red de colegios K-12 de aprendizaje personalizado acelerado por IA toma decisiones de currículum e instrucción a diario: qué secuencia adoptar, qué intervención escalar, qué práctica retirar. Esas decisiones se apoyaban en resúmenes generados con IA que mezclaban, en una sola operación, dos cosas que jamás debían ir juntas: la extracción fiel de lo que un estudio reportaba y la interpretación pedagógica de lo que ese estudio significaba para el aula.

El resultado era imposible de auditar. No se podía distinguir lo que el autor de un paper había medido de lo que el modelo había inferido, ni rastrear una recomendación hasta el estudio concreto que la sostenía.

Las conclusiones llegaban como narrativa fluida y sin anclaje: frases huérfanas que sonaban con autoridad pero que nadie podía verificar.

A esto se sumaba un problema de identidad: el sistema previo referenciaba estudios por su posición en un arreglo, y cada reordenamiento corrompía silenciosamente esas referencias.

Para una institución que se define por estar basada en evidencia, el cuello de botella no era producir texto, sino producir texto del que se pudiera responder.

02 EL OBJETIVO

TRAZABILIDAD A NIVEL DE AFIRMACIÓN

OBJETIVO PRIMARIO

Construir un sistema operativo de evidencia, no un generador de resúmenes: que **toda conclusión publicable tuviera trazabilidad a nivel de afirmación**. Ninguna frase del documento final podía existir sin un mapeo estructurado a uno o más estudios fuente identificados por un `refId` estable.

Junto a eso, tres metas concretas: separar de forma irreversible la extracción de la interpretación, aplicar un marco de calidad metodológica determinista y reproducible, y mantener un corpus vivo y versionado capaz de incorporar estudios nuevos o re-evaluaciones sin reconstruirse entero.

El producto debía cerrar el ciclo completo, desde el paper crudo hasta la decisión institucional documentada, de modo que un revisor humano pudiera intervenir en cualquier punto sin romper la cadena de evidencia. Es una herramienta interna de admin para dos personas: el operador del pipeline y una investigadora del equipo de learning science incorporada como par de revisión.

03 EL SISTEMA

UN PIPELINE DE NUEVE ETAPAS AUDITABLE

El sistema se organiza como un pipeline de nueve etapas encadenadas, donde cada etapa tiene su propio agente, su contrato de entrada y salida, y un artefacto persistido y versionado. La decisión central, documentada como **ADR-001**, fue prohibir que extracción, evaluación, clustering, síntesis y redacción ocurrieran en una misma llamada al modelo: la separación de capas es lo que hace el sistema auditable.

Ingesta→Extracción→Clustering→Síntesis→Triangulación→Claims→Capítulos→Final→Editorial

