INTERNAL TRAINING DOCUMENT — NOT FOR CLIENT DISTRIBUTION

Mock Case Study: Houston HVAC & Chiller Mechanical Contractor

How this agency would help an $18M, 130-person full-service mechanical company grow throughput and margin with the same workforce — five plays, the ROI math behind each one, the pitches to sell them, the questions to uncover them, and exactly how engineering would build them.

How to use this document

FOR TRAVIS (SALES / OPERATIONS)

Read the prospect snapshot, the "why this industry" section, each play's **pitch** and **discovery questions**, and the engagement shape. Everything is written in business vernacular you can say out loud in a room.

FOR THE ENGINEERING TEAM

Read each play's **how we build it** block and the reference architecture (p. 12). Stack names are standard placeholders — swap in the client's actual systems during the paid audit.

The prospect (composite, anonymized). Houston-based full-service mechanical contractor: HVAC and chiller manufacturing, installation, service, and preventive maintenance (PM) contracts. Approximately **$18M revenue, 130 internal employees plus subcontractors**. Well-defined culture and operations, strong written SOPs, leadership that invests in the company. Core ERP and industry software largely in place; little to no AI connectivity today. Open to exploring AI for ROI beyond 2026 — **explicitly without replacing or laying off the current workforce**.

What this document proves, play by play

Play	What it does	Illustrative annual value*
1. After-hours emergency intake	Never miss a 2 a.m. chiller failure again	$112K–$200K recovered revenue
2. PM contract compliance	Every contracted visit happens, on time	$60K–$110K protected renewal revenue
3. Field-to-office paperwork	Truck voice notes become invoices	$55K–$95K labor + faster cash cycle
4. Quote turnaround	Replacement quotes in 48 hours, not a week	$40K–$80K incremental gross profit
5. Knowledge assistant (RAG)	Senior-tech knowledge in every tech's pocket	$30K–$60K avoided callbacks & ramp time

*Illustrative ranges from stated assumptions inside each play (p. 4–9). Audit hypotheses to verify with the client's real numbers — never quote them as promises. Roll-up on p. 10.

This prospect scores 68–78 on your ICP qualifier — pursue, likely straight to a paid audit

Run the prospect through the same five hard gates and 100-point scorecard from the ICP qualifier checklist. Unknowns score zero — the range below reflects what the description confirms versus what call 1 still has to verify.

The five hard gates

Gate	Status	Evidence / what to verify
1. Funded pain	**Likely pass**	Leadership “values investment into the company.” Verify a named pain has a budget owner, not just interest.
2. Economic-buyer access	**To verify**	You need the owner/GM or VP Operations in the room by the audit readout — not only a service manager.
3. Real system of record	**Pass**	ERP and industry software in place. This is the single biggest predictor of build speed — AI wires into it, nothing gets ripped out.
4. Named process owner	**To verify**	The open gate. Someone must own dispatch/PM day to day and spend 2–3 hrs/week with us during the pilot.
5. Reachable data	**Likely pass**	Strong SOPs + ERP history means work orders, PM schedules, and manuals are digitized. Audit confirms export/API access.

Scorecard estimate

Block	Points	Reasoning
Firmographics & vertical fit (30)	22–25	$18M sits in the lower tier ($15–25M); 130 employees is in-band; mechanical contracting is a strong-fit vertical (p. 3).
Tech maturity & process readiness (30)	23–26	ERP present, SOPs written, documented workflows. Loses points only until API/integration access is confirmed.
Pain, budget & buying readiness (40)	23–27	Investment-minded leadership and an explicit “explore AI for ROI” mandate; budget figure not yet named, hence the deduction.
Total	**68–78**	**70+ = pursue with paid audit.** Even the low end earns a pilot-wedge. This is not a nurture-and-wait prospect.

