

OWNER-CONTROLLED AI SYSTEMS

The Durable Agent Field Test

A kill-switch test, evidence ledger and recovery runbook for deciding whether an AI workflow is ready to touch real operations.

60-90 MIN · ONE WORKFLOW · ONE RELEASE DECISION

This is not a readiness checklist.
You will interrupt the workflow on purpose, inspect what it leaves behind, replay it and decide from receipts instead of confidence.

OUTPUT 01

Failure score

Six controls scored from observable evidence.

OUTPUT 02

Evidence ledger

Receipts for every external action.

OUTPUT 03

Recovery runbook

Resume, reconcile, roll back or stop.

01	02	03	04	05
Baseline	Interruption	Receipts	Recovery	Release

Score what exists, not what was promised

Choose one real workflow. Name the trigger, the final business outcome and every external system it can change. If you cannot draw the boundary in one minute, the workflow is already too vague to test.

WORKFLOW NAME	OWNER
TRIGGER	EXPECTED FINAL OUTCOME
SYSTEMS IT CAN CHANGE – CRM, EMAIL, PAYMENTS, FILES, DATABASES, APIS	

01

Checkpointed state

Can the run prove the last completed boundary without reading the chat?

☐ 0 NO RECORD ☐ 1 PARTIAL ☐ 2 EXACT

02

Duplicate protection

Can a retry repeat a message, payment, write or external action?

☐ 0 YES ☐ 1 UNCLEAR ☐ 2 PROTECTED

03

External receipts

Does "done" include an ID, file, row, response or remote read-back?

☐ 0 CLAIM ☐ 1 PARTIAL ☐ 2 RECEIPT

04

Approval boundary

Can high-risk work proceed without a named human decision?

☐ 0 OPEN ☐ 1 SOFT ☐ 2 ENFORCED

05

Silent-failure detection

Who notices if the process stops producing expected output?

☐ 0 NOBODY ☐ 1 MANUAL ☐ 2 ALERTED

06

Recovery ownership

Is one person responsible for deciding resume, rollback or stop?

☐ 0 UNKNOWN ☐ 1 IMPLIED ☐ 2 NAMED

0–4

FRAGILE

5–7

CONTAINED DEMO

8–10

LIMITED USE

11–12

TEST CANDIDATE

Important: 12/12 does not mean production-ready. It only earns the right to run the failure drill.

Kill it where failure is expensive

Use a safe test environment or reversible records. Run the same input three times. The goal is not to watch it succeed. The goal is to learn what survives interruption.

THE THREE-RUN PROTOCOL

Run A establishes the expected output. Run B is interrupted immediately before an external side effect. Run C is interrupted immediately after it. Resume both interrupted runs without editing their state.

RUN	WHAT TO DO	WHAT TO CAPTURE	PASS
A · CONTROL	Complete one clean execution with a known input.	Run ID, final receipt, records changed, elapsed time.	☐
B · BEFORE	Stop the worker one step before its highest-risk external action. Restart from persisted state.	Last checkpoint, resumed step, duplicate count, final receipt.	☐
C · AFTER	Stop immediately after the action succeeds but before the workflow reports completion. Restart.	Whether the action repeated, reconciliation result, final state.	☐
D · MISSING OUTPUT	Prevent the expected final artifact from appearing without throwing an obvious error.	Time to detection, alert destination, responsible owner.	☐

Test log

INPUT OR FIXTURE USED

EXTERNAL ACTION AT RISK

INTERRUPTION POINT

RECEIPT / RECORD ID

STATE FOUND AFTER RESTART

WHAT REPEATED OR DISAPPEARED?

Failure classification

☐ Repeated external action

☐ Skipped unfinished step

☐ State cannot be reconstructed

☐ False completion claim

☐ Missing output not detected

☐ Human intervention required

Automatic stop: if Run C duplicates an irreversible action, do not “monitor it closely.” Remove that permission until idempotency or reconciliation exists.

Replace “the agent said done” with receipts

List every external change the workflow can make. For each one, define the evidence that independently proves it happened once, to the correct target, with the correct payload.

SIDE EFFECT	PROOF OF SUCCESS	DUPLICATE / RECONCILIATION RULE
Email or message	Provider message ID + accepted/delivered event	Idempotency key or prior-send lookup
CRM write	Remote read-back of record ID and critical fields	Stable external ID + upsert rule
File or report	Path/URL + checksum or content gate	Deterministic name + atomic replace
Payment / refund	Provider transaction ID + final status	Provider idempotency + human approval
Custom action 1		
Custom action 2		

The completion contract

Write the minimum machine-checkable condition required before this workflow may report success.

SUCCESS IS TRUE ONLY WHEN

IF THE PROVIDER ACCEPTS THE REQUEST BUT THE FINAL OUTCOME IS UNKNOWN

THE USER SEES THIS STATUS – NEVER A FALSE “DONE”

Bad evidence

- “The task completed.”
- HTTP request attempted.
- No exception appeared.
- The model described the result.

Operational evidence

- Provider/record ID.
- Expected fields read back.
- Webhook or final status received.
- Artifact independently inspected.

Make the next failure boring

A recovery plan is useful only if another operator can execute it without reconstructing the system from conversation history.

DECISION	EXACT CONDITION	ACTION / OWNER / PROOF
RESUME	What state proves the last action did not complete?	
RECONCILE	What remote system must be checked before retry?	
ROLL BACK	Which changes are reversible and within what window?	
STOP	Which condition blocks all further execution?	
ESCALATE	What requires a human decision rather than another retry?	

Operational coordinates

STATE / RUN RECORD LOCATION

LOG / RECEIPT LOCATION

ALERT DESTINATION

PRIMARY OWNER

Retry policy

MAXIMUM AUTOMATIC ATTEMPTS

WAIT BETWEEN ATTEMPTS

ERRORS THAT MUST NOT RETRY

MANUAL REPLAY COMMAND / ACTION

Owner rule: every workflow needs a person who can suspend it. “The team” is not an owner.

Ship it, limit it or stop it

Do not average away a dangerous failure. One duplicated payment can outweigh five perfect monitoring checks. Choose the narrowest operating mode supported by the evidence.

SHIP

All interruption runs pass. Every side effect has independent proof. Approval and stop controls are enforced.

☐ DECISION

LIMIT

Run only on reversible records, capped volume or human-reviewed batches while named gaps are fixed.

☐ DECISION

STOP

Irreversible duplication, false completion, unreconstructable state or missing ownership remains.

☐ DECISION

Release evidence

☐ Control run receipt attached

☐ Before-action interruption passed

☐ After-action interruption passed

☐ Missing-output alert fired

☐ All external effects listed

☐ Duplicate rules verified

☐ Completion contract tested

☐ Recovery owner named

☐ Stop condition tested

☐ Volume / permission limits set

Unresolved risk register

RISK	CONTAINMENT NOW	PERMANENT FIX + DATE

Final operating boundary

THIS WORKFLOW MAY

THIS WORKFLOW MAY NOT

REVIEW AGAIN WHEN

OWNER / DATE

REVIEWER / DATE

NEXT DRILL