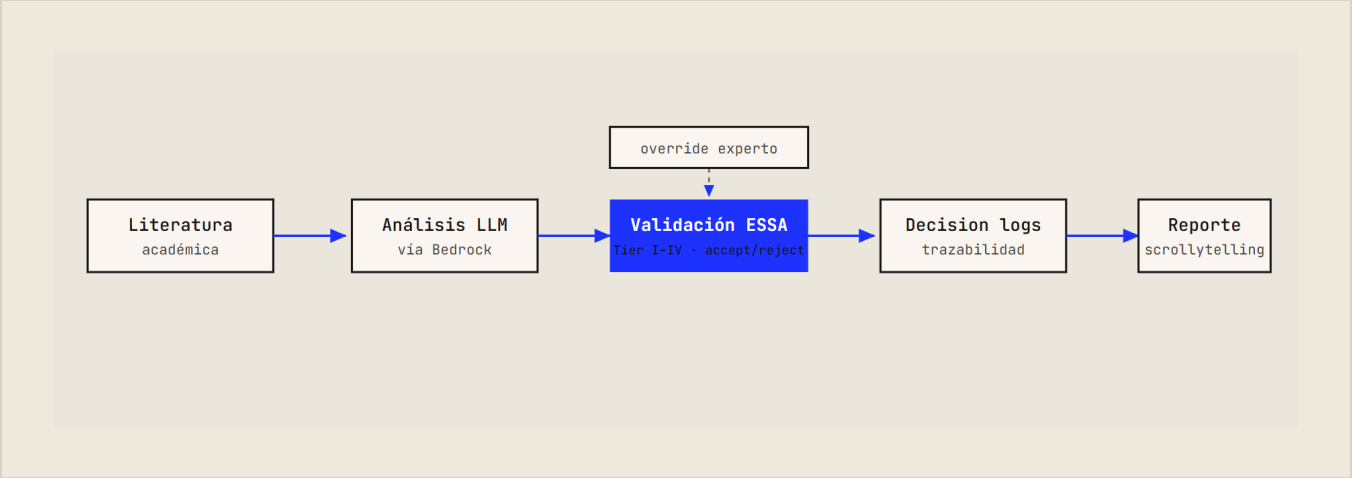


FIG. 1 — PIPELINE LITERATURA → LLM → VALIDACIÓN ESSA → DECISION LOGS → REPORTE · CADA ETAPA PERSISTE UN ARTEFACTO VERSIONADO CON IDENTIDAD POR REFID INMUTABLE, NUNCA POR ÍNDICE POSICIONAL

Separar extracción de interpretación — nunca en la misma llamada al modelo

Contexto. El sistema anterior mezclaba en una operación qué medía el estudio y qué significaba para el aula, y referenciaba estudios por su posición en un arreglo — cada reordenamiento corrompía las referencias.

Decisión. El marco de calidad (ADR-012) bifurca el *appraisal* según el tipo de estudio —ESSA para cuantitativo, CASP para cualitativo, ambos para mixto— y codifica el árbol de decisión ESSA como reglas deterministas explícitas en el prompt: `randomAssignment + controlGroup + muestra adecuada → Sólida`. El agente no infiere el nivel con libertad.

Trade-off aceptado. Más etapas y más artefactos que un resumidor de un paso — pero es lo que vuelve cada conclusión rastreable hasta el estudio que la sostiene. Las 21 decisiones quedaron registradas como ADRs.

04 EL FUNDAMENTO

TRES ESTÁNDARES DE SÍNTESIS, VUELTOS SOFTWARE

La arquitectura traduce dos estándares reconocidos de síntesis de evidencia al dominio del software, sumando un tercero para lo cualitativo. El principio rector — **fidelidad primero**— viene directo de la disciplina de la revisión sistemática: representar el estudio antes de interpretarlo.

MARCO	QUÉ APORTA	CÓMO OPERA EN EL SISTEMA
ESSA	Fuerza de la evidencia en cuatro niveles	Árbol de decisión determinista codificado en prompt; clasifica cada referencia.
PRISMA 2020	Transparencia de revisiones sistemáticas	Auditoría automática de 12 ítems contra los artefactos del sistema al cierre.
CASP	Appraisal de evidencia cualitativa	Bifurca el análisis para que cuanti, cuali y mixto no se fuercen a una misma lógica.

Los cuatro niveles ESSA estructuran todo el corpus, del más fuerte al de menor evidencia empírica:

TIER 1 · SÓLIDA — RCT TIER 2 · MODERADA — cuasi-experimental

TIER 3 · PROMISORIA — correlacional

TIER 4 · LÓGICA — modelo lógico

> *Lo cuantitativo se sintetiza por efectos y moderadores; lo cualitativo por temas y mecanismos; lo mixto se integra explícitamente mediante triangulación. Representar el estudio antes de interpretarlo.*

05 LA VALIDACIÓN

EL CORAZÓN FUNCIONAL: REVISIÓN HUMANA REFERENCIA POR REFERENCIA

La validación ocurre en **/research** mediante un flujo explícito de revisión humana. Tras correr la validación ESSA, la investigadora y el operador revisaban cada referencia una por una, con tres veredictos posibles — **Aceptada**, **Advertencia** o **Rechazada**— sobre el nivel que el agente había asignado según el árbol determinista. El sistema permite *override* humano: el juicio del modelo es una propuesta, no una sentencia.

