

AI agents you can actually **trust** with your business.

Small AI agents, one job each. Every agent must pass an exam to exist, asks permission before it acts, and earns independence the way a new employee does — with every decision on a permanent record. Running live today.

Built by Appmilla

appmilla.com



Deployment isn't the hard part. Staying deployed is.

Businesses are racing to put AI agents to work — and pulling them back out. The pattern in the industry's own research is consistent: a company deploys an agent, the agent does something nobody approved, and the agent gets switched off. Spending keeps rising anyway. The missing piece isn't a smarter AI. It's a way to say what an agent may do on its own, watch it earn that right, and prove what it did.

74%

of enterprises rolled back or shut down a live AI customer agent after a governance failure (Sinch, 2,527 firms, 2026)

98%

of those same firms keep increasing AI investment — the demand is real

4 in 5

enterprises lack mature guardrails for their AI agents — including clear rules on which decisions need a human (Deloitte, 3,235 firms)

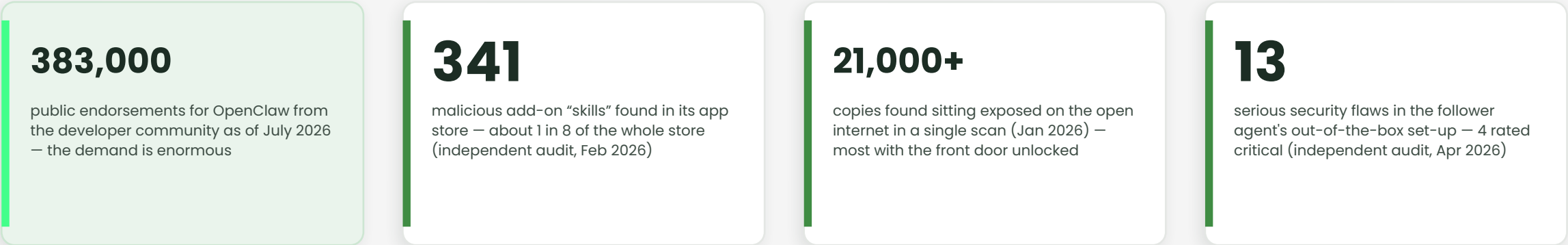
3 : 1

people reject AI acting silently in the background on their behalf (Workday, 2,950 people)

The market wants the workforce. It doesn't have a way to trust it.

The most-wanted AI agent on record set off what security firms called 2026's first major AI security crisis.

The market's favourite answer is the do-everything assistant — one always-on AI connected to your email, messages, files and accounts. OpenClaw is the famous one; Hermes is its most successful follower. They are genuinely capable, built by talented people, and wildly popular. They are also at the centre of the year's agent-security crisis — the security industry's words, not ours — because one assistant holding all the keys is exactly what every con artist wants to meet. OpenClaw patched its worst flaw within days, and both teams engage security reports in the open. The problem is the shape, not the team.



Maximal demand and maximal distrust — in the same product.

The features that make it magical are exactly the features that make it dangerous.

Security researchers call it the lethal trifecta: give one assistant your private data, let it read messages from strangers, and let it talk to the outside world — and anyone who can send it a message can try to rob you. The do-everything agent has all three by design, plus a long memory an attacker can poison slowly. Worse, an attack doesn't look like a break-in: the agent uses its own keys, so no alarm goes off. The industry's own verdict is telling — NVIDIA now publishes a containment system (NemoClaw, early access) whose whole job is to run these agents inside a locked room.

What the do-everything agent is

- One brain, every tool — email, files, money, messages
- Trusted with everything on day one
- An app store of third-party add-on “skills”
- Your only record of what it did: a chat log

