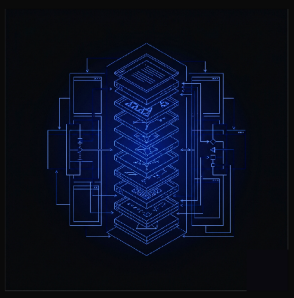
Evidence you can answer for

RESEARCH INTELLIGENCE PLATFORM



An evidence operating system that turns academic literature into traceable institutional decisions — one claim at a time, anchored to the study that supports it.

STATUS	STACK	AI	VALIDATION
Production	Next.js 16 · React 19 · Firebase	AWS Bedrock · Claude	ESSA + 26 suites

01 THE PROBLEM

CONCLUSIONS NO ONE COULD ANSWER FOR

A network of AI-accelerated personalized-learning K-12 schools makes curriculum and instruction decisions daily: which sequence to adopt, which intervention to scale, which practice to retire. Those decisions relied on AI-generated summaries that blended, in a single operation, two things that should never go together: the faithful extraction of what a study reported and the pedagogical interpretation of what that study meant for the classroom.

The result was impossible to audit. There was no way to tell what a paper's author had measured from what the model had inferred, nor to trace a recommendation back to the specific study that supported it.

The conclusions arrived as fluid, unanchored narrative: orphan sentences that sounded authoritative but that no one could verify.

On top of this was an identity problem: the previous system referenced studies by their position in an array, and every reordering silently corrupted those references.

For an institution that defines itself as evidence-based, the bottleneck was not producing text, but producing text you could answer for.

02 THE GOAL

CLAIM-LEVEL TRACEABILITY

PRIMARY GOAL

Build an evidence operating system, not a summary generator: that **every publishable conclusion had claim-level traceability**. No sentence in the final document could exist without a structured mapping to one or more source studies identified by a stable `refId`.

Alongside that, three concrete goals: irreversibly separate extraction from interpretation, apply a deterministic and reproducible methodological quality framework, and maintain a living, versioned corpus capable of incorporating new studies or re-evaluations without rebuilding entirely.

The product had to close the full cycle, from the raw paper to the documented institutional decision, so that a human reviewer could intervene at any point without breaking the chain of evidence. It is an internal admin tool for two people: the pipeline operator and a researcher from the learning science team brought in as a review peer.

03 THE SYSTEM

AN AUDITABLE NINE-STAGE PIPELINE

The system is organized as a chained nine-stage pipeline, where each stage has its own agent, its input and output contract, and a persisted, versioned artifact. The central decision, documented as **ADR-001**, was to forbid extraction, evaluation, clustering, synthesis and writing from happening in a single model call: the separation of layers is what makes the system auditable.

Ingestion→Extraction→Clustering→Synthesis→Triangulation→Claims→Chapters→Final→Editorial