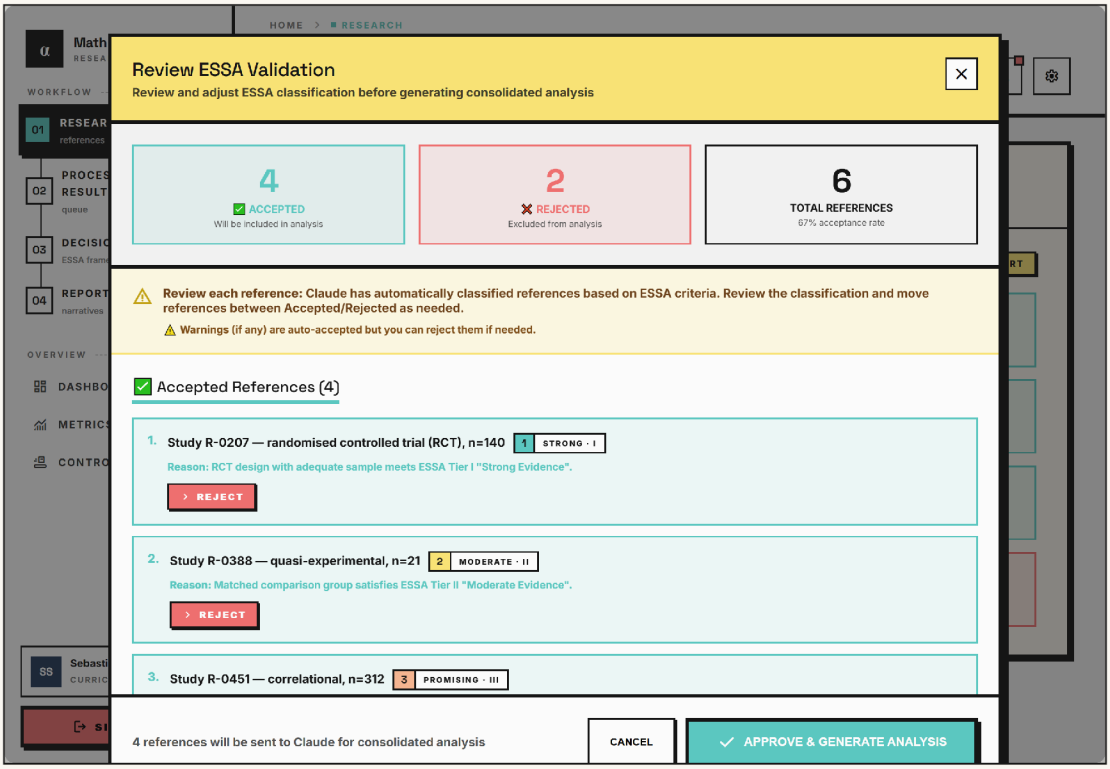


FIG. 2 — /RESEARCH CON ESSA VALIDATION · CADA REFERENCIA MUESTRA SU NIVEL PROPUESTO Y EL OVERRIDE ACCEPTED / WARNING / REJECTED · NINGUNA SELECCIÓN SE APRUEBA SIN PASAR EL FILTRO HUMANO

06 EL SEGUNDO FILTRO

AUDITORÍA PRISMA AUTOMÁTICA Y DECISION LOGS

El cierre del documento final añade una segunda capa de control automático, la **auditoría PRISMA**, que verifica doce ítems del estándar contra los artefactos del sistema —desde criterios de elegibilidad y riesgo de sesgo por estudio hasta resultados de síntesis y limitaciones— y reporta un cumplimiento total, parcial o no conforme. Así, ninguna investigación se publica sin pasar el doble filtro: revisión humana sobre la evidencia y verificación automática sobre el reporte. La vista **/decision-logs** registra cada decisión institucional con su cadena de evidencia intacta.