> **What this means for deal shape.** At $18M, do not sell the full discovery → pilot → retainer motion built for $25–75M accounts. The play is a **productized wedge**: a fixed-scope paid audit, one pilot on one named process, then a modest retainer ($1.5–3K/mo). Pricing detail on p. 10.

Mechanical contracting is a near-ideal agency vertical — and this company is the right shape inside it

Four structural reasons the ROI is real here

1. The constraint is labor supply, not demand. Skilled chiller and HVAC technicians are scarce in Houston; a contractor's growth ceiling is tech hours, not lead flow. Any tool that gives techs and dispatchers more billable throughput per day attacks the actual bottleneck. This is also why the no-layoffs frame is honest: there is no surplus labor to cut. The sell is "your 130 people do the work of 160," and owners believe it because they live the shortage.

2. The work is high-value and time-critical. A down chiller in a Houston August — hospital, data center, office tower — is an emergency with a four-figure ticket and penalty clauses behind it. Response speed is directly monetizable in a way it is not in slower industries.

3. The paperwork burden is enormous and unbillable. Service notes, PM checklists, refrigerant logs, safety forms, and invoices are retyped between truck, office, and ERP. Every retyped hour is pure overhead, and the information degrades at each hand-off. Voice-and-document automation is a mature, low-risk build here.

4. Revenue is substantially recurring. PM contracts renew annually. Protecting completion and renewal rates compounds: a point of churn avoided this year is revenue kept every year after. Recurring-revenue businesses get more value from compliance automation than one-off job businesses do.

Why AI has not already solved this inside the company

- **The systems exist; the connective tissue does not.** ERP, field-service software, and phones each work. Nobody's job is wiring them together — that integration layer is precisely the agency product.
- **Off-the-shelf tools stop at the demo.** Generic answering services and chatbots do not know their PM calendar, their on-call rotation, or their ERP's work-order fields. The value is in the wiring to *their* data.
- **Tribal knowledge is retiring.** The senior tech who can diagnose a 20-year-old centrifugal chiller by sound is a single point of failure. Capturing that knowledge is a stated anxiety for every mechanical contractor owner — a buying trigger SaaS cannot manufacture.

Where this prospect sits in the portfolio

Mechanical contracting was not in the original top tier (insurance, dental/medical, CPA) — it sits with construction as a strong second tier, roughly #5 overall, and Houston is an unusually good market for it: energy, industrial, and commercial building stock, brutal cooling loads, and a deep bench of $15–75M contractors. A win here is a referenceable logo in a relationship-driven local industry.

Play 1 — After-hours emergency intake: the fastest ROI story you will ever sell

The pain

Chillers fail at 2 a.m., not 2 p.m. Today, after-hours calls typically land on an answering service that takes a message, a rotating on-call phone, or voicemail. Each path leaks: messages sit until morning, the wrong tech gets woken, and a frustrated facility manager calls the competitor whose name is second in their phone. A single missed emergency is a **$2,500–$5,000 ticket** — plus SLA penalties on PM contracts and a renewal conversation that starts with “where were you that night?”

Workflow: today vs. with the build

Today: Call → answering service / voicemail → message relayed (or not) → dispatcher triages next morning → work order created hours later.
With AI: Call answered on ring 1, 24/7 → agent identifies the caller and site against the ERP customer list → structured triage (what equipment, what alarm, water on the floor?) → severity scored against the client’s own SOP → on-call tech woken with a complete brief → work order drafted in the ERP before the tech rolls.

Why it pays

- **Recovered revenue:** every captured emergency that would have leaked is found money — the job, the pull-through parts margin, and often a new PM contract.
- **Contract protection:** response-time SLAs on PM agreements are where renewals are won and lost; documented instant response is renewal ammunition.
- **Tech morale:** on-call techs get woken only for real emergencies, with the full picture — fewer false alarms, less burnout, better retention of the scarcest resource.