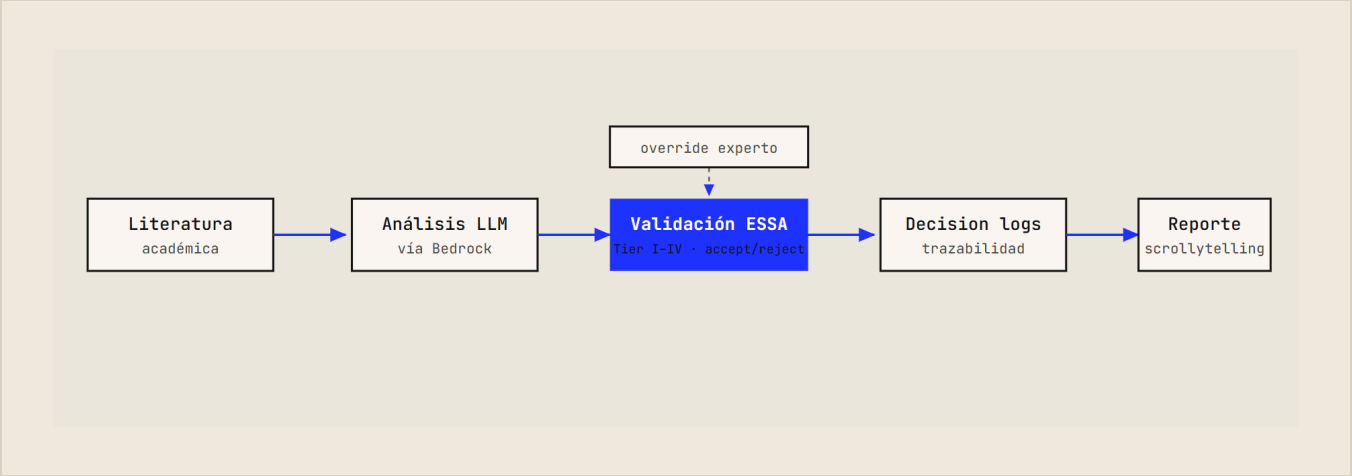


FIG. 1 — PIPELINE LITERATURE → LLM → ESSA VALIDATION → DECISION LOGS → REPORT · EACH STAGE PERSISTS A VERSIONED ARTIFACT WITH IDENTITY BY IMMUTABLE REFID, NEVER BY POSITIONAL INDEX

Separate extraction from interpretation — never in the same model call

Context. The previous system blended in one operation what the study measured and what it meant for the classroom, and referenced studies by their position in an array — every reordering corrupted the references.

Decision. The quality framework (ADR-012) forks the *appraisal* by study type —ESSA for quantitative, CASP for qualitative, both for mixed— and encodes the ESSA decision tree as explicit deterministic rules in the prompt: `randomAssignment + controlGroup + adequate sample → Strong`. The agent does not infer the tier freely.

Accepted trade-off. More stages and more artifacts than a one-step summarizer — but it is what makes every conclusion traceable back to the study that supports it. The 21 decisions were recorded as ADRs.

04 THE FOUNDATION

THREE SYNTHESIS STANDARDS, TURNED INTO SOFTWARE

The architecture translates two recognized evidence-synthesis standards into the software domain, adding a third for the qualitative. The guiding principle –**fidelity first**– comes straight from the discipline of systematic review: represent the study before interpreting it.

FRAMEWORK	WHAT IT BRINGS	HOW IT OPERATES IN THE SYSTEM
ESSA	Strength of evidence in four tiers	Deterministic decision tree encoded in the prompt; classifies every reference.
PRISMA 2020	Transparency of systematic reviews	Automatic 12-item audit against the system's artifacts at closing.
CASP	Appraisal of qualitative evidence	Forks the analysis so that quantitative, qualitative and mixed are not forced into a single logic.

The four ESSA tiers structure the entire corpus, from the strongest to the lowest empirical evidence:

TIER 1 · STRONG — RCT TIER 2 · MODERATE — quasi-experimental

TIER 3 · PROMISING — correlational

TIER 4 · RATIONALE — logic model

> *The quantitative is synthesized by effects and moderators; the qualitative by themes and mechanisms; the mixed is explicitly integrated through triangulation. Represent the study before interpreting it.*

05 THE VALIDATION

THE FUNCTIONAL CORE: HUMAN REVIEW REFERENCE BY REFERENCE

Validation happens in **/research** through an explicit human review flow. After running ESSA validation, the researcher and the operator reviewed each reference one by one, with three possible verdicts —**Accepted**, **Warning** or **Rejected**— on the tier the agent had assigned according to the deterministic tree. The system allows a human *override*: the model's judgment is a proposal, not a verdict.

