From one-hour quotes to five-minute quotes — without ever inventing a price

An internal AI assistant that turns plain-English requests into source-backed draft quotes, with one hard rule: the model never sets the number.

1 hr → 5 min
Quote drafting time

118 + 23
Unit tests + behavioral evals, run in CI

100%
Quotes human-reviewed before a customer sees them

$0
Prices the model can set — pricing is deterministic only

THE PROBLEM

Masters Telecom's sales engineers built every quote by hand from an intake conversation — slow, and risky if a price was ever wrong or invented. They needed speed without putting a bad number in front of a customer.

WHAT WE BUILT

- A Zulip assistant that turns a plain-English request (“draft a quote for ABC Dental: 2 HaaS boxes + 8 SIP lines”) into a clean, sectioned draft quote.
- Pricing pulled only from structured catalog files through deterministic Python — never generated by the model.
- Internal cost and margin data suppressed from anything customer-facing.
- A human reviews and approves every quote before it reaches a customer.

> **The safeguard.** The model assists with understanding and wording; the price comes only from catalog files and deterministic code. The model is structurally prevented from producing a final number — and that boundary is enforced by tests, not policy.

THE RESULT

Hardened from a “B−” audit into a tested, eval-gated tool with a human in the loop. Quote drafting dropped from roughly an hour to about five minutes — and not one price originates with the model.

Have a quote-heavy workflow where a wrong number is expensive?

