Field truth, scored by code — an AI that can't invent a price, a part, or a pass

A mobile-first, offline field tool: cellular readings and phone-line inventories captured on a phone, scored by a deterministic on-device engine — with an AI that may narrate the recommendation, never make it.

−30%
Survey time — team-reported vs. the prior process

>50%
Survey accuracy gain — team-reported

~2,096
Automated assertions, 0 failing across 8 offline suites

$0
Prices or parts the model can introduce — quarantined on sight

THE PROBLEM

Connectivity-dependent services live or die on signal at the demarc: sell past a weak site and the hardware comes back, the deal sours, and life-safety lines end up on links that can't carry them. Yet qualification data was the industry's throwaway — the tools carriers publish deliberately store nothing, field readings lived in text messages and memory, and “false start” surveys, begun without the right tools or information, burned trips and revisits. What was needed: capture that survives, scoring that can't be sweet-talked, and a hard wall between measurement and quote.

WHAT WE BUILT

- A deterministic on-device scoring and placement engine — outcome classes, A–E confidence, a solution ladder, a life-safety signal floor. Pure code; no model in the math.
- POTS-replacement sizing and bill of materials computed per demarc — capacity math applied at each demarc, never summed across a site.
- Work-order dispatch: the engineer creates the order, the tech gets a short link + 4-digit PIN, capture is pre-seeded, and submission produces an expected-vs-found diff — life-safety lines first.
- A server-side AI relay: the engine's redacted payload goes to the model for the narrative — validated against the catalog, capped at engine confidence, quarantined on violation.
- Four survey modes in one adaptive flow, plus Starlink failover sizing and Excel template round-trip.

> **The safeguard.** Two structural walls. The installer's guided flow is capture-only — price, BOM, and AI output cannot render in it. And the AI is fail-closed: the deterministic engine runs first and is authoritative; any model reply carrying a dollar amount, an uncataloged part, or life-safety compliance assurance is quarantined on sight.