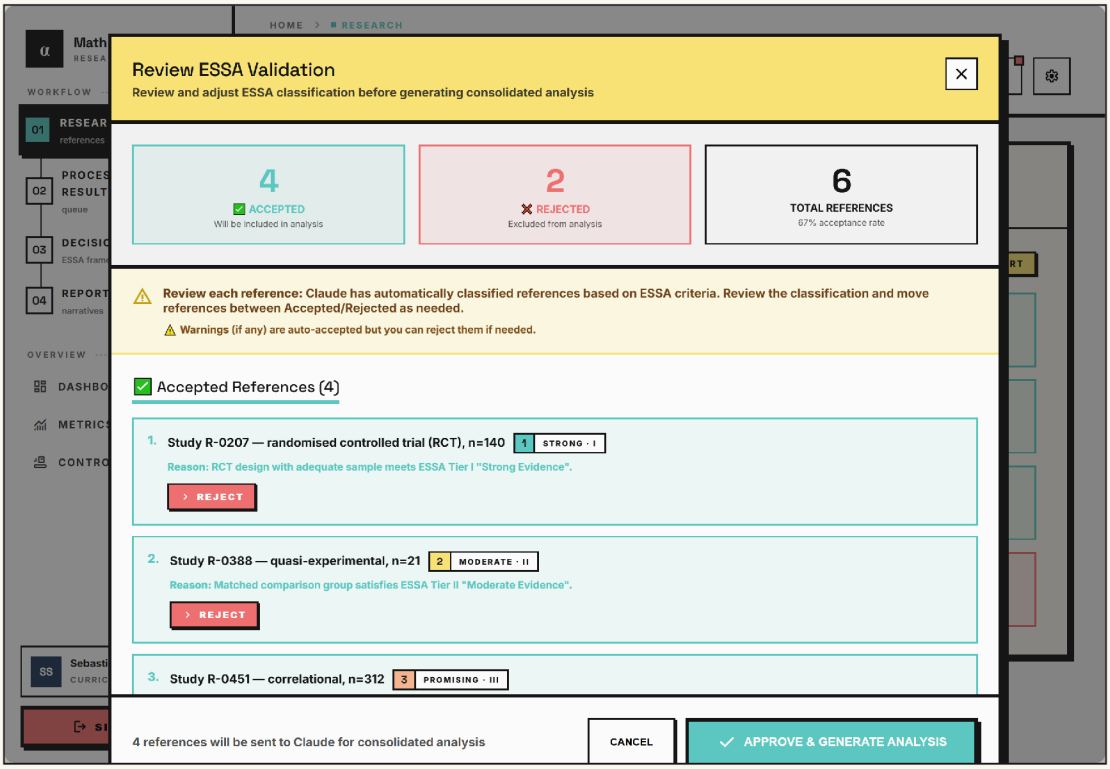


FIG. 2 — /RESEARCH WITH ESSA VALIDATION · EACH REFERENCE SHOWS ITS PROPOSED TIER AND THE ACCEPTED / WARNING / REJECTED OVERRIDE · NO SELECTION IS APPROVED WITHOUT PASSING THE HUMAN FILTER

06 THE SECOND FILTER

AUTOMATIC PRISMA AUDIT AND DECISION LOGS

The closing of the final document adds a second layer of automatic control, the **PRISMA audit**, which verifies twelve items of the standard against the system's artifacts —from eligibility criteria and per-study risk of bias to synthesis results and limitations— and reports total, partial or non-compliant adherence. Thus, no research is published without passing the double filter: human review over the evidence and automatic verification over the report. The **/decision-logs** view records each institutional decision with its chain of evidence intact.