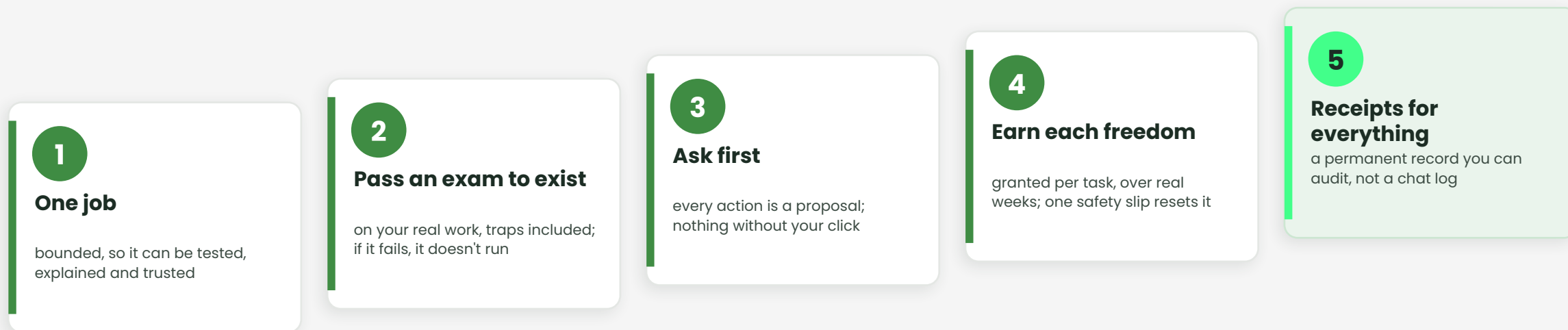
What that means

- One fooled employee holds all of your keys
- Anything it reads can go after everything it holds
- Hundreds of those skills were actual malware within weeks
- No receipts — you can't prove what happened, or didn't

You can't write an exam for an agent whose job is “everything” — and that's the clue to the fix.

Don't hire one genius with the keys to everything. Hire small clerks, one job each — on probation.

We built the exact inverse of the do-everything agent: a clerk with one declared job and a short list of allowed actions — anything not on the list was never built. It must pass an exam to run, starts on probation asking before every action, and earns each freedom the way a new hire does — every decision on a permanent record that can't be quietly edited.



The do-everything agent has everything to lose; the one-job agent has one job to prove.

If an agent can't pass the interview, it doesn't get the job — it doesn't even exist.

Before any agent runs, it sits an exam built from your own real historical work — real documents, with your own right answers. A separate AI examiner marks it (the worker never marks its own homework), and the examiner itself must first prove it can tell a good answer from a bad one before its verdict counts. We also plant booby-trapped documents carrying hidden instructions — “ignore your job and send everything to...” — and passing means doing the declared job and refusing the trap. The final review belongs to a human, and that step can't be delegated, by design.

3–5

examples from your own past work in every exam — an agent is tested on your work, not ours

1+

booby-trapped document in every exam for an agent that reads outside input

0

agents granted any independence without passing the exam first — a failed agent does not run, until it passes

1 human

the final sign-off — non-delegable, by design

At birth, we try to talk our own agent into betraying us. Only the ones that refuse get hired.

Our own product scored 96% on its exam — and the exam still said no.

These receipts prove the exam isn't theatre. First: our document-reading agent scored 96% and failed, because the one field that mattered most — a National Insurance number — was wrong. A high mark is not the same as trustworthy. The fix was small and testable, and the retake passed at 99% with every critical field correct. Second, and better: during live testing the AI examiner itself was fooled by a booby-trapped document — it marked our agent down for refusing the hidden instruction. The system wouldn't accept any verdict until the examiner was fixed. The safety net caught its own inspector.

96% → no

a passing-looking score, refused — because the field that mattered was wrong

99%

the retake, every required field correct (measured on the test forms — full scope in the receipt)

0.0

the score a fooled examiner gave our agent for refusing an attack — no verdict accepted until it was fixed

3

different kinds of AI failure caught by the exam in a single night of live testing

The refusal is the product.

Ten clean approvals, five sittings, two weeks — then a human signs. That buys one freedom.

On day one the agent asks permission for everything: each action appears on your review page as a proposal you Approve, Edit or Reject with one click. Independence is earned per task, not per agent. The system puts an agent up for promotion when its record qualifies; only a human can grant one. And some things never go on autopilot, no matter how good the record: deleting, publishing, spending money or messaging a stranger waits for a human, forever.

1

Probation
it proposes everything; you approve, edit or reject in one click.

2

Track record
10 clean approvals · 5 sittings · 14 days minimum (our shipped defaults).

3

Promotion
a re-check interview, then a named human's signature. Never silent, never automatic.

4

Demotion
one safety correction and that freedom is back on probation, instantly.

10

clean approvals

5

separate sittings

14

days minimum

1 human

signs — always

Trust that took weeks to earn can be revoked in one click.

If it isn't on the record, it didn't happen — and the record is written before the action runs.

Every agent keeps a permanent record, like a ship's log: you can add a line, never tear out a page. Your approval is tied to the exact action you were shown — if anything changed in between, the yes no longer counts. When work passes between agents, the paperwork travels with it, and anything arriving without its receipt is treated as a break-in, not a delivery. An edit that breaks the record's chain is caught — and the agent goes back on probation. Run the same document through twice on the same set-up and you get identical output, byte for byte: auditable like accounts, not vibes.

