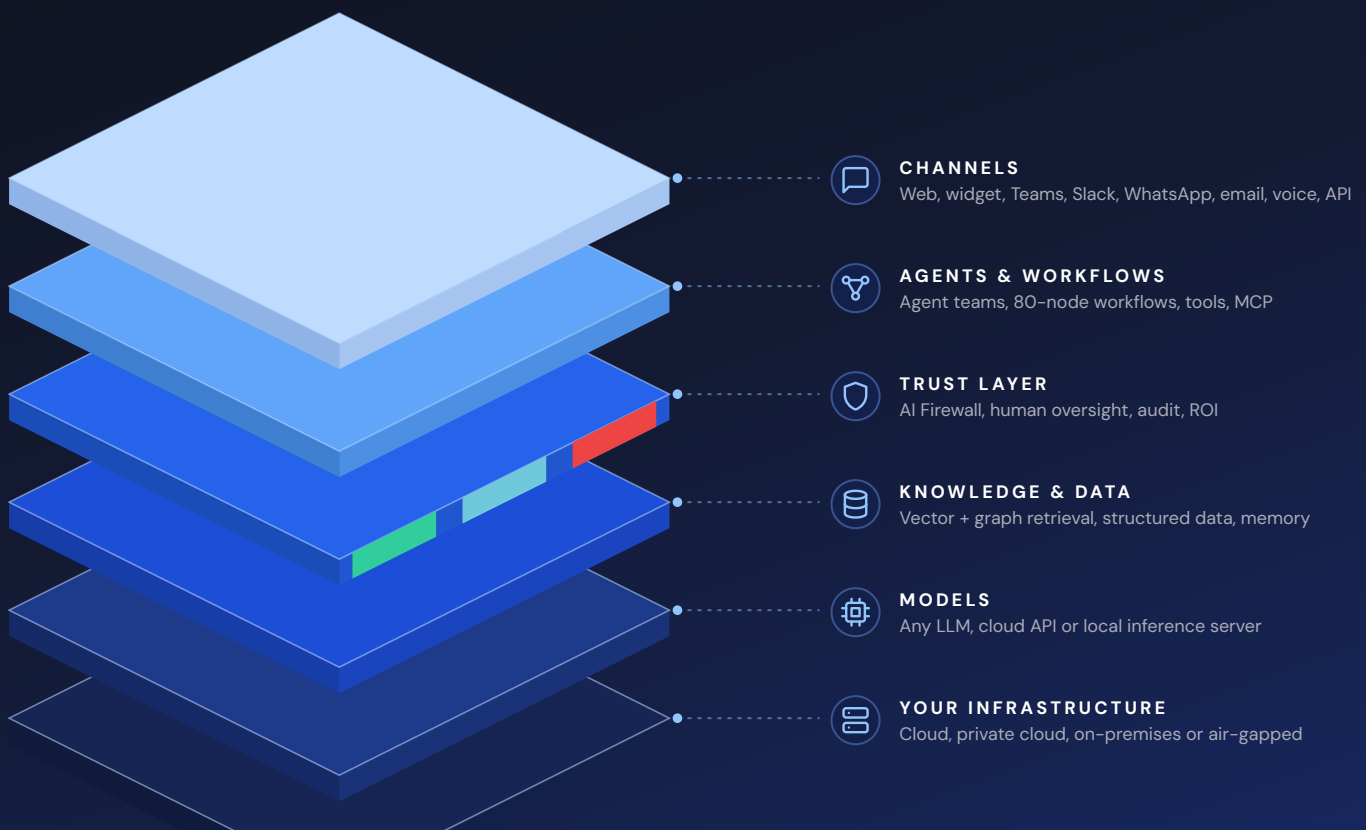


Technical Overview.

How Evolvable.ai works.

Architecture, security controls, integrations and deployment for the **sovereign AI agent platform**. One install on your infrastructure runs your agents, your workflows and the controls around them.