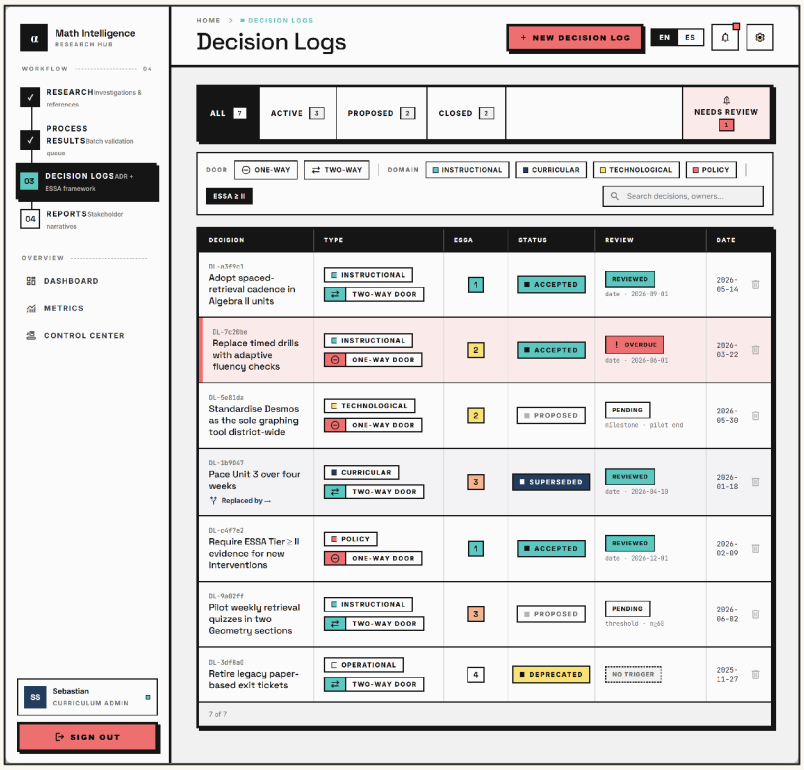


FIG. 3 — /DECISION-LOGS · EACH DECISION IS ANCHORED TO THE RESEARCH AND TO THE STUDIES THAT SUPPORT IT · TRACEABILITY FROM THE PAPER TO THE DOCUMENTED DECISION

07 FLAGSHIP CASE

MEASURING THE 2X FACTOR

Measuring the 2x Factor: Speed and Mastery Depth Metrics for Adaptive HS Math. The thesis: the claim that adaptive math platforms can double learning speed **cannot be verified or refuted** with the available evidence as it is structured today —

no study operationalizes speed and mastery depth simultaneously against an elite benchmark.

FINDINGS, WITH EFFECT SIZES

d 0.10–0.68 · g 0.37 — Adaptive platforms produce moderate positive effects on math performance, but no study measured them against elite benchmarks like 800 on SAT Math or 5 on AP.
speed ≠ depth — Advancing quickly within a platform can produce shallow understanding; speed metrics must be separated from conceptual mastery metrics.
d 0.30–1.34 — Interleaved and spaced practice produce the most robust effects on long-term retention; distributed practice doubles or triples retention versus massed practice.
depth > acceleration — Deep precalculus mastery predicts performance in college calculus with more than double the impact of simply having taken the course.
wheel-spinning — Extensive practice without achieving mastery is predictable from early indicators; time on task and number of exercises are insufficient to evaluate real learning.

08 FLAGSHIP CASE

THE PROPOSED FRAMEWORK AND ITS EVIDENCE BASE

DUAL-AXIS SPEED-MASTERY MEASUREMENT FRAMEWORK

(1) a valid speed metric distinct from time on task; (2) a mastery depth metric anchored in verified foundational competencies; and (3) a calibrated elite-performance benchmark.

The unresolved tension. The speed–depth contradiction remains open because no study operationalized both metrics at once against an elite benchmark; the effects for high-performing students under extreme pressure remain undocumented. The implication is direct: any platform that claims acceleration («2x») should empirically demonstrate the operationalization of its metrics and report effects on external assessments before declaring itself effective. It is the framework that makes the model's central promise *falsifiable*.

Evidence clusters	29
Mapped claims	45
Sources	45 · all STRONG
Report confidence	Moderate