ROI model (illustrative — verify in audit)

Assumption	Conservative	Moderate
After-hours emergency calls / month	25	40
Currently leaked or >1 hr delayed	15%	20%
Average emergency ticket	$2,500	$3,200
Recovered revenue / year	**~$112K**	**~$307K**

Excludes SLA penalty avoidance and new-contract pull-through. Even the conservative case pays for the pilot several times over.

Play 1 — pitch, questions, and the engineering build

SAY THIS

"When a chiller goes down at 2 a.m., what happens to that call right now — honestly? A machine can answer on the first ring, pull up their contract and equipment history while it's talking to them, wake the right tech with the full picture, and have the work order already sitting in your system. You don't miss the job, you don't burn out your on-call guys on false alarms, and nobody's job goes away — your dispatchers just stop starting every morning with an apology."

Discovery questions BIZ ENG

- BIZ "Walk me through the last after-hours emergency that went sideways. What did it cost?" *(Listen for: a named dollar figure and a named night.)*
- BIZ "How many emergency calls a month never become work orders — do you even know?"
- JOINT "Which contracts carry response-time SLAs, and have you ever paid a penalty or lost a renewal over one?"
- ENG "How does the answering service hand off today — call, text, email? Is any of it structured data?"
- ENG "Can the ERP create a work order via API, and can we look up a customer by inbound phone number?"

How engineering builds it

Stack (placeholders): telephony/voice layer [Twilio Voice or client's existing VoIP provider] → voice agent (speech-to-text + LLM + text-to-speech) → integration layer → ERP [e.g., Sage 300 CRE / Viewpoint / Foundation — placeholder] and field-service module [ServiceTitan / ServiceTrade — placeholder]. On-call roster from [scheduling tool / shared calendar].

Data flow: (1) Inbound call hits the agent; caller ID matched to ERP customer/site records. (2) Agent runs the client's written triage SOP as a scripted decision tree — equipment type, alarm codes, safety issues. (3) Severity score routes: true emergency → immediate call + SMS to on-call tech with a structured brief; non-urgent → work order queued for morning. (4) Draft work order created in ERP via API; full transcript and recording attached. (5) Every call logged with disposition for the morning huddle.

Guardrails: agent never quotes prices or promises arrival times beyond the SOP's approved windows; safety keywords (gas smell, electrical fire, flooding near panels) trigger instant human escalation; any caller can say "person" and get transferred; tech must confirm acceptance — no confirmed acceptance in 10 minutes escalates to the backup and then the service manager.

Build & rollout: 3–4 weeks. Week 1: SOP digitization + ERP sandbox integration. Week 2: voice agent + triage tree. Week 3: shadow mode (agent drafts, answering service still answers; compare). Week 4: live on the after-hours line only, business-hours rollout later if wanted.

> **Why this is the wedge.** Smallest scope, fastest measurable win, touches no one's daytime workflow — and it produces the call-log data that sells Plays 2 and 3. Lead with this one.

Play 2 — PM contract compliance: silent churn is the most expensive line item nobody tracks

The pain

Preventive maintenance contracts are the annuity under the whole business — scheduled visits, contractually owed on specific cadences. In practice, scheduling lives in a dispatcher's head, a spreadsheet, or ERP reminders someone must remember to run. Visits slip when emergencies crowd them out; a slipped visit becomes a lapsed obligation; a lapsed obligation becomes a renewal meeting where the customer asks what exactly they paid for.

Workflow: today vs. with the build

Today: Contract signed → visits manually scheduled → slippage discovered at invoicing or renewal → scramble.
With AI: Contract terms read into a compliance calendar at signing → agent proposes visit dates inside each contract's window → dispatcher approves with one click → cancellations auto-backfilled → 30/14/7-day at-risk flags → renewal pack auto-generated showing every visit completed, on time, with reports attached.

Why it pays

- **Protected recurring revenue:** PM books at a company this size plausibly run $4–7M/yr. Holding completion above 98% and cutting churn by even 2 points protects six figures of high-margin revenue annually.
- **Capacity smoothing:** backfilling cancellations inside contract windows flattens the tech calendar — more billable hours from the same crew.