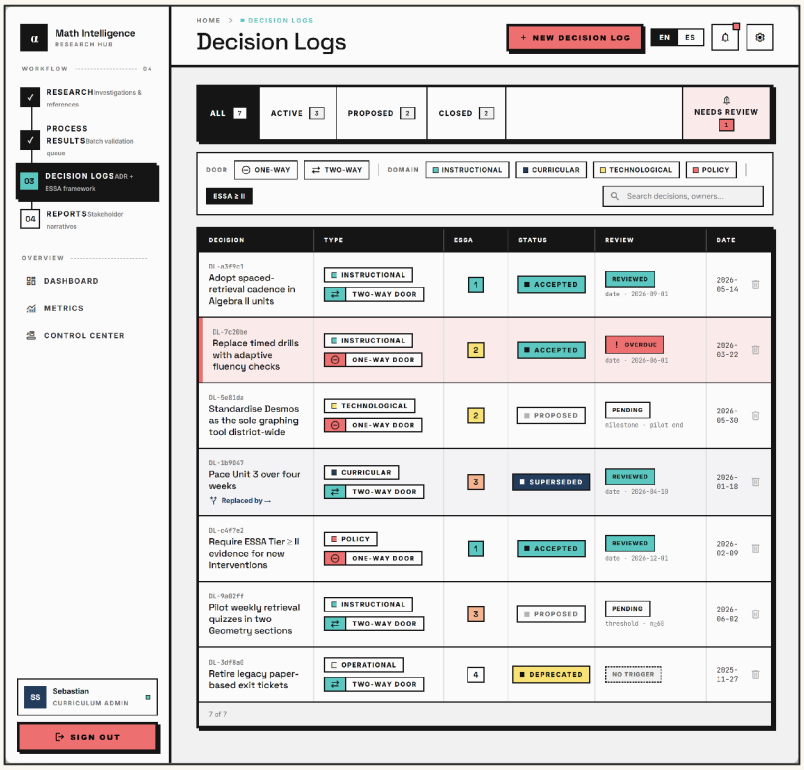


FIG. 3 — /DECISION-LOGS · CADA DECISIÓN QUEDA ANCLADA A LA INVESTIGACIÓN Y A LOS ESTUDIOS QUE LA SOSTIENEN · TRAZABILIDAD DEL PAPER A LA DECISIÓN DOCUMENTADA

07 CASO EMBLEMA

MEASURING THE 2X FACTOR

Measuring the 2x Factor: Speed and Mastery Depth Metrics for Adaptive HS Math. La tesis: la afirmación de que las plataformas adaptativas de matemáticas pueden duplicar la velocidad de aprendizaje **no puede verificarse ni refutarse** con la

evidencia disponible tal como está estructurada hoy — ningún estudio operacionaliza velocidad y profundidad de dominio simultáneamente contra un benchmark de élite.

HALLAZGOS, CON EFFECT SIZES

d 0,10–0,68 · g 0,37 — Las plataformas adaptativas producen efectos positivos moderados sobre el rendimiento matemático, pero ningún estudio los midió contra benchmarks de élite como 800 en SAT Math o 5 en AP.
velocidad ≠ profundidad — Avanzar rápido dentro de una plataforma puede producir comprensión superficial; hay que separar métricas de velocidad de métricas de dominio conceptual.
d 0,30–1,34 — La práctica intercalada y el espaciado producen los efectos más robustos sobre retención a largo plazo; la práctica distribuida duplica o triplica la retención frente a la masiva.
profundidad > aceleración — El dominio profundo de precálculo predice el rendimiento en cálculo universitario con más del doble de impacto que simplemente haber cursado el curso.
wheel-spinning — La práctica extensa sin logro de dominio es predecible desde indicadores tempranos; tiempo de uso y número de ejercicios son insuficientes para evaluar el aprendizaje real.

08 CASO EMBLEMA

EL MARCO PROPUESTO Y SU BASE DE EVIDENCIA

MARCO DE MEDICIÓN DE DOBLE EJE VELOCIDAD-DOMINIO

(1) una métrica de velocidad válida distinta del tiempo en tarea; (2) una métrica de profundidad de dominio anclada en competencias fundacionales verificadas; y (3) un benchmark de rendimiento de élite calibrado.

La tensión sin resolver. La contradicción velocidad–profundidad permanece abierta porque ningún estudio operacionalizó ambas métricas a la vez contra un benchmark de élite; los efectos para estudiantes de alto rendimiento bajo presión extrema siguen sin documentarse. La implicación es directa: toda plataforma que afirme aceleración («2x») debería demostrar empíricamente la operacionalización de sus métricas y reportar efectos sobre evaluaciones externas antes de declararse efectiva. Es el marco que vuelve *falsable* la promesa central del modelo.