100⁺

integrations, native and MCP

60⁺

AI Firewall detectors

80

workflow node types

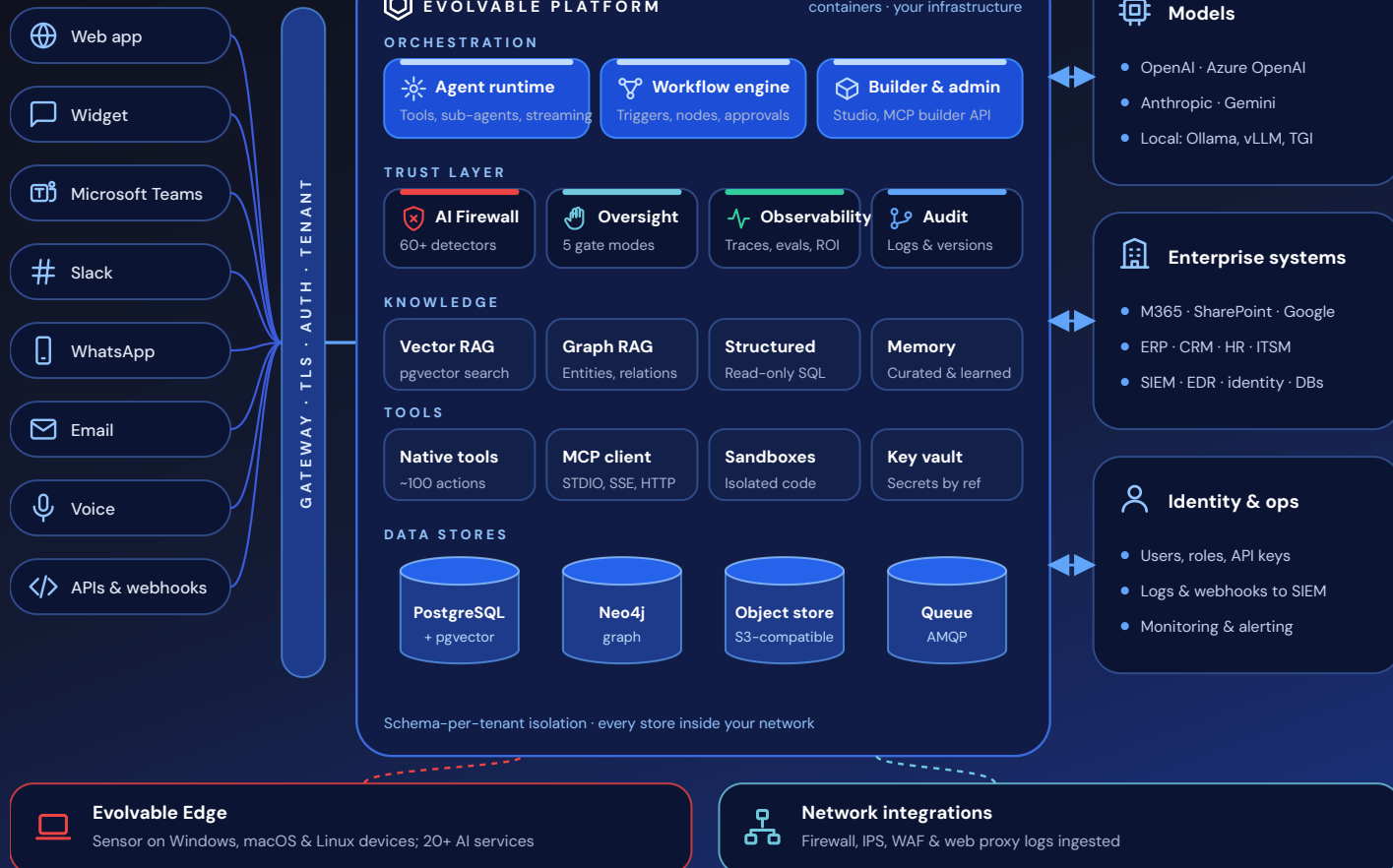
4

ways to deploy, incl. air-gapped

01 Reference architecture.

A containerised platform inside your estate. Every request enters through one gateway, runs through the trust layer and touches only the systems and models you connect.