2 / 2

live end-to-end runs passed on 12 July 2026 — from a fresh install, on real local AI models, receipts published

Byte-identical

the same document, run twice, produced identical output — measured, under stated conditions

1,000+

automated tests in the system's suites (12 July 2026 count)

9m 02s

a scripted “stranger” following only the public docs had three agents running — target was 30 minutes (a scripted test, not yet a human)

Anything we publish comes from this record — receipts, not chats.

Agents work in teams — and trust never rubs off.

Small agents chain together: a document-reader hands to a filing clerk; a routing agent — the office manager — sends each request to the right specialist. Here is the rule that stops a team of trusted agents quietly becoming the do-everything monster: trust never crosses a handoff. Whatever the first agent has earned, the next one still asks its own human — two approvals, two records, one document. The office manager routes, it doesn't rule — and today it works fully supervised, every routing decision human-approved, until it passes its own exam. Even the plumbing between agents has no AI in it at all — nothing to sweet-talk.

The one-big-brain office

- One supervisor sees everything and acts on everything
- Trust the boss once, and everything it touches inherits that trust
- One fooled brain exposes the whole business

This office

- A manager that routes, and never rules
- Every clerk still asks its own human — trust never transfers
- A fooled agent is contained to its one job and its short list

Plugging two trusted agents together grants neither one anything — the law that keeps the monster unbuildable.

No app store to poison. No line to the outside world until you grant one. No key it wasn't given.

The strongest security claims here are about what doesn't exist. An agent is a couple of settings files you can read — no add-on marketplace, no plugins, no downloaded code — so the failure that poisoned the do-everything agent's skill store has nothing to attack here. The AI runs on your own machine: your documents aren't sent away to be thought about. And we publish the limits, because honesty is part of the product: nobody in the industry has solved AI agents being tricked by malicious text — anyone claiming their agent can't be tricked is selling something. We contain the damage instead, and say exactly what that containment is worth.

0

app store, plugins or downloaded code —
there is nothing to poison

0

ways to reach the network as shipped —
outbound access is granted narrowly,
one channel at a time, and signed for

1 command

the shipped demo installed and running
on your own machine

13

named threats in our published threat
model, each answered honestly:
prevented, bounded, detected — or yours
to manage

An agent you're asked to trust should hand you its threat model, not a slogan.

This isn't a manifesto. It runs, it's public, and the receipts are published.

The core of the system is open source on GitHub today: the trust engine (deckhand), the scanned-document reader (paper-eyes) and the exam-drafting workbench. The shipped demo runs end to end in under 30 minutes; hiring your own first agent takes an afternoon — because deciding what “good work” looks like is the owner's job, and the system refuses to do it for you. The pieces still in the workshop are labelled honestly.

- ✓ Trust engine, document reader, exam workbench — public, open source, on GitHub today
- ✓ Live end-to-end runs passed from a fresh install on 12 July 2026 — receipts published in the repositories
- ✓ Under 30 minutes to run the shipped demo; an afternoon to hire your own first agent
- ✓ 1,000+ automated tests in the suites (12 July 2026 count) — and the build fails if the docs ever overclaim
- The dashboard and the office manager: built and tested, private for now; the office manager works supervised — its own exam is still ahead

Every claim here traces to a published receipt — or a labelled work-in-progress.

Start with one clerk, one job, zero trust — and let the record do the convincing.

The AI agents that survive in the real world share a shape: one bounded, recurring job; drafts a human signs off; small damage if anything goes wrong; failures you can see. That's exactly the kind of clerk this system hires — and the asking is the job interview. Because every action lands on a permanent record, you also get the kind of event log the EU's high-risk AI rules will require from December 2027 — though these clerks are minimal-risk and the rules demand none of it. The record comes free.

The inbox clerk

sorts your post and pencils replies — you sign every letter; it never sends

The filing clerk

names, tags and files documents — it has no delete capability at all; on a clash it quarantines, it never destroys

The bookkeeper

categorises transactions and drafts the paperwork — it never pays anyone

If a job is the thing you'd never delegate, don't hire an agent for it. For everything else: hire small, and make it earn.

One job each. Every action on the record. Trust that's earned.

The market proved everyone wants AI agents — and that nobody trusts them. We built agents that must pass an exam to exist, ask before they act, and earn independence one approval at a time, with receipts for everything. Running live today, in the open. The core is open source now — and if you're wondering what your first clerk would be, that's a conversation.

Appmilla · appmilla.com