Clusters de evidencia	29
Afirmaciones mapeadas	45
Fuentes	45 · todas SÓLIDA
Confianza del reporte	Moderada

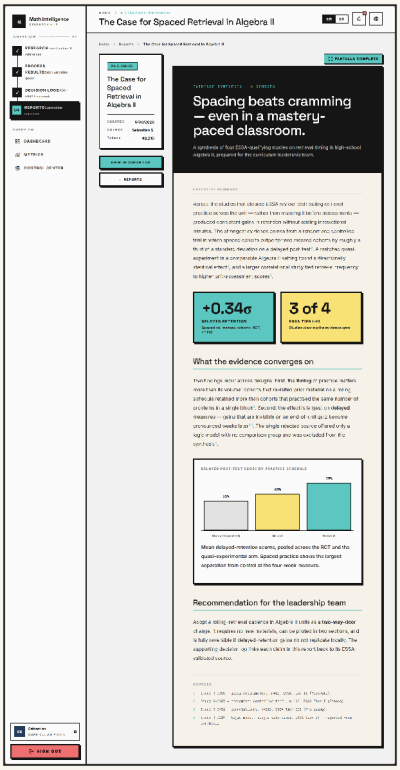


FIG. 4 — REPORTE SCROLLYTELLING GENERADO · UNA INVESTIGACIÓN CONVERTIDA EN NARRATIVA NAVEGABLE, CON CADA AFIRMACIÓN ANCLADA A SUS FUENTES · EXPORTABLE A PDF

09 GÉNESIS V0

SIETE ESTUDIOS MANUALES QUE JUSTIFICARON LA PLATAFORMA

Antes de la plataforma hubo una prueba de concepto: **siete estudios sintetizados a mano con Gemini**. Demostraron que convertir literatura en investigación accionable era valioso — y que hacerlo a mano no escalaba ni dejaba trazabilidad auditable. Ese v0 es lo que justificó construir el sistema operativo de evidencia.

1. **Academic Audit Report.** Auditoría longitudinal de desempeño en matemáticas de secundaria (SAT Math y AP) frente a instituciones de referencia. cap. 07
2. **Cognitive Architecture del SAT.** La brecha 650→800 como fluidez estructural y arbitraje con Desmos. cap. 08
3. **AP/SAT Curricular Intersection.** El sistema radicular algebraico que sostiene Cálculo y Estadística avanzados. cap. 09
4. **Automation Threshold Roadmap.** El umbral de automaticidad como puente hacia Cálculo acelerado. cap. 10
5. **MS Persistence vs SAT Stamina.** De la persistencia en 7º grado a la resistencia cognitiva del 1550. cap. 11
6. **Elite Freshman Profiles.** Auditoría recursiva de la compresión del SAT Math y la preparación STEM. cap. 12
7. **Technical Calculation Protocol.** Métricas propietarias para el sprint diario de matemáticas. cap. 13

Los siete reaparecen como capítulos propios de este book: el v0 manual fue el origen del corpus que la plataforma luego escaló y formalizó.

10 EL CORPUS, DE UN VISTAZO

LA SALIDA DE LA PLATAFORMA A ESCALA

De siete estudios manuales a un cuerpo de investigación validado, vivo y versionado. Esto es lo que la plataforma produjo:

19
Investigaciones
542
Estudios sintetizados
465
Referencias

FIG. 5 — CORPUS AT A GLANCE · CADA CIFRA ES PRODUCTO DEL PIPELINE DE NUEVE ETAPAS CON DOBLE FILTRO · EL PROTOTIPO MANUAL PREVIO CON GEMINI PARTIÓ DE 7 ESTUDIOS

Cada una de estas investigaciones pasó por el árbol de decisión ESSA, la revisión humana referencia por referencia y la auditoría PRISMA automática de doce ítems. La interfaz de validación con override humano y la auditoría PRISMA son las dos garantías que distinguen este sistema de un generador de resúmenes: cada conclusión que sale es rastreable hasta los estudios que la sostienen. Es, en la práctica, la infraestructura que alimenta el corpus de investigación de los capítulos siguientes.