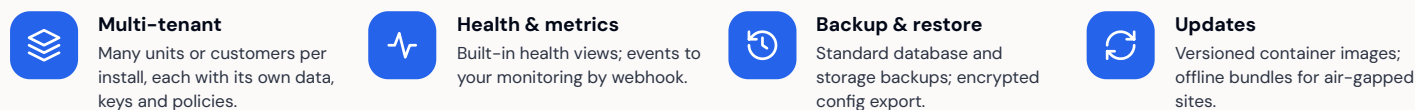
PEOPLE & CHANNELS



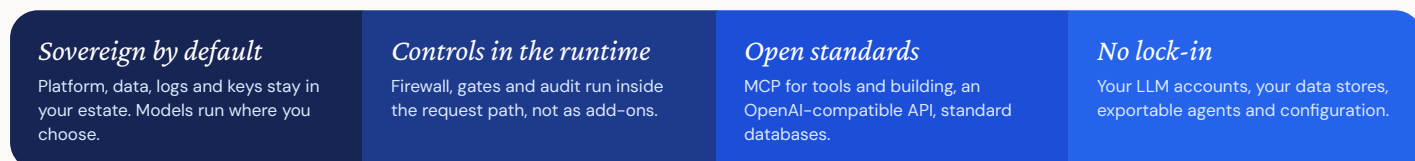
BUILT ON PROVEN OPEN TECHNOLOGY



OPERATING THE PLATFORM

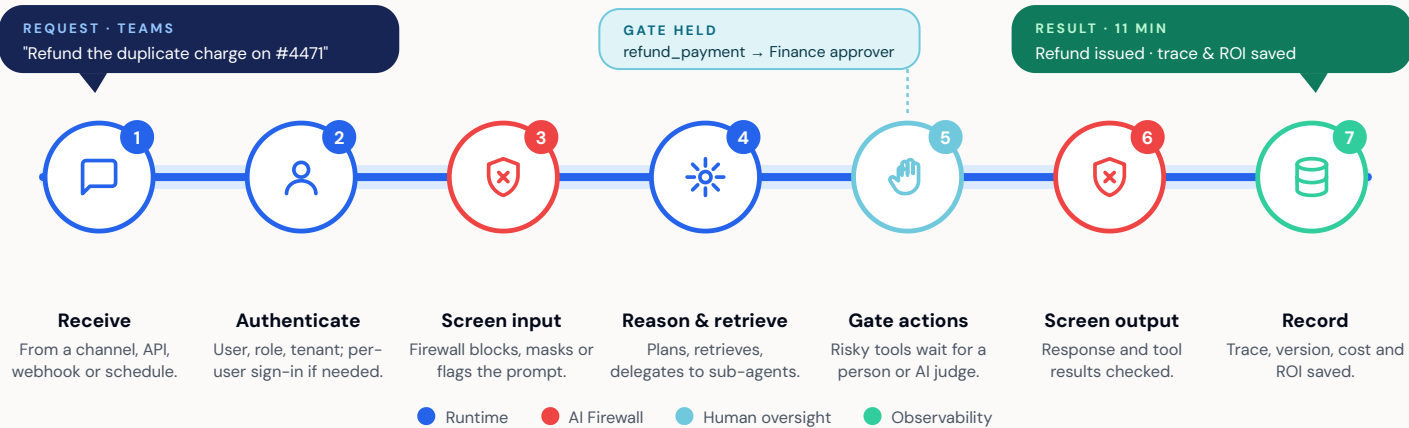


DESIGN PRINCIPLES



02 The runtime: *life of a request.*

Every request follows the same seven steps. The controls are part of the runtime, not a proxy in front of it, so nothing reaches a model, a tool or a user without passing them.



THREE ENGINES, ONE RUNTIME

Agent teams

- Sub-agents, agent-to-agent calls and parallel fan-out
- Skills, JSON schemas, model per agent
- Versions, diff and restore; evals before go-live

Workflows

- ~80 node types incl. agents, MCP, SQL, Python
- Triggers: webhook, schedule, Slack, mailbox
- Durable approval and decision nodes

Dual RAG & memory

- pgvector semantic + Neo4j graph retrieval
- Read-only text-to-SQL on structured tables
- Curated, learned and scoped memory