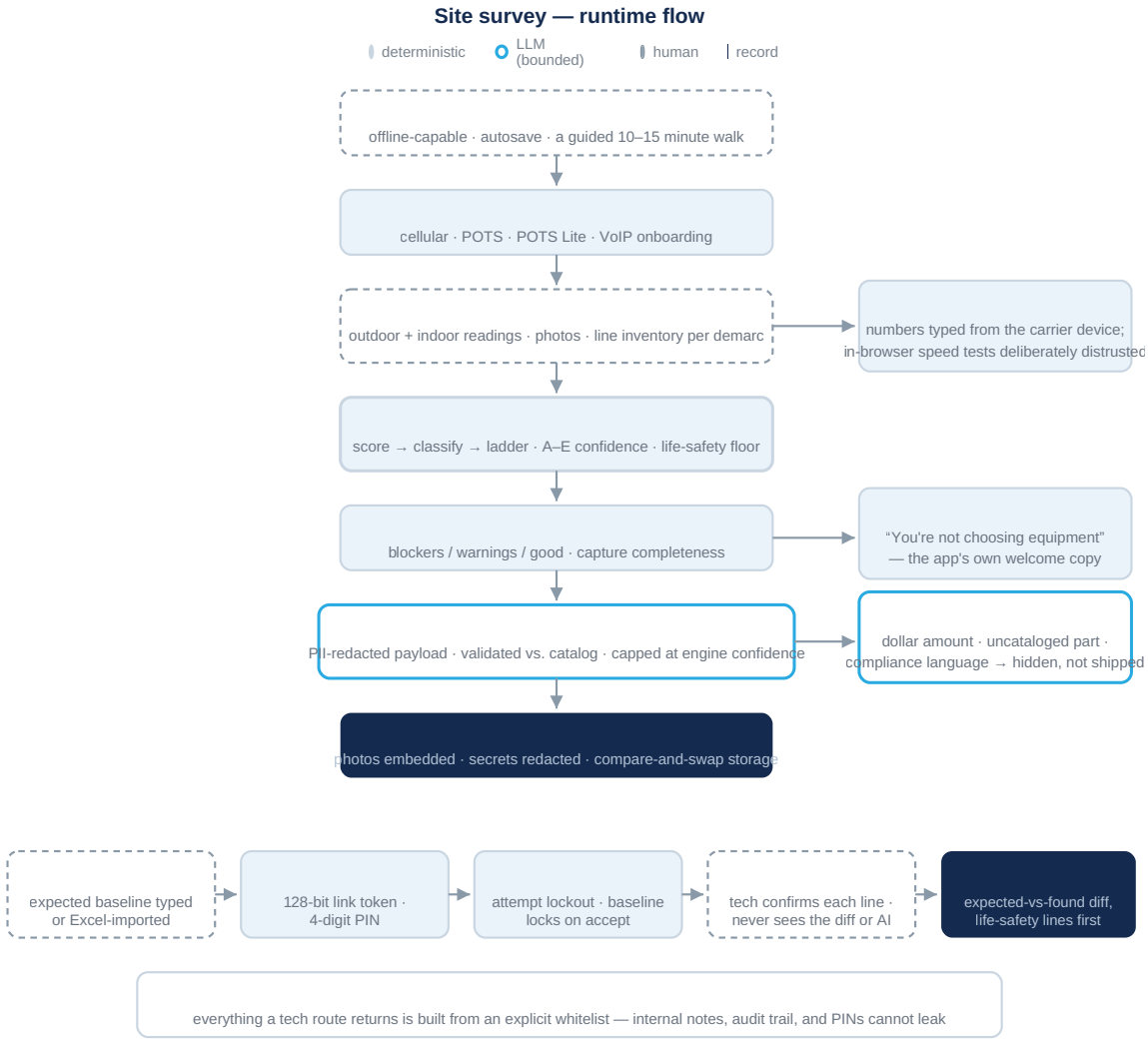
THE RESULT

The gating logic lives in auditable deterministic code behind ~2,096 assertions — dispatch live, every survey persisted. In field use, the team reports survey time down ~30%, accuracy up more than 50%, and false-start trips sharply reduced; an instrumented measurement window will harden these into in-app figures.

Does your field truth live in text messages and memory? **Start with an AI Workflow Teardown →**

HOW IT WORKS — CAPTURE, SCORE, DISPATCH

A tech opens the link on a phone; the page works offline and autosaves. The deterministic engine scores every reading before any model runs; the AI sees only a redacted payload and returns prose that must survive a validator built from the engine's own predicates. Dispatch wraps the same capture in a PIN-gated work order.



The engine decides; the model narrates. There is no LLM anywhere in the scoring, sizing, or gating math.

Two design choices stand out. **The capture-only boundary:** a field tech is never put in the position of quoting a job — enforced by the render layer, not by convention. **The mirror-image validator:** the model's reply is graded against the same predicates its instructions were built from, with one corrective retry — and if the repair grades worse, the original is kept.

AI DESIGN — THE ENGINE IS THE PRODUCT; THE MODEL IS A CONSTRAINED RENDERER

The tool never asks an LLM to make the call. An on-device engine, ported from an internal spreadsheet ruleset, computes the outcome class, a confidence *cap*, and a list of per-survey directives — deterministic, pre-computed facts and framing rules. Those directives are handed to the model as authoritative, and the model's structured reply is then run through a validator that grades the *same* predicates the directives were built from: confidence may not exceed the engine's cap; no dollar amounts; no part absent from the provided catalog; no positive life-safety compliance language. A single corrective retry feeds the validator's own messages back as a repair prompt — and if the repair grades worse, the original is kept. The instruction builder and the validator are deliberately mirror images.

The second, equally deliberate idea is the capture-only boundary: the installer's guided UI is structurally forbidden from surfacing price, bill of materials, or AI output, so a field tech is never put in the position of quoting a job — a rule enforced by the render layer and reinforced in dispatch, where a tech-facing response is built from an explicit whitelist and survey type locks on accepted jobs.