11 APÉNDICE

ÍNDICE DEL CORPUS — LAS 19 INVESTIGACIONES

INVESTIGACIÓN	ESTUDIOS	REFS.	FECHA
Measuring the 2x Factor: Speed and Mastery Depth Metrics for Adaptive HS Math	211	195	2026-05-07
Critical Variables for Big Bang Adaptive Math Curriculum Implementation in High School	58	53	2026-05-07
Change Management and Leadership Resistance in Non-Traditional School Reform	28	14	2026-05-07
Educational Data Mining and Learning Analytics in Mathematics	23	20	2026-04-06
Advanced Placement Mathematics Achievement Factors	23	13	2026-04-06
Mastery-Based Learning and Spaced Practice in Mathematics	19	11	2026-04-06
Interleaved Practice and Problem-Type Discrimination	18	8	2026-04-06
From Computation to Proof: Pedagogical Transitions for Advanced Adolescents in Self-Directed Environments	17	17	2026-04-21
Computational Fluency and Mathematical Modeling Integration in Advanced Secondary Curricula	16	16	2026-04-21
Self-Regulated Learning in Digital Mathematics Environments	16	11	2026-04-06
Beyond AP and SAT: Assessment Frameworks for Mathematical Maturity — International Models and Admissions Signaling	15	15	2026-04-21
Sequencing Advanced Mathematics for Adolescent STEM Readiness: International Comparative Analysis	15	15	2026-04-21
Conceptual vs. Procedural Knowledge Development in Mathematics	15	8	2026-04-06
Mathematics Anxiety and Emotional Regulation in Digital Learning	14	9	2026-04-06
Sustaining Mathematical Motivation: Wellbeing, Coach Support, and Resilience in Accelerated Adolescent Math Learning	11	19	2026-04-21
Large Language Models and Conversational Tutoring in Mathematics	11	11	2026-04-06
Cognitive Load Management in Autonomous Digital Learning	11	9	2026-04-06

Intelligent Tutoring Systems Effectiveness in Secondary Mathematics	11	11	2026-04-06
Projected Learning Gains of LLM-Based Conversational Tutoring in Secondary Mathematics: A Monte Carlo Simulation Based on AutoTutor Evidence	10	10	2026-04-12

12 CONSTRUCCIÓN Y RESULTADO

AI-FIRST, SOMETIDO A CONTRATOS DE DATOS

Construí la plataforma de forma AI-first y en alta cadencia, fines de semana incluidos, sobre un stack deliberadamente moderno: Next.js 16 con App Router, React 19 y Tailwind v4, asumiendo sus breaking changes. El motor de inferencia es AWS Bedrock: **Claude Sonnet 4.6** para todas las etapas de clasificación y publicación, y **Opus** para los asistentes interactivos. La disciplina que sostiene la confianza en el output es la validación con Zod en cada borde de datos, en lectura y en escritura, contra contratos centralizados: un artefacto que no valida no entra al corpus.

Los runners de cada agente están desacoplados del cliente Bedrock mediante una interfaz inyectable, lo que permite correr el pipeline completo contra un mock en los tests. La CI ejecuta typecheck, lint, tests y build en cada push, con **26 suites** entre contratos, integración, regresión y unitarias. La plataforma quedó en producción como app de admin, sirviendo el ciclo completo desde el paper hasta la decisión documentada.

Commits	`~773 · 31-mar → 5-jun 2026`
Etapas del pipeline	**9** · ingesta → editorial
Decisiones de arquitectura	**21 ADRs** registradas
Marcos de calidad	ESSA + CASP + PRISMA 2020 (12 ítems)
Validación de datos	Zod en lectura y escritura, en cada borde
Suites de test en CI	**26** · contratos, integración, regresión, unitarias
Usuarios admin	2 · operador + investigadora de learning science
Estado	**En producción**

APRENDIZAJES

- **De arquitectura.** La auditabilidad no se agrega después, se diseña desde la primera capa: separar extracción de interpretación y anclar la identidad en un refId inmutable resolvió de raíz una clase entera de bugs que ningún parche posterior habría contenido.
- **De IA y juicio experto.** Codificar el árbol ESSA como reglas deterministas dentro del prompt, en lugar de dejar que el modelo razonara el nivel libremente, hizo el resultado reproducible y dejó el juicio donde corresponde: en el override humano referencia por referencia.
- **De diseño.** Trabajar el flujo de validación junto a una investigadora de learning science afinó la interfaz más que cualquier especificación escrita — ver a alguien recorrer la revisión en vivo mostró exactamente dónde la trazabilidad ayudaba y dónde estorbaba. Y los 21 ADRs probaron su valor al operar sobre un stack con breaking changes.