INTERFACES FOR YOUR DEVELOPERS

Chat & task APIs
Sync, streaming (SSE) and async tasks

OpenAI-compatible
Drop-in for existing OpenAI SDKs

Builder MCP API
Build agents from any MCP client

Signed webhooks
HMAC-signed, with retry & replay

ANATOMY OF AN AGENT · CONFIGURED, NOT CODED

Identity
Instructions, model, behaviour, output schema

Knowledge
Documents, databases and memory it may use

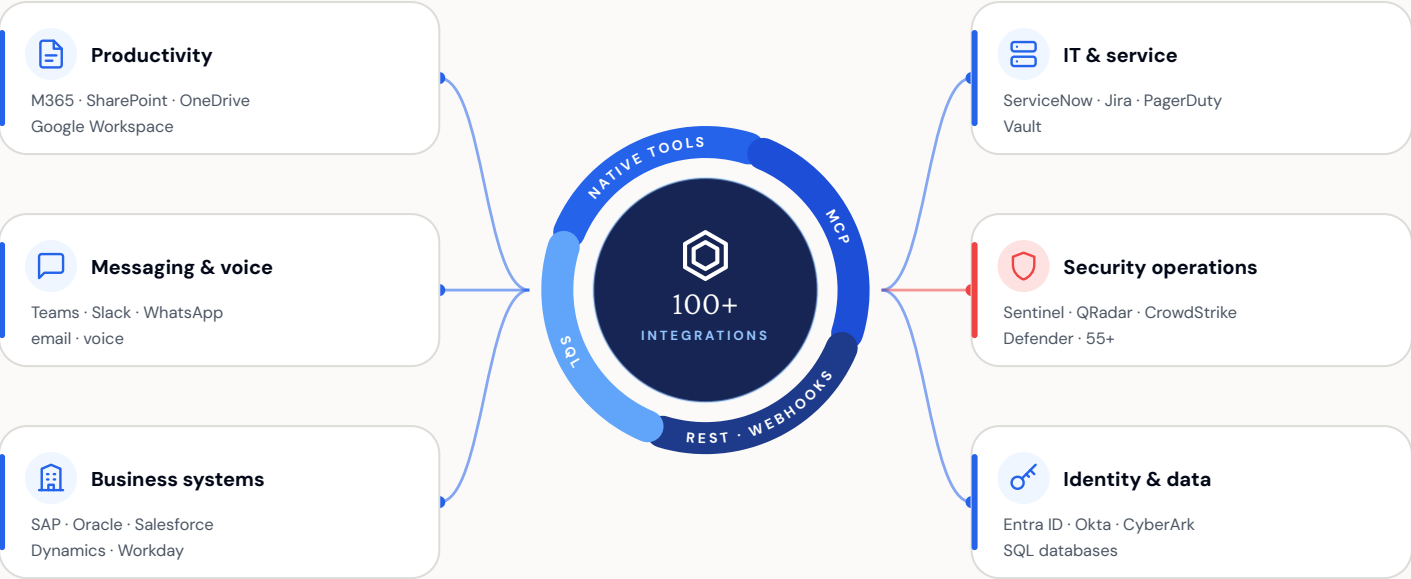
Capabilities
Tools, connectors, skills, sub-agents, sandbox

Governance
Firewall policies, approval gates, risk tier

Value
ROI tags, evals, monitoring, versions

03 Integrations & models.

Connections are set once per tenant, enabled per agent, and called with a shared service account or each user's own identity. Four methods reach anything your agents need.



Examples shown; systems without a ready connector are reached through MCP or REST.

BRING YOUR OWN MODEL, PER AGENT

OpenAI CLOUD

Azure OpenAI CLOUD

Anthropic CLOUD

Google Gemini CLOUD

Ollama LOCAL

vLLM LOCAL

TGI LOCAL

Per agent
Each agent picks its model; tenant default as fallback.

Your bill
Your own LLM accounts, no mark-up. Spend tracked per model.

Air-gap ready
Local models via Ollama, vLLM or TGI on your GPUs.