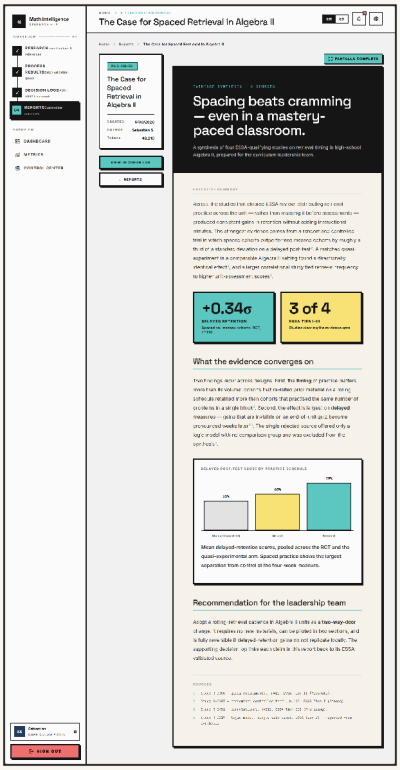


FIG. 4 — GENERATED SCROLLYTELLING REPORT · A PIECE OF RESEARCH TURNED INTO NAVIGABLE NARRATIVE, WITH EACH CLAIM ANCHORED TO ITS SOURCES · EXPORTABLE TO PDF

09 V0 GENESIS

SEVEN MANUAL STUDIES THAT JUSTIFIED THE PLATFORM

Before the platform there was a proof of concept: **seven studies synthesized by hand with Gemini**. They proved that turning literature into actionable research was valuable — and that doing it by hand neither scaled nor left an auditable trace. That v0 is what justified building the evidence operating system.

1. **Academic Audit Report.** Longitudinal audit of high-school math performance (SAT Math and AP) against reference institutions. ch. 07
2. **Cognitive Architecture of the SAT.** The 650→800 gap as structural fluency and arbitrage with Desmos. ch. 08
3. **AP/SAT Curricular Intersection.** The algebraic root system that supports advanced Calculus and Statistics. ch. 09
4. **Automation Threshold Roadmap.** The automaticity threshold as a bridge to accelerated Calculus. ch. 10
5. **MS Persistence vs SAT Stamina.** From 7th-grade persistence to the cognitive stamina of the 1550. ch. 11
6. **Elite Freshman Profiles.** Recursive audit of SAT Math compression and STEM readiness. ch. 12
7. **Technical Calculation Protocol.** Proprietary metrics for the daily math sprint. ch. 13

The seven reappear as their own chapters in this book: the manual v0 was the origin of the corpus that the platform later scaled and formalized.

10 THE CORPUS, AT A GLANCE

THE PLATFORM'S OUTPUT AT SCALE

From seven manual studies to a validated, living, versioned body of research. This is what the platform produced:

19
Research pieces
542
Studies synthesized
465
References

FIG. 5 — CORPUS AT A GLANCE · EACH FIGURE IS THE PRODUCT OF THE NINE-STAGE PIPELINE WITH A DOUBLE FILTER · THE PREVIOUS MANUAL PROTOTYPE WITH GEMINI STARTED FROM 7 STUDIES

Each of these research pieces went through the ESSA decision tree, the human review reference by reference, and the automatic twelve-item PRISMA audit. The validation interface with human override and the PRISMA audit are the two guarantees that distinguish this system from a summary generator: every conclusion that comes out is traceable back to the studies that support it. It is, in practice, the infrastructure that feeds the research corpus of the following chapters.

11 APPENDIX

CORPUS INDEX — THE 19 RESEARCH PIECES

RESEARCH	STUDIES	REFS.	DATE
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 BUILD AND RESULT

AI-FIRST, SUBJECTED TO DATA CONTRACTS

I built the platform AI-first and at high cadence, weekends included, on a deliberately modern stack: Next.js 16 with App Router, React 19 and Tailwind v4, taking on their breaking changes. The inference engine is AWS Bedrock: **Claude Sonnet 4.6** for all classification and publishing stages, and **Opus** for the interactive assistants. The discipline that sustains confidence in the output is validation with Zod at every data edge, on read and on write, against centralized contracts: an artifact that does not validate does not enter the corpus.

Each agent's runners are decoupled from the Bedrock client through an injectable interface, which makes it possible to run the full pipeline against a mock in the tests. CI runs typecheck, lint, tests and build on every push, with **26 suites** across contracts, integration, regression and unit tests. The platform shipped to production as an admin app, serving the full cycle from the paper to the documented decision.