Qué demuestra el portafolio, en conjunto

EL CRITERIO ES EL PRODUCTO.

Seis piezas para un mismo programa de matemáticas. Vistas de lejos, no son seis apps: son seis ejercicios del mismo músculo — decidir qué medir, traducir ciencia del aprendizaje a software y sostener la calidad sin degradarla.

01 EL ARCO

DE VER, A ENTRENAR, A FUNDAMENTAR

El portafolio recorre un programa entero, no una función suelta. Primero **ver**: una capa de analítica que volvió legible el riesgo de ~1.600 estudiantes y la convirtió en acción matinal sin trabajo manual Sec. 1. Luego **entrenar**: dos plataformas que atacan los saltos de dificultad más caros del examen —la fluidez con la calculadora del SAT y la argumentación del AP— traduciendo distinciones de ciencia del aprendizaje en mecánicas de software Sec. 2. Y por último **fundamentar**: una infraestructura que convierte literatura académica en decisiones de currículum trazables, cada afirmación anclada a su fuente Sec. 3.

Cada producto resolvió su problema y, a la vez, habilitó al siguiente: la analítica detectó la brecha que el entrenamiento atacó; la necesidad de fundamentar esas decisiones empujó la plataforma de evidencia.

No es un catálogo de demos: es un sistema pensado por capas, donde la decisión de qué construir vino siempre de un diagnóstico, no de una lista de funciones.

02 EL HILO COMÚN

LO QUE SE REPITE EN LAS SEIS PIEZAS

CRITERIO TRAZABLE

Cada decisión estructural quedó documentada con su *por qué* y su trade-off aceptado — del Desmos embebido y persistente al scoring determinista por completitud. El artefacto muestra el juicio, no solo el resultado.

LEARNING SCIENCE VUELTA SOFTWARE

CERC como tipo de datos, las cinco métricas del riesgo, el fading de andamiaje por racha, los estadios de razonamiento. Principios de aprendizaje convertidos en mecánicas medibles, no en eslóganes.

CALIDAD INSTITUCIONALIZADA

Gates que fallan la build, doble filtro humano + validación de evidencia, contratos de datos en CI. La calidad como propiedad del sistema, no como revisión manual que se erosiona con el ritmo.

HONESTIDAD DE DISEÑO

Elegir lo defendible sobre lo vistoso: un scoring honesto sobre un evaluador semántico impresionante pero opaco; un stub fiel a sus contratos antes que una integración fingida.

03 LO QUE SE MIDIÓ, Y LO QUE NO

EL ENCUADRE HONESTO DEL IMPACTO

Cada pieza llegó a producción y pasó validación: peer-review matemático e instruccional, gates automáticos, baterías de pruebas y contratos de datos. **Eso sí se midió — la calidad del producto.**

> *Medir el efecto sostenido en el aprendizaje a escala habría exigido un rollout y una ventana de evaluación que no controlé. No hay, por tanto, métrica de outcome de alumno; y fingirla sería justo la deshonestidad que el resto del trabajo evita.*

Lo que este book sí puede defender, bajo cualquier repregunta, es el criterio: cómo se decidió qué medir, cómo se tradujo la teoría a software y cómo se sostuvo la calidad. Es lo que queda cuando se apagan los servidores — y es, exactamente, lo que estos productos fueron diseñados para demostrar.

04 CODA

HECHO EN SEIS MESES, CON LA IA COMO PALANCA

Todo se construyó en una ventana de seis meses con un modelo de desarrollo AI-first: la IA aceleró el *cómo construirlo*; la decisión de **qué medir y por qué** fue siempre propia. El portafolio es la evidencia de esa división del trabajo — y de que, bien dirigida, esa palanca permite a una sola persona razonar y entregar a la escala de un equipo.