ARTIFICIAL INTELLIGENCE AND AUTOMATION

Jarvis and RIALMS

The agent platform we run our own company on: a named fleet, routing across hosted and local models, outbound mail nobody can send without approval, and a memory that can tell you what it believed last Tuesday.

47	1.69.27	0	5
Named agent personas	Version in use internally	External messages sent unapproved	Failures before quarantine, not a loop

WHERE AI PROGRAMS ACTUALLY FAIL

Three questions decide whether an AI program survives contact with a business: what the system remembers, what it is allowed to do on its own, and what happens when a model is down. A demo answers none of them and still looks good, which is the trap.

Jarvis is our answer, and it is not a product for sale. It is how Robles Industries runs itself, which is the only honest test of an operations platform. What it demonstrates is the capability, and the capability transfers.

THE FLEET

An orchestrator with a roster of named agents, each with its own persona, role and voice, covering general management, operations, quality assurance, systems, correspondence and executive support. Agents take instructions, call tools, write to memory, raise issues and hand work to each other.

A live console shows who is active, who has failed, how much each has done in the last day and which model backed the last call. That console is the difference between an agent fleet and a folder of scripts nobody is watching.

ROUTING, SPEND AND THE FAILURE PATH

- **Routing is task-aware and availability-aware.** Interactive work prefers one class of model, background and fleet work stays on cheaper local models and escalates only when the work needs it, and heavy document, design or code jobs go to a frontier worker.
- **A backend with no credential or no health pass is skipped,** not retried into a wall. An hourly cap and a circuit breaker bound the spend.
- **Reliability is policy, not string matching.** Every backend failure is classified once into retryable, should-fall-back and should-mark-exhausted, and the fallback logic branches on those booleans rather than on the text of an error message.
- **Loops are bounded.** Per-turn guardrails warn and then block repeated identical failing tool calls, and a turn ends with one final tool-less call so it always produces an answer.

THE APPROVAL GATE

An agent can compose an email. It lands as a draft in an approval queue, and nothing auto-sends to anyone outside the company. Uncertain send receipts are reconciled from read-only provider evidence, and a message whose outcome is uncertain is never resent.

RIALMS, THE MEMORY UNDERNEATH

An append-only event log, gapless and ordered per memory realm, with corrections, retractions, irreversible redaction, idempotency keys and daily quotas. Screening runs before a write and fails closed. Everything derived from the log is a projection that can be rebuilt, so a bad extraction is a rebuild rather than an archaeology exercise.

- **Retrieval runs two lanes.** A semantic lane and a full-text lane, fused by rank, because the question that needs an exact string and the question that needs a meaning are both asked by the same user in the same minute.
- **A closed predicate vocabulary,** with generic predicates forbidden outright, so nobody writes relates_to and calls it knowledge.
- **Corroboration instead of overwrite.** Two agents asserting the same fact record it through an attestation table rather than one erasing the other's provenance.
- **Retraction is a tombstone,** not a delete, and point-in-time belief history renders supersession explicitly. You can ask what was true as of a date and get an honest answer.
- **An immutable extraction ledger,** so rebuilding a projection never calls a model provider twice, and an event that fails deterministic processing five times goes to quarantine rather than looping.

HOW IT IS BUILT

Python throughout, containerized, with PostgreSQL and a vector extension as the single substrate for the event log and every projection. Embeddings run in-process.

A small stateless gateway is the only authenticated write door. It authenticates the calling agent, takes the trust ceiling from the token rather than the request body, screens the text, applies a salience gate, rate-limits per caller, and filters every read by the realms that caller may see. Clients hold no database connections, and when the gateway is unreachable they spool to disk and replay.

Python · PostgreSQL · Vector search · In-process embeddings · Docker · Hosted and local models · Ollama · Append-only event log

WHAT WE DO NOT PUBLISH

No infrastructure detail, host names, addresses or counts appear in any public description of this system, including this one. An internal platform that runs the company is an attack surface before it is a marketing asset.

What we will do is show it. The console, the approval queue, a belief history rendered for a date you pick, and the classification table behind the failure handling are all demonstrable in a call.

ONE RULE WORTH STEALING

A generic predicate is rejected at write time. A memory system that accepts relates_to as a fact will be unusable within a year, and by then the log is too big to clean.

WHAT WE CAN START ON

An agent that works inside your operations with a human approving anything irreversible. Retrieval and memory for an assistant that has to stay correct over months rather than minutes. Or a candid review of an AI program that is not producing.

512-900-4871 · div42.com · roblesindustries.com