ROI model (illustrative)

Assumption	Conservative	Moderate
Annual PM contract book	$4.0M	$6.0M
Churn reduction from completion + renewal packs	1.5 pts	2.5 pts
Protected revenue / year	**~$60K**	**~$150K**
Plus: dispatcher hours returned / week	4–6 hrs	6–10 hrs

SAY THIS

"Your PM contracts are the annuity under this business — but who can tell you today which visits are at risk of slipping this month? We'd watch every contract's calendar so nothing lapses quietly, and walk into every renewal with proof of every visit."

How engineering builds it

Build: (1) Contract extraction: LLM reads signed PM agreements [SharePoint/Drive — placeholder] into structured terms — sites, equipment, cadence, windows, SLA clauses — a human verifies each extraction once. (2) Scheduling engine: a rules layer (not the LLM) computes due windows and proposes dates against tech availability; dispatcher approves. (3) Watchdog: nightly job flags at-risk visits at 30/14/7 days and drafts reschedule outreach for approval. (4) Renewal pack: visit history + reports auto-assembled into a branded PDF, 60 days before expiry.

Guardrails: the agent proposes, humans dispose — no customer-facing message sends without dispatcher approval in v1; penalty-clause contracts route to the service manager. **Effort:** 4–6 weeks, mostly contract-extraction QA and calendar edge cases.

Play 3 — Field-to-office paperwork: kill the retyping, keep the people

The pain

Every service call generates a paper trail: notes, readings, parts used, photos, sign-offs. Today it moves truck → phone-typed notes → office retypes a clean report → billing re-keys an invoice. Three touches, each slower than the last, and details fall out at every hand-off — where billing disputes and unbilled "while you're here" work come from. Office staff hired to serve customers spend their days as human copy machines.

Workflow: today vs. with the build

Today: Tech scribbles notes → office retypes report → billing re-keys invoice → invoice goes out days later.
With AI: Tech dictates a 90-second voice note in the truck (readings, work done, parts, recommendations) → agent transcribes and structures it into the service-report template, cross-checks parts against the work order → office reviews and approves on screen → draft invoice generated in the ERP the same day.

Why it pays

- **Labor returned, not removed:** at 40+ calls/day, saving 25 office minutes per call returns 15+ hours a day — capacity that moves to dispatch, follow-up, and collections. Same people, higher-value work.
- **Faster cash:** same-day invoicing typically pulls days-sales-outstanding down 3–7 days; on $18M revenue each DSO day is roughly $49K of cash freed.

ROI model (illustrative)

Assumption	Conservative	Moderate
Service calls / day	35	50
Office minutes saved / call	20	30
Loaded office cost / hour	$28	$32
Labor value returned / year	**~$85K**	**~$200K**

Capacity only — DSO and recovered add-on billing are measured in the audit.

SAY THIS

"Your tech finishes a job and the real work starts — notes, reports, invoices, three people touching the same information. What if he just talks to his phone for ninety seconds in the truck, and the report and draft invoice are waiting for your office to approve? Nobody loses a job — your office team stops retyping and starts doing the customer work you hired them for."

How engineering builds it

Build: (1) Mobile capture: tech records audio in the existing field-service app (or a lightweight web link — no new app to learn); photos attach as today. (2) Pipeline: speech-to-text → LLM extraction into the service-report template (equipment, readings, work, parts, recommendations) → validation flags missing fields back to the tech *before he leaves the site*. (3) Review queue: office approves on one screen; approved report writes to ERP and spawns the draft invoice. (4) Learning loop: office edits feed the templates, so accuracy climbs monthly.

Guardrails: nothing bills without human approval; parts cross-checked against the work order with mismatches flagged, never silently fixed; regulated logs stay on existing compliant forms. **Effort:** 4–5 weeks; the risk is template variety — the audit counts report types first.