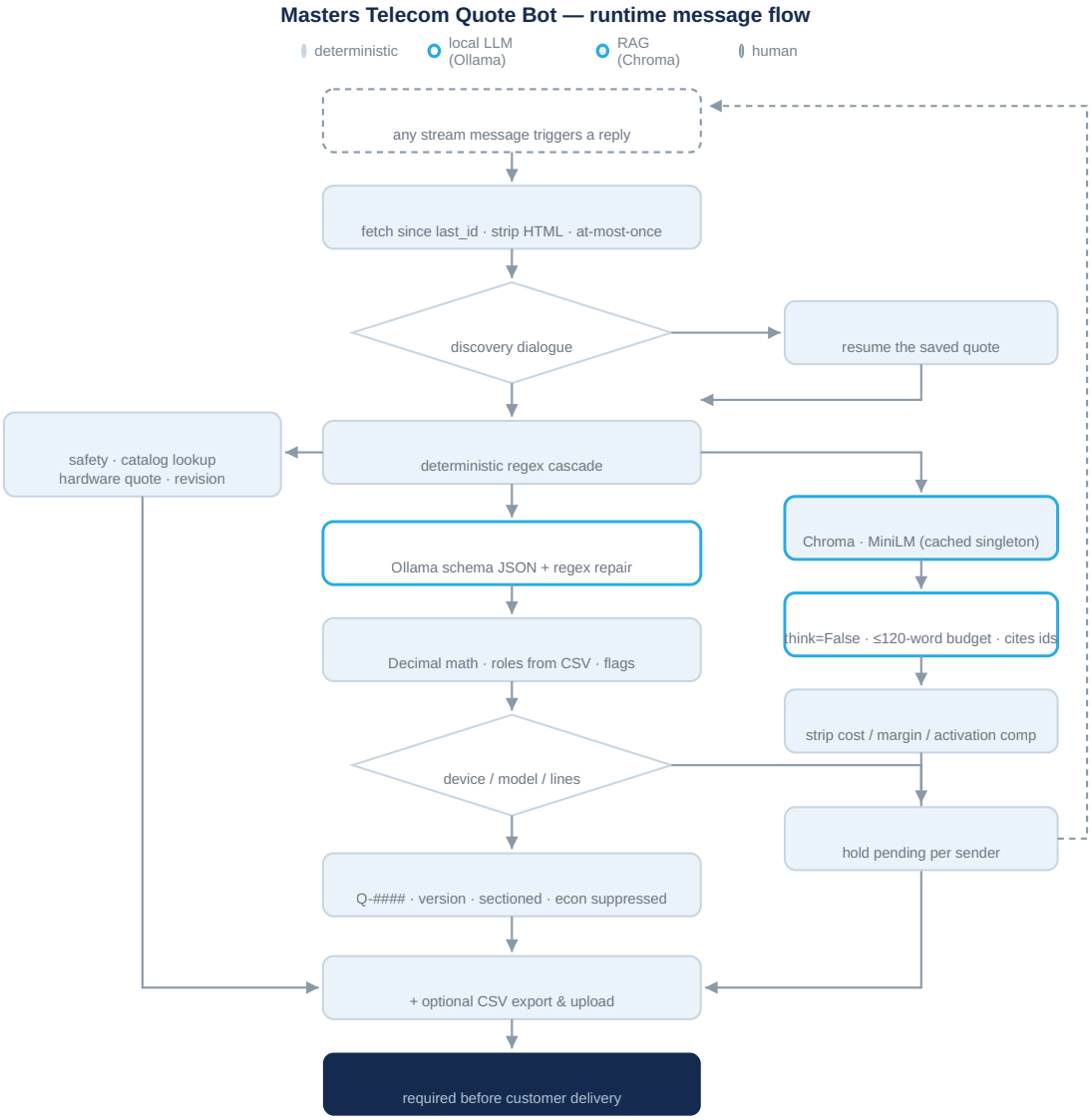
Start with an AI Workflow Teardown →

Pete Dunn · PD Insights · pete@pdinsights.ai · pdinsights.ai

Full technical detail on the following pages

HOW IT WORKS — RUNTIME FLOW

A message in the watched Zulip stream is polled, cleaned, and routed by a deterministic intent classifier. Catalog lookups and quote math run in pure Python; the local LLM is used only to extract structured fields and to answer open questions over retrieved documents. A final sanitizer strips any internal economics before the reply leaves the system.