RELIABILITY AND GUARDRAILS

MECHANISM	WHAT IT DOES
Engine-first determinism	Score, classify, and solution-ladder run in pure on-device code before any model call — no LLM in the decision path
AI hard-quarantine validator	Rejects and hides model output containing a dollar amount, an uncataloged model number, a missing primary recommendation, or a pre-staged BOM on a manual-review job
Life-safety assurance guard	Clause analysis blocks any language implying code compliance, authority approval, or certified readiness — checklist data is capture evidence only
Capture-only guided gate	The installer UI cannot render price, BOM, device counts, or AI output; those exist only in explicit engineer mode
PII redactor	Strips customer phone numbers and site contacts from every model-bound payload; full data persists only in the storage archive
Fail-closed relay auth	The AI relay refuses to serve unless a configured trust check passes — deny-by-default, with rate limits and a hard spend cap behind it
Dispatch PIN + lockout	A 4-digit PIN is checked before any customer data leaves; wrong attempts advance a compare-and-swap counter, and the fifth locks the work order
Whitelist projection	Tech-facing responses are assembled from an explicit whitelist — internal notes, audit trails, and PINs structurally cannot leak into them

INTEGRATION STACK

LAYER	TOOL	ROLE
Client	Single-file HTML app (7,100 lines)	Capture UI + deterministic engine; offline; autosave to local storage, photos to on-device DB
Hosting	Cloudflare Pages, identity-gated	Serves the app and the dispatched-tech path
Backend	Cloudflare Worker (2,102 lines)	Server-side AI relay — the key never reaches the browser — plus storage, admin queue, dispatch API
Storage	Cloudflare R2	One JSON object per survey or work order; compare-and-swap on every mutation
Model	Anthropic Claude, via the relay	Recommendation narrative and customer summaries — validator-bounded, never in the math
Import	SheetJS	Excel template round-trip for baselines and inventories
CI	GitHub Actions	Eight offline suites on every push; a separate manually-triggered live model A/B gate

WHAT THE HUMAN BUILT — VS. WHAT THE AGENTS DID

- **The rules and thresholds** — the deterministic engine's scoring logic, ported from the internal device ruleset, with the life-safety signal floor and per-demarc sizing math.
- **The boundaries** — the capture-only installer wall, the fail-closed relay auth model, and the dispatch state machine with PIN-before-data.
- **The ops decisions** — a deploy script written specifically to defeat a silent-preview deploy bug; human judgment encoded as tooling.
- **Agent-executed review cycles** — the repository carries an unusually deep trail of adversarial AI audit rounds across multiple models, each round's named findings closed and folded back into the single file.
- **The honest line** — the exact human-vs-agent authorship split isn't measured, and won't be claimed without the build log.

BUILD-STATE EVIDENCE

EVIDENCE	DETAIL
Automated assertions	~2,096 passing, 0 failing across 8 offline suites — engine 1,483; worker, work-order, PDF, AI-core, and audit suites the rest (each run individually, 2026-07-10)
Footprint	1 runtime dependency; the entire app in one 7,100-line file — zero build step, fully offline
Pricing catalog drift	0 — the in-app price table verified in sync with the source catalog by a sync-check script
Dispatch subsystem	Live — merged 2026-07-10, with end-to-end and concurrent-accept smoke tests: one accept wins, concurrent attempts are rejected
Field outcomes (team-reported)	Survey time ~30% lower, survey accuracy up >50%, false-start surveys sharply reduced — reported by the operating team, July 2026; an instrumented in-app measurement window is still queued

TRADEOFFS AND LIMITS

One enormous file. 7,100 lines is hard to navigate — and it's also zero build step, fully offline, and one thing to ship and cache in a metal closet. v2 adds a module split without losing single-file deployment.

Field outcomes are team-reported, not yet instrumented. The ~30% and >50% figures come from the operating team's observation of real field use; v2 instruments duration and gate-saves in-app so every number carries its own measurement basis.

The live AI path is exercised by a manual gate, not CI. Offline suites stub the model. Acceptable because the deterministic engine is authoritative and the validator quarantines a bad reply rather than shipping it; v2 wires an automated live evaluation into CI under a spend cap.

The simplified guided flow is designed, not built. A shorter three-stop capture exists as a dated spec; installers walk the full flow today. The current flow is functionally complete and gated.

CONSULTING TAKEAWAY

The math was never the risk — 1,483 engine assertions saw to that. The bug that mattered lived in the plumbing: a mid-submit resume path that could silently blank a survey a tech had just spent an hour capturing. Caught in review, fixed with one explicit guard — the class of failure that never shows up in a demo and always shows up in a parking lot. A field tool's sync, filters, and auth have to be held to the same standard as its scoring, because the seam is where the surveys you built the tool to catch go missing.

And one bet worth naming: the qualification tools this industry publishes deliberately store nothing. This one persists every reading — because a measurement you can't retrieve is a measurement you'll pay to take twice.