Play 4 — Quote turnaround: in replacement work, the first good quote usually wins

The pain

Replacements and retrofits are the highest-margin, most quote-intensive work a mechanical contractor does. An estimator pulls equipment history, specs a replacement, prices parts and labor, and builds a proposal. At a week-plus per quote, estimators become the bottleneck that caps growth, and the job goes to whichever competitor answered in 48 hours.

Workflow: today vs. with the build

Today: Request → estimator gathers history from 3 systems → manual takeoff & pricing → proposal in 5–10 days.
With AI: Request → agent assembles the equipment/service history pack in minutes → drafts scope and line-item estimate from pricing templates → estimator approves → proposal out in 24–48 hours.

Why it pays

- **Win rate:** speed wins emergency-adjacent replacement work — the pitch is not cheaper quotes, it is *first credible* quotes.

ROI model (illustrative)

Assumption	Conservative	Moderate
Replacement/retrofit quotes / month	12	20
Average job gross profit	$9K	$14K
Win-rate lift from 48-hr turnaround	+3 pts	+5 pts
Incremental gross profit / year	**~$39K**	**~$168K**

SAY THIS

"When a customer's unit is dead in August, how many days until your quote is in their hands — and who else is quoting while they wait? We're not replacing your estimator. We hand him the full history and a drafted proposal before he finishes his coffee, so his judgment is the only slow part left — and it isn't slow."

How engineering builds it

Build: (1) History pack: agent queries ERP service history and PM reports (Play 3's output strengthens this over time) into a one-page brief per asset. (2) Draft generation: LLM works from the client's estimate templates and rate tables [estimating tool/ERP — placeholder], producing a scoped draft with standard exclusions language. (3) Estimator workspace: side-by-side draft-vs-source view; approved proposal renders to the branded template and logs to CRM [HubSpot/Salesforce — placeholder], with win/loss tracked on a simple dashboard.

Guardrails: no quote leaves without estimator sign-off; rate tables are read-only to the agent (it proposes quantities, never rates); large quotes require a second approver per existing policy. **Effort:** 3–5 weeks once Play 3 data exists — sequence it after Play 3.

> **Discovery dependency.** This play only pencils if quote volume and current turnaround are real. If they already quote in 48 hours, drop it and sell Play 5 — the audit finds that out before anyone promises anything.

Play 5 — Knowledge assistant (RAG): the senior tech's brain, in every tech's pocket

The pain

A junior tech in front of a 20-year-old centrifugal chiller has three options: call the senior tech (who is on another roof), flip through a 400-page OEM manual, or guess. All three waste billable time — and the first quietly taxes the most valuable employee in the company. The senior tech's knowledge — this building's quirks, this model's failure patterns, the fix that worked in 2019 — exists nowhere written down.

What gets built

A question-answering assistant over the company's own documents: OEM manuals and bulletins, internal SOPs, past service reports (Play 3's archive), and PM checklists. The tech asks in plain language — "high discharge pressure on a [model], what's the check sequence?" — and gets an answer drawn from those documents **with the source cited**, plus relevant past jobs at that site. This is a RAG system: in business terms, "a search box that understands our stuff and shows its work."

Why it pays

- **Fewer callbacks:** first-time-fix rate is the silent margin killer — every return trip is unpaid. Even a 10% callback reduction on hundreds of calls a month is real money.
- **Faster ramp:** a six-month tech performs closer to a two-year tech — in this labor market, ramp speed is a growth strategy, and the knowledge stays when seniors retire.

ROI model (illustrative)

Assumption	Conservative	Moderate
Service calls / month	800	1,100
Callback rate reduction	8%	12%
Loaded cost per callback trip	$180	$220
Avoided callback cost / year	~$138K gross, pre-overlap	~$348K gross, pre-overlap

Gross figures overlap with Plays 3–4. Page 10 nets the portfolio down — do not sum the play tables raw in front of a prospect.