GPU SIZING FOR LOCAL MODELS · INT8 VRAM

7–8B Mistral 7B, Llama 3.1 8B	8 GB	1× 24 GB
12B Mistral / Gemma 12B	12 GB	1× 24 GB
~17B Llama 4 Scout	17 GB	1× 32 GB
32B Qwen 2.5 / 3	32 GB	1× 40–80 GB
70B+ Llama 70B class	70 GB	Multi-GPU

No GPU needed when you use cloud model APIs.

EVERY CHANNEL YOUR PEOPLE USE

Web & widget
Signed widget tokens

Teams
Bot Framework, Adaptive Cards

Slack
Messages trigger workflows

WhatsApp
Business Cloud API

Email
SMTP, IMAP, push

Voice
STT + streaming TTS

Your apps
REST, SSE, OpenAI API

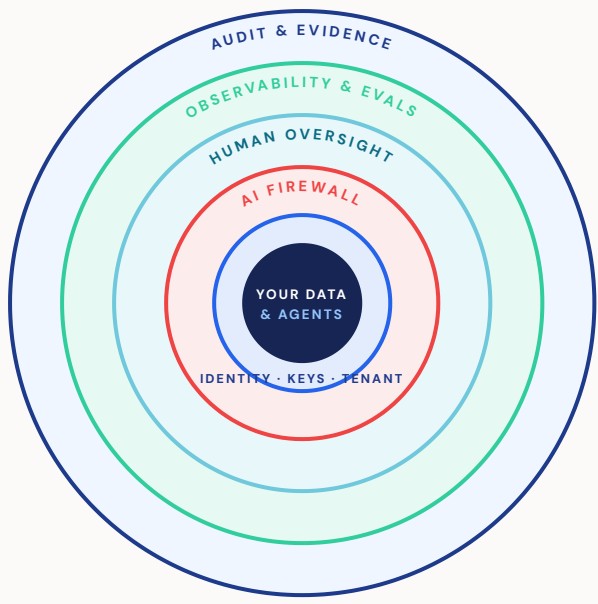
HOW AGENTS SIGN IN TO YOUR SYSTEMS

Shared service account
One account per connector, held in the key vault. Best for back-office automation and workflows.

Per-user identity
Each user signs in with their own Microsoft or Google account; the agent acts with their permissions only.

04 Security & governance.

Five layers of control wrap every agent and workflow. Each layer writes its own evidence, so audit answers come from records, not interviews.



- Identity, keys & tenant**
Schema-per-tenant isolation · JWT sessions and granular roles · SHA-256-hashed API keys with rotation · key vault secrets by reference · per-user OAuth for connectors · TLS everywhere
- AI Firewall & Edge**
Input and output policies on selected or all agents · 60+ detectors: PII and national IDs, secrets, injection, jailbreak, exfiltration, 14-language profanity incl. Arabic, LLM judges · block, mask or flag · Evolvable Edge on devices
- Human oversight**
Pre-approval, veto window, AI judge, post-review or blocked · on any tool, MCP call or workflow step · approver groups and SLA escalation · risk tier and coverage per agent
- Observability & evals**
End-to-end traces · error rate, guardrail hits and eval pass rate per agent and model · evals before go-live · supervisor monitoring
- Audit & evidence**
LLM, tool and decision logs · firewall executions with review status · every agent and workflow versioned; fleet as of any date

AUDIT EVIDENCE ON RECORD

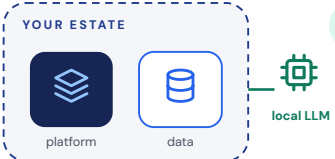
**LLM call log**
Prompt, response, model, tokens, cost

**Tool & decision log**
Every action, approver, reason, timing

**Firewall executions**
Detector, action, review status

**Change history**
Every version; fleet as of any date

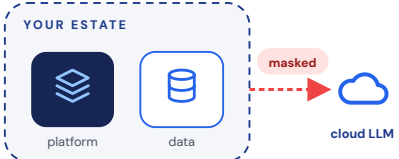
WHERE YOUR DATA GOES



AIR-GAPPED

Nothing leaves

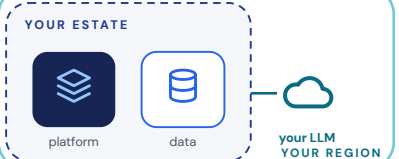
Platform, data and local models inside your network. No internet in or out.



