From two-hour hand-built quotes to a validated quote in under ten minutes

The deterministic pricing engine under the quoting stack: 4,600+ SKUs across seven vendors, compiled into one catalog that catches end-of-life hardware, surfaces the real deal economics, and validates an entire bill of materials in a single go / confirm / stop call.

2 hrs → 10 min
Full quote preparation — ~92% faster

4,600+
SKUs across 7 vendors, auto-refreshed from price books

100%
Quotes with margin rules applied automatically

97 tests
Plus an independent code review

THE PROBLEM

A reseller managing roughly 1,000 router models across seven manufacturers quoted by hand, with pricing scattered across vendor spreadsheets and PDFs. Three failures compounded on every quote: discontinued products slipped in unnoticed — a re-quote at best, a credibility hit at worst; the largest profit source, one-time carrier activation compensation, was invisible at quote time, so deals were priced blind to their own economics; and margin rules varied with whoever built the quote that day.

WHAT WE BUILT

- A unified catalog of 4,600+ SKUs across 7 vendors, auto-refreshed from raw vendor price books — no hand-keyed pricing anywhere in the path.
- Lifecycle intelligence that flags end-of-life hardware before it's quoted and recommends the active replacement.
- An activation-compensation engine surfacing real deal economics — held in an internal-only view, with a negative-margin flag.
- A customer-pricing engine applying the firm's exact margin rules identically every time — three customer-facing options driven by activation tier.
- A one-call bill-of-materials check returning a clear go / confirm / stop decision.

> **The safeguard.** Pricing and lifecycle facts come from the unified catalog and deterministic rules — not memory, and not the model. An entire bill of materials is validated in one call; ambiguity escalates to a human as **confirm** instead of becoming a guess.

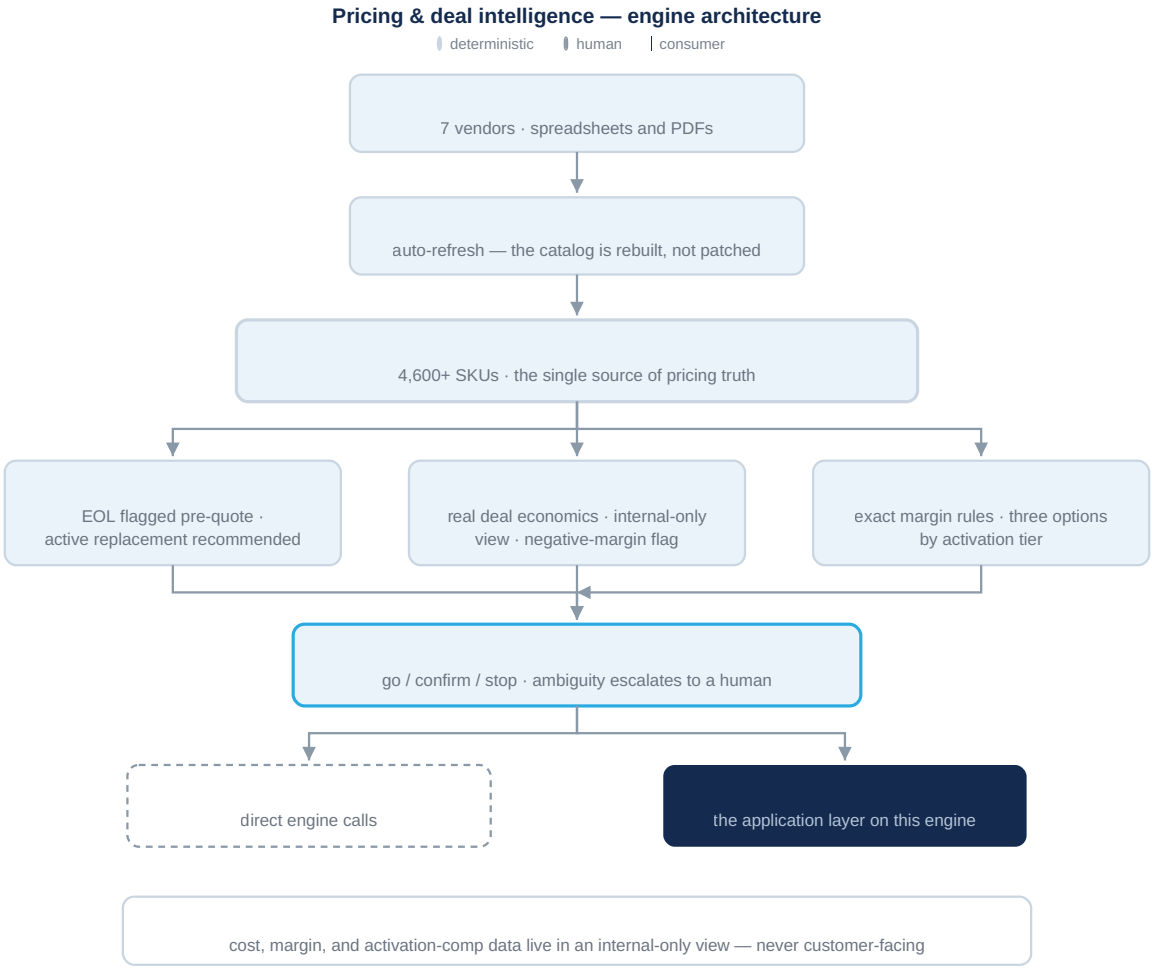
THE RESULT

Quote preparation dropped from ~2 hours to under 10 minutes, with discontinued hardware caught automatically, full deal economics visible on every quote, and margins applied consistently. Built to last: 97 automated tests and an independent code review — and the same engine now powers a conversational quoting assistant (see the Quote Bot case study).