Commits	~773 · Mar 31 → Jun 5 2026
Pipeline stages	9 · ingestion → editorial
Architecture decisions	21 ADRs recorded
Quality frameworks	ESSA + CASP + PRISMA 2020 (12 items)
Data validation	Zod on read and write, at every edge
Test suites in CI	26 · contracts, integration, regression, unit
Admin users	2 · operator + learning science researcher
Status	In production

LESSONS LEARNED

- **On architecture.** Auditability is not added later, it is designed from the first layer: separating extraction from interpretation and anchoring identity in an immutable refId resolved an entire class of bugs at the root that no later patch would have contained.
- **On AI and expert judgment.** Encoding the ESSA tree as deterministic rules inside the prompt, instead of letting the model reason out the tier freely, made the result reproducible and put judgment where it belongs: in the human override reference by reference.
- **On design.** Working the validation flow alongside a learning science researcher tuned the interface more than any written specification — watching someone walk through the review live showed exactly where traceability helped and where it got in the way. And the 21 ADRs proved their worth when operating on a stack with breaking changes.

What the portfolio demonstrates, taken together

JUDGMENT IS THE PRODUCT.

Six pieces for a single mathematics program. Seen from a distance, they are not six apps: they are six exercises of the same muscle — deciding what to measure, translating learning science into software, and sustaining quality without letting it erode.

01 THE ARC

FROM SEEING, TO TRAINING, TO GROUNDING

The portfolio spans an entire program, not an isolated feature. First **seeing**: an analytics layer that made the risk of ~1,600 students legible and turned it into morning action with no manual work Sec. 1. Then **training**: two platforms that tackle the exam's most expensive difficulty leaps —the SAT's calculator fluency and the AP's argumentation— translating learning-science distinctions into software mechanics Sec. 2. And finally **grounding**: an infrastructure that turns academic literature into traceable curriculum decisions, every claim anchored to its source Sec. 3.

Each product solved its own problem and, at the same time, enabled the next: the analytics surfaced the gap the training attacked; the need to ground those decisions drove the evidence platform.

It is not a catalog of demos: it is a system designed in layers, where the decision of what to build always came from a diagnosis, not from a feature list.

02 THE COMMON THREAD

WHAT RECURS ACROSS THE SIX PIECES

TRACEABLE JUDGMENT

Every structural decision was documented with its *why* and its accepted trade-off – from the embedded, persistent Desmos to deterministic scoring by completeness. The artifact shows the judgment, not just the result.

LEARNING SCIENCE TURNED SOFTWARE

CERC as a data type, the five risk metrics, scaffolding fading by streak, the reasoning stages. Learning principles turned into measurable mechanics, not slogans.

INSTITUTIONALIZED QUALITY

Gates that fail the build, a double human filter + evidence validation, data contracts in CI. Quality as a property of the system, not a manual review that erodes under pressure.

DESIGN HONESTY

Choosing the defensible over the flashy: honest scoring over an impressive but opaque semantic evaluator; a stub faithful to its contracts before a faked integration.

03 WHAT WAS MEASURED, AND WHAT WASN'T

THE HONEST FRAMING OF IMPACT

Every piece reached production and passed validation: mathematical and instructional peer review, automated gates, test batteries and data contracts. **That was measured — product quality.**

> *Measuring the sustained effect on learning at scale would have required a rollout and an evaluation window I did not control. There is, therefore, no student-outcome metric; and faking one would be exactly the dishonesty the rest of the work avoids.*

What this book can defend, under any line of questioning, is the judgment: how it was decided what to measure, how theory was translated into software, and how quality was sustained. It is what remains when the servers go dark — and it is, precisely, what these products were designed to demonstrate.

04 CODA

BUILT IN SIX MONTHS, WITH AI AS LEVERAGE

Everything was built within a six-month window using an AI-first development model: AI accelerated the *how to build it*; the decision of **what to measure and why** was always my own. The portfolio is the evidence of that division of labor — and that, well directed, this leverage lets a single person reason and deliver at the scale of a team.