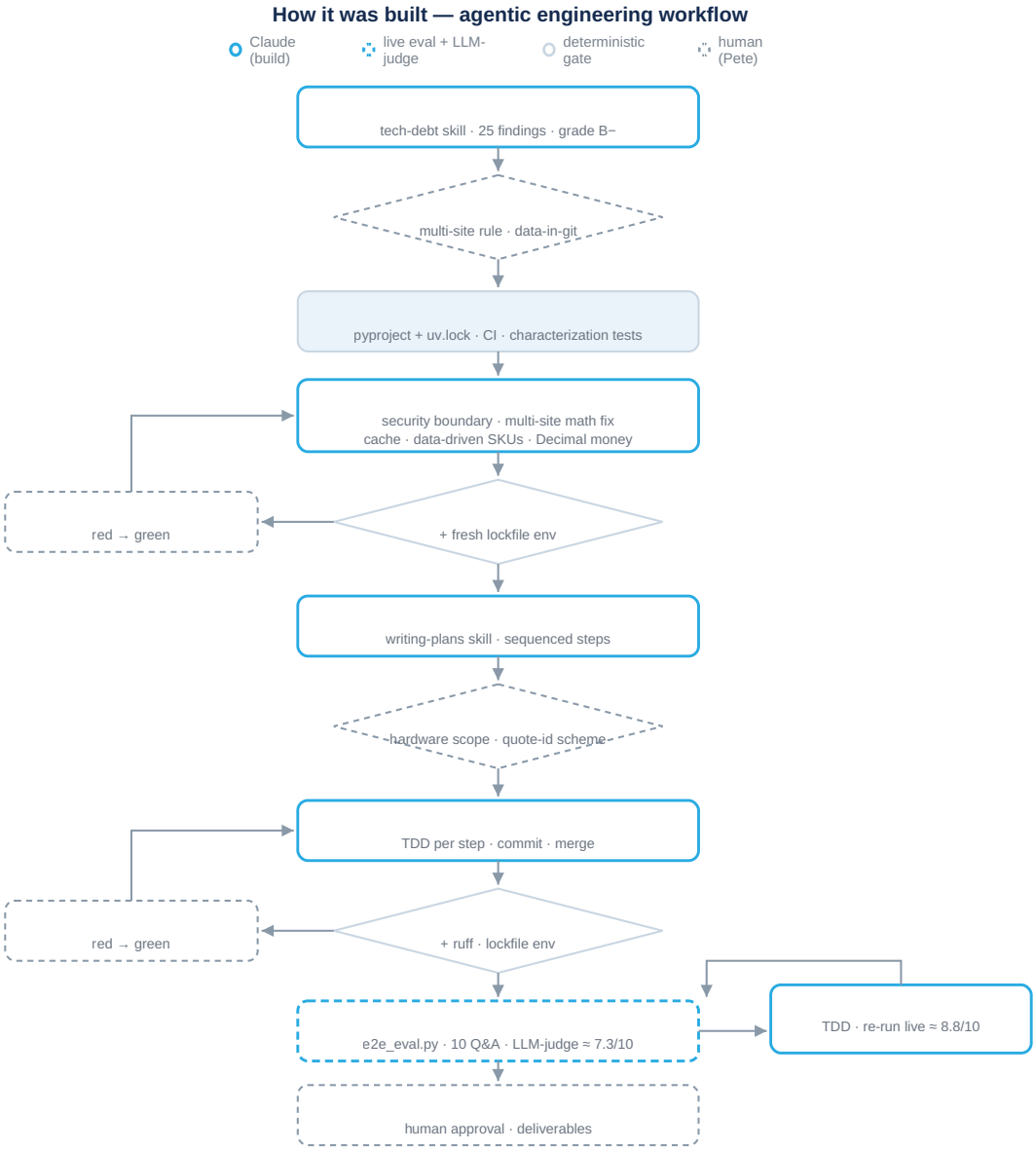


Pricing and quote math are pure deterministic Python; the local LLM only assists extraction and freeform answers, and never produces final numbers.

Two design choices stand out. **The discovery loop:** when required information is missing, the bot holds the partial quote and asks one question rather than guessing or refusing. **The suppression boundary:** customer-facing output is mechanically scrubbed of cost, margin, and activation-compensation data — and that scrubbing is asserted by tests.

HOW IT WAS BUILT — AGENTIC ENGINEERING WORKFLOW

The work ran as an agentic engineering workflow: an AI engineer (Claude, in an agent harness) performed the audit, planning, and implementation under human direction, with deterministic quality gates and an independent live evaluation that had to pass before any human merge. Every code change followed test-first discipline — a failing test, then the minimal change to pass it.



The shape mirrors a classic analyst–critic pipeline: build, gate, and an independent evaluation feeding a bounded revise loop. Humans owned the decisions code cannot settle — how multi-site quantities should be interpreted, whether sensitive data may live in the repository, and the final merge.

WHAT WAS DELIVERED

- **Hardware / FWA quoting** — routers and fixed wireless priced from the 4,900-row catalog using the real three-option, activation-tier-driven pricing model, with an internal-only deal-economics view and a negative-margin flag.
- **More quoting domains** — Verizon One Talk "Red Phone" handsets and typed fax / alarm / elevator / modem SIP lines as first-class quote items.
- **Conversational revision** — "add 2 more SIP lines," "switch to outright" edit a saved quote in place, producing a new version with a totals delta.
- **One-question discovery** — a missing math-critical fact triggers one targeted question, not a dump of gaps.
- **Quote identity** — stable Q-#### ids, versioned history, PDF-style sections; "show quote Q-1024" recalls any version.
- **Tighter answers** — answer-first replies on a ~120-word budget with inline source citations.

OUTCOMES & METRICS

METRIC	RESULT
Audit grade → state	B− → reproducible, tested, hardened
Automated tests	0 → 118 unit + 23 behavioral evals, all green in CI — incl. penny-exact "golden" quotes reproduced from real historical quotes
Quote extraction latency	~33s → ~4s per quote
Multi-site pricing	~3× silent overcharge → correct totals + review flag on ambiguity
Money arithmetic	Binary floats → exact Decimal, half-up rounding
Live model-judged quality	~7.3 → ~8.8 / 10, measured on a repeatable 10-question live evaluation
Data boundary	PDFs + cost columns in git → written residency policy + pre-commit secret guard

PROBLEMS WORTH HIGHLIGHTING

The 3× quote bug. Multi-site requests applied the total quantity to every site, so "3 sites, 6 boxes" priced 18. The fix distributes totals across sites and raises a review flag whenever the interpretation is ambiguous — caught by a test written before the fix.

The model that thought itself silent. Capping output length made the default reasoning model return empty answers — it spent its entire budget thinking. Disabling chain-of-thought for extraction and short answers fixed correctness and cut latency ~8×.

A latent ranking bug. Category names were normalized on one side of a comparison but not the other, so promotional SKUs systematically outranked real router bundles in catalog search. Found during refactoring; fixed with the search evals still green.

Evaluation-driven iteration. Ten questions, scored by an independent model judge; the five lowest-scoring behaviors became five test-first fixes, and a re-run confirmed the lift from ~7.3 to ~8.8 — a measurable, repeatable loop rather than a subjective "looks better."

KNOWN LIMITS

Documented honestly rather than hidden: conversational revision currently supports single-site quotes; hardware quotes are not yet persisted to the versioned quote store; authentication and a managed deployment remain future work (today, Zulip stream membership is the access boundary, documented as such). None of these affect the core safety guarantee — **the model never sets a price** — and each is recorded with a clear path forward.