Quoting from scattered spreadsheets where errors cost margin? **Start with an AI Workflow Teardown →**

HOW IT WORKS — ONE CATALOG, THREE ENGINES, ONE CONTRACT

Raw vendor price books are ingested and normalized into a unified catalog; three engines run on top of it, and an entire bill of materials validates in one call. Note what is absent from this diagram: a model. No LLM sets a number anywhere in the pricing path — pricing flows from the catalog through pure code.



Exact money arithmetic throughout: Decimal math with half-up rounding — no binary-float pennies.

Two design choices stand out. **The go / confirm / stop contract:** the system never guesses its way past a judgment call — ambiguity escalates as **confirm**. **One engine, many surfaces:** the same catalog and rules power direct validation and the conversational quoting assistant, with zero divergence.

RELIABILITY AND GUARDRAILS

MECHANISM	WHAT IT DOES
Deterministic pricing only	No model sets a number, anywhere in the platform — pricing comes from the catalog through pure code
Exact money arithmetic	Decimal math with half-up rounding — no binary-float pennies
Auto-refresh from source books	Pricing updates flow from vendor price books, not hand edits — the catalog is rebuilt, not patched
EOL gate, pre-quote	Discontinued hardware is flagged before a quote exists, with the active replacement recommended
go / confirm / stop contract	Ambiguity escalates to a human as **confirm** — the system never guesses its way past a judgment call
Internal-economics suppression	Cost, margin, and activation-comp data live in an internal-only view, kept out of anything customer-facing
97 automated tests + independent review	The engine's behavior is asserted, not assumed — and reviewed independently of the author

INTEGRATION STACK

LAYER	TOOL	ROLE
Engine	Python, deterministic code paths	Catalog compilation, pricing, validation
Money math	Decimal, half-up rounding	Penny-exact totals
Ingestion	Vendor price books → normalized catalog	Per-vendor mapping; a format change means a re-map, not silent corruption
Consumers	One-call validation + the quoting assistant	Engine layer serving multiple surfaces from one source of truth

THE PLATFORM CONNECTION

This engine is the data layer of a two-layer platform. The application layer — a conversational quoting assistant that turns plain-English requests into human-reviewed draft quotes in about five minutes — is documented in its own case study (Quote Bot). The metrics are scoped deliberately: this study's 2 hours → under 10 minutes covers full quote *preparation* — research, validation, pricing, and economics; the assistant's 1 hour → 5 minutes covers draft *assembly* on top of this engine. Two numbers, two workflows, one platform.

OUTCOMES & METRICS

METRIC	RESULT
Full quote preparation	~2 hours, hand-built → under 10 minutes, validated (~92% faster)
Discontinued hardware reaching quotes	Slipped in unnoticed → caught automatically, pre-quote, with the active replacement recommended
Deal economics visibility	Activation comp invisible at quote time → surfaced on every quote; negative-margin deals flagged
Margin-rule consistency	Varied by quote builder → 100%, applied identically on every quote — deterministic by construction
Engineering rigor	97 automated tests plus an independent code review

TRADEOFFS AND LIMITS

Confirm outcomes still route to a human. By design — ambiguity is escalated, not guessed. The share of bills of materials that hit **confirm** is the number to watch as the catalog and rules mature.

Ingestion is mapped per vendor format. Seven vendors, seven known formats; a vendor changing its price-book layout means a re-map, not silent corruption — a maintenance cost accepted in exchange for zero hand-keyed pricing.

The catalog covers the firm's seven-vendor portfolio. It matches what the firm actually sells; off-catalog and custom line items remain a human judgment call outside the engine.

CONSULTING TAKEAWAY

The profit wasn't missing. It was invisible. Activation compensation — the largest profit source on these deals — existed the whole time; it just wasn't visible at the only moment it mattered: when the quote was being priced. The team wasn't leaving money on the table out of carelessness. The table was in another room.

That's the real product of a pricing engine — not speed, *sight*. Deal economics at the moment of pricing. End-of-life status at the moment of selection. Margin rules at the moment of calculation, identical no matter who's calculating. The 92% time saving is real, but it's the side effect: when nothing has to be looked up twice and nothing can be applied inconsistently, speed is just what's left over. And an engine compounds — the same catalog and rules now power a conversational quoting assistant: one source of pricing truth, multiple surfaces, zero divergence.