SAY THIS

"When your best chiller tech retires, how much of what he knows leaves with him? We can put your manuals, SOPs, and service history into something every tech can ask in the field — with the source attached, so nobody trusts a guess. Your juniors get better faster, and your senior stops being the unpaid help desk."

How engineering builds it

Build: (1) Ingestion: manuals/SOPs/reports [SharePoint or Drive — placeholder] parsed, chunked, embedded into a vector store, with equipment/site/date metadata per chunk. (2) Retrieval: hybrid keyword + semantic search filtered by the equipment on the current work order. (3) Answer layer: LLM answers only from retrieved passages, cites document + page, and says "not in your documents" rather than improvising — the anti-hallucination rule is a product requirement. (4) Capture loop: a one-tap "add to knowledge base" drafts a note when a tech solves something undocumented (senior reviews weekly) — the database compounds.

Guardrails: safety-critical procedures surface the official SOP verbatim; full audit log of questions and answers.
Effort: 5–7 weeks; document quality is the long pole — budget real time for the manual library.

The portfolio case: what the audit would be testing

The five plays overlap (the same service calls appear in several models), so the honest roll-up discounts the raw sum. Presented the way you would present it to the owner — hypotheses with assumptions showing, not promises.

Value stream	Low	High	Cash or capacity?
Play 1 — recovered emergency revenue	$112K	$200K*	Cash (revenue)
Play 2 — protected PM renewals	$60K	$110K*	Cash (retained revenue)
Play 3 — office capacity + faster billing	$55K	$95K*	Capacity + working capital
Play 4 — incremental quote wins	$39K	$80K*	Cash (gross profit)
Play 5 — avoided callbacks (net of overlap)	$30K	$60K	Cost avoidance
Portfolio total (discounted for overlap)	**~$296K**	**~$545K**	

*High ends trimmed from each play's moderate case to remove double-counting. Against that: audit $5–8K one-time, pilot $10–25K, retainer $1.5–3K/mo. **Even the low portfolio case repays year-one fees roughly 4–6x.** Mid-market buyers typically demand ROI inside 8 months — this clears that bar if any single play performs at its low end.

Engagement shape for this prospect

Step	Price / timing	What happens
1. Paid audit	$5–8K, 2–3 weeks	Engineers ride along on the five workflows, pull real call/quote/PM volumes from the ERP, and replace every illustrative number in this document with the client's. Output: ROI-ranked build plan with fixed pilot scope.
2. Pilot-wedge	$10–25K, 30–90 days	Play 1 (after-hours intake) live on the real phone line, shadow-tested first. Success metric agreed in writing before build: capture rate and response time vs. the answering-service baseline.
3. Expand + retainer	$1.5–3K/mo	Plays 2–3 next (Play 4 after 3, Play 5 last — it feeds on the others' data). Retainer covers monitoring, tuning, and the quarterly ROI review.

> **The no-layoffs commitment, said out loud early:** "Nothing in this plan reduces headcount. Every play is aimed at throughput — more jobs completed, more contracts kept, quotes out faster — with the same 130 people. In your labor market, that's not a concession; it's the whole strategy." Say it in the first meeting, before they wonder.

Discovery bank for this prospect — one 60-minute meeting covers all five plays

Bold lines are what to actually say. Tags mark who leads: BIZ Travis ENG engineering JOINT either, together.

Open with pain and why-now (0–10 min)

- BIZ "**Where does growth pinch first — tech hours, dispatch, estimators, or the office?**" *The answer ranks the plays for you.*
- BIZ "**What made AI worth a conversation now, versus two years from now?**"
- BIZ "**If we could fix exactly one workflow this quarter, which one pays you back fastest?**"

Walk one real job end to end (10–30 min)

- JOINT "**Take the last emergency call that came in after hours — walk me through it minute by minute.**" *The "one real example" rule: no hypotheticals, only the last actual instance.*
- JOINT "**Same exercise: the last PM visit that slipped. How did anyone find out?**"
- BIZ "**From job done to invoice sent — how many hands, how many days?**"
- BIZ "**How long does a replacement quote take today, and who's the bottleneck?**"