HYBRID

Only masked model calls leave

Prompts reach your chosen cloud LLM over TLS, sensitive data masked by policy first.



YOUR CLOUD

Your tenancy, your region

In your subscription and region, e.g. UAE or KSA, under your cloud controls.

EVIDENCE FOR THE FRAMEWORKS YOU REPORT AGAINST

ISO/IEC 42001

NIST AI RMF

EU AI Act

OWASP LLM Top 10

UAE PDPL

DIFC Reg. 10

KSA PDPL & SDAIA

IIA

Compliance mapping · coming soon

SHARED RESPONSIBILITY

Evolvable.ai
Releases & security fixes · detector updates · deployment guides · support · sizing & hardening help

You
Infrastructure & OS · network controls · backups · TLS & DNS · identity & approvers · model accounts & GPUs

05 Deployment & sizing.

Containers for Docker Compose or Kubernetes with Helm. The same build runs everywhere; separate environments are separate instances or tenants.

Public cloud
AWS, Azure, Google Cloud or a local-region cloud

Private cloud
OpenShift, VMware or sovereign cloud

On-premises
Your data centre, cloud or local models

Air-gapped
No internet; local models; offline updates

PLATFORM SIZING

PILOT · MINIMUM

12 CPU cores x86_64

24 GB RAM

100 GB SSD, grows with knowledge

Databases co-located on one host.

PRODUCTION · INDICATIVE

2⁺ app nodes x 12 cores

32–64 GB RAM per node

500 GB+ SSD

Dedicated PostgreSQL (HA) and Neo4j. Confirmed in a sizing workshop.

SOFTWARE

Ubuntu 22.04+ / RHEL 8+

Docker 24+ & Compose v2

Kubernetes 1.27+ & Helm 3

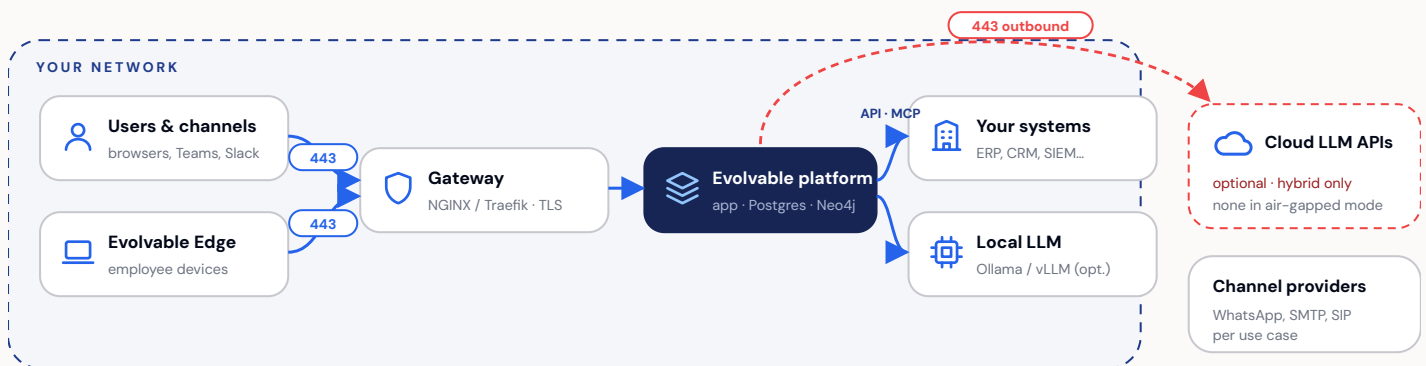
PostgreSQL 15+ & pgvector

Neo4j 5+

NGINX / Traefik

Ollama / vLLM / TGI (opt.)

NETWORK FLOWS



TYPICAL ROLLOUT



WHAT WE NEED FROM YOU BEFORE DAY ONE

- ✓ Linux hosts or a Kubernetes cluster sized as above
- ✓ DNS names and TLS certificates
- ✓ Identity provider and role mapping
- ✓ LLM accounts or GPU servers
- ✓ Service accounts for the systems to connect
- ✓ Firewall rules for the flows above