

A pilot plan you can run without us in the room.

Thirty days observing, thirty tuning, thirty enforcing — structured so the part that can go wrong comes last, and so every phase produces a document somebody outside the project can read. Take it as written or cut it down; it is meant to be used, not admired.

The shape of it

Most governance pilots fail in one of two ways: they enforce on day one and break something, or they never enforce at all and end as a slide deck. The phases below exist to make the second failure visible early — if you cannot get through phase one, the problem is worth knowing about before anyone signs anything.

DAYS 0–30 **Observe**

OBSERVE mode

Sentinel evaluates every decision and enforces nothing. The caller always receives APPROVE; the real verdict is recorded. Zero enforcement risk, because there is no enforcement.

SUCCESS CRITERION

A shadow-mode report quantifying decisions that fell outside policy, with counts by rule and reason — a number the business does not currently have.

WHAT WE DO

- Deploy the gate alongside one workflow, in shadow.
- Load the template rules pack and its eval set.
- Wire the audit chain and confirm evidence exports verify.

WHAT YOU DO

- Name a policy owner. One person, accountable for the verdicts.
- Route one workflow's traffic through the gate.
- Nothing else. Enforcement is not on the table this month.

FLAG and ESCALATE are enforced; BLOCK is downgraded to FLAG and annotated. Nothing stops in production while thresholds are calibrated against your own traffic rather than our defaults.

SUCCESS CRITERION

False-positive rate on your traffic, measured — not asserted. Plus a QA sampling run with a disagreement rate your second line has actually looked at.

WHAT WE DO

- Tune thresholds against observed traffic.
- Extend the eval set with cases drawn from your real declines.
- Stand up QA sampling and the disagreement metric.

WHAT YOU DO

- Review flagged decisions weekly. This is the work.
- Rule on the false positives: which are wrong, which are policy.
- Decide whether the remaining rate is one you would defend.

The computed verdict is applied and a BLOCK stops the action before it executes — but only where the previous sixty days justify it. Promotion is gated on a checklist, not a decision to be brave.

SUCCESS CRITERION

An examiner walkthrough of one evidence bundle, end to end: the decision, the rule that fired, the citation, the hash chain, verified by a tool that imports nothing from the vendor.

WHAT WE DO

- Confirm the promotion checklist: policy owner, passing eval gate, QA sampling.
- Run the kill-switch drill and time it.
- Hand over the evidence bundle and the standalone verifier.

WHAT YOU DO

- Promote one workflow to ENFORCE. One.
- Walk an evidence bundle through your own second line.
- Decide whether this is worth expanding, on evidence.

Scope outline for an agreement

Not a contract and not legal advice — the headings a pilot agreement needs so that neither side discovers a disagreement in month three. Give it to your counsel as a starting point.

CLAUSE	WHAT IT SHOULD SAY
Term	90 days from deployment, extendable once by written agreement. Not auto-renewing into a subscription.
Scope	One named workflow, one agent type, one environment. Expansion is a new scope, not a change request.
Enforcement authority	OBSERVE for days 0–30. Any move to ADVISORY or ENFORCE requires written sign-off from the named policy owner. No inline BLOCK authority is granted by this agreement on signature.
Success criteria	The three above, stated as deliverables rather than aspirations: a shadow report, a measured false-positive rate, a verified evidence bundle.
Data handling	What leaves your environment and what does not. Retention period. Deletion on termination, with confirmation.
Exit	Either party terminates on 14 days' written notice. On termination the gate is removed and traffic returns to its prior path; audit records already written remain yours and remain exportable.
Availability	What happens when the gate is unreachable, stated explicitly: fail closed or fail open, chosen by you rather than defaulted.
Kill switch	Your operators can disable enforcement globally, by agent type, or by agent, without contacting us. Drill it during the pilot.
Commercials	Pilot fee, whether it credits against a subsequent subscription, and pricing held through the first renewal.
Publicity	Default to none. A named reference is a separate conversation after the pilot, not a term of it.

What this pilot will not tell you

Stated up front, because finding out in week ten is how pilots end badly.

Known limits

- **Ninety days is not a fair-lending study.** A pilot can show that controls fire and that evidence verifies. It cannot establish disparate impact, which needs volume and a time series neither side will have.
- **Injection detection is pattern matching.** We publish precision of 1.00 and recall of 0.03 against a public corpus. It catches known attack shapes; a novel phrasing passes the stage. The rule engine behind it is the control that carries the weight.
- **One workflow does not generalise.** A clean pilot on collections outreach says little about a credit decisioning deployment, where the scrutiny and the cost of a wrong BLOCK are both higher.
- **We are early.** Three design partners, not three hundred customers. If you need a vendor with a decade of references, that is a legitimate reason to wait, and we would rather you said so now.

Start the conversation

We are taking three design partners on bespoke terms, with pricing locked through the first renewal. If a workflow in your book fits the shape above, that is the route.