Systems, data, and gates (30–45 min)

- ENG "**What's the system of record — where do work orders, PM schedules, and invoices actually live?**" [ERP placeholder]
- ENG "**API or export access — have you ever integrated anything with it? Who's your vendor contact?**"
- ENG "**Where do SOPs and manuals live, and are they current?**" *Strong SOPs were claimed — verify with one document.*
- JOINT "**Who owns dispatch and PM scheduling day to day — and can they give us two hours a week?**" *This is hard gate #4. No owner, no pilot.*
- ENG "**After-hours telephony: who answers, and is there call recording or any log we can baseline against?**"

Money and close (45–60 min)

- BIZ "**Who signs off on a $6K audit — and who needs to see the readout?**" *Hard gate #2: economic-buyer access.*
- BIZ "**What would success have to look like in 90 days for this to be an easy yes to expand?**" Write the answer down — it becomes the pilot's success metric.
- BIZ "**Headcount question, head-on: is anyone worried this replaces people? Here's our position…**" then deliver the no-layoffs frame from p. 10 verbatim.

> **Scoring on the drive home:** named pain with a dollar figure (+), process owner identified (+), ERP access path confirmed (+), economic buyer in the readout meeting (+). Three or four of those = send the audit proposal within 48 hours — speed matters here for the same reason it does in Play 4.

Reference architecture — the pattern engineering reuses across all five plays

Placeholder stack (swap in the audit)

Layer	Placeholder	Role in the build
ERP / system of record	[Sage 300 CRE / Viewpoint Vista / Foundation]	Customers, contracts, work orders, invoicing
Field service	[ServiceTitan / ServiceTrade / ERP module]	Dispatch board, tech mobile app, visit calendar
Telephony	[Twilio / existing VoIP provider]	Play 1 voice intake, call recording, SMS
Documents	[SharePoint / Google Drive]	SOPs, manuals, contracts, proposal templates
CRM	[HubSpot / Salesforce / none]	Quote pipeline, renewal dates, win/loss log

The shape of every build

Trigger (call, schedule event, completed job, expiring contract) → **Agent layer** (LLM + client SOPs as scripts) → **Rules layer** (deterministic checks — code, not prompts) → **Human approval** (anything customer-facing or money-moving in v1) → **Write to system of record** (API into ERP — never a side database) → **Log everything** (for the ROI review).

Standing guardrails (all plays)

- **Human-in-the-loop by default.** Agents draft and propose; people approve customer-facing and financial actions until the client explicitly widens the leash with data in hand.
- **No side databases.** Everything lands in the ERP/field-service system the staff already opens. Adoption dies the day someone has to check a second system.
- **Citations or silence; safety in code.** The knowledge assistant answers from client documents with sources, or says it does not know; gas/electrical/flooding/lockout triggers are rule-based instant human escalations, never model judgment.
- **Audit trail.** Every agent action is logged with inputs and approver — this makes the quarterly ROI review, and renewal of our own retainer, a numbers conversation.

Sequencing logic (pitch this as discipline, not upsell)

Play 1 first (fastest proof, generates call data) → Play 2 (protects the annuity) → Play 3 (creates the structured service archive) → Play 4 (consumes that archive) → Play 5 (compounds it). Each play makes the next cheaper and smarter — and that ordering is itself a credibility asset: it tells the owner you sequence for his ROI, not your invoice size.

> **Honest limits to say out loud.** Every dollar figure in this document is an illustrative hypothesis until the audit replaces it with the client's real volumes. Integration effort depends on the actual ERP's API surface — the audit's first week answers that. And none of this works without the named process owner (gate #4); that is a client-side commitment we ask for explicitly, in writing, before the pilot starts.

Internal training — composite mock prospect; figures are stated-assumption illustrations, not promises.