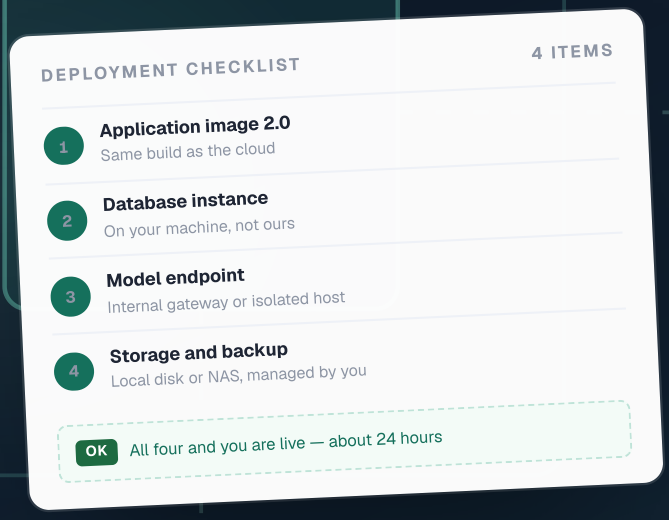
Onsites AI

DEPLOYMENT GUIDE · EDITION 2.0

Self-Hosted Guide

Your data, your servers

One image, two kinds of network: connected, or completely offline. You supply the model and the storage, so we charge nothing for either. We do the install, the updates, and the support.



Version
Edition 2.0

Published
November 2026

Hosts
Linux · Windows · macOS

Document ID
ONS-SH-EN-2.0

© 2026 Onsites AI · Regenerated with each release; latest edition at onsitesai.com/download.html

Contents

01 **When to self-host** Three kinds of organisation, one decision table 03

02 **Architecture and requirements** One image, four hops, offline and upgrades 04

03 **Four steps to live** Assessment, install, connect, pilot — with deliverables 05

04 **Licensing, cost and responsibility** Price, the difference from cloud, who owns what 06

At a glance

$20 /seat/mo
$240 per year
minimum 10 seats

24 h
install, configure,
migrate and test

60 days
full-product pilot
before you buy

$0 AI + storage
you supply both;
we charge neither

Your situation	Run it	Why
Customer data, call records, and invoices must stay in-country	Self-hosted	If it cannot leave the building, only a local install satisfies it.
Plant floor, defence, or an office network with no internet	Self-hosted, offline	Requires offline install and offline upgrade packages — not "activate online once".
You must use your own model and storage	Self-hosted	Model, keys, and disk are yours; no points are purchased.
Isolation required, but you want us to operate it	Dedicated cloud	Isolated instance and region choice, without running servers.
A normal team, starting now	Cloud SaaS	Live in 15 seconds; the free plan is enough to test the workflow.

This covers most selection conversations; the finer judgement calls are in chapter 1.

PURPOSE

For IT, security, and procurement to judge feasibility, resourcing, and cost. Usable as an appendix to an internal proposal.

PRICING BASIS

USD, annual. Final pricing is set by the order; changes are reflected here and on the price page.

OUT OF SCOPE

Customer-side network and hardware procurement, and third-party compliance testing.

FEEDBACK

Edition 2.0, November 2026. Assessment questions: support@onsitesai.com.

01

WHEN TO SELF-HOST

Three cases the cloud cannot solve

Self-hosting is not a more expensive SaaS plan. It exists for organisations whose data cannot leave the network, or who must own the model and the storage. The test is not budget — it is whether the data is allowed to travel.

1.1 Three kinds of organisation

Case one

Data cannot leave

Government, banking, healthcare, and tightly regulated subsidiaries: customer records and call history must stay inside the country and inside their own data centre.

Case two

Network is isolated

Plant floors, defence sites, internal-only offices. The software has to run without internet and upgrade from an offline package, not "connect once to activate".

Case three

Own model and storage

The model, the keys, and the files are supplied internally. No points are bought, and customer data does not leave the network by default.

1.2 The line of reasoning

Question 01

May the data leave?

No → self-hosted. Yes, but with a named region and isolation → dedicated cloud. Either is fine → cloud.

Question 02

Must it run offline?

Yes → self-hosted with offline upgrades. No → any of the three, decided on operating cost.

Question 03

Who supplies the model?

You → self-hosted, and we charge nothing for AI or storage. Us → cloud or dedicated.

Question 04

Who operates it?

Us → cloud or dedicated. Your infrastructure team, with our support → self-hosted.

Figure 1-1 Four questions and the answer usually settles itself. Any answer of "we must own this" collapses the choice to self-hosted.

1.3 When it is the wrong answer

- **"We care about security" but nothing forbids the cloud.** The cloud has TLS 1.3, encryption at rest, seat permissions, workspace isolation, and audit trails too.
- **There is no infrastructure team.** Someone has to own servers, patching, and backups. Without that, dedicated cloud is the honest middle: isolated instance and region choice, still operated by us.
- **You have not yet proven the workflow.** Start on the free plan, prove that one queue can absorb all channels, then decide where it should run.

Rule of thumb Organisations that genuinely need self-hosting mention "export controls", "data residency", or "air gap" in the first conversation. If none of those come up, start on cloud or dedicated cloud — lower cost, lower risk.

02

ARCHITECTURE

One image, four hops

The whole system is a single image: no middleware to maintain separately, no vendor runtime to install. Cloud hosts and internal servers run the same build, so upgrades, backups, and troubleshooting work identically.

2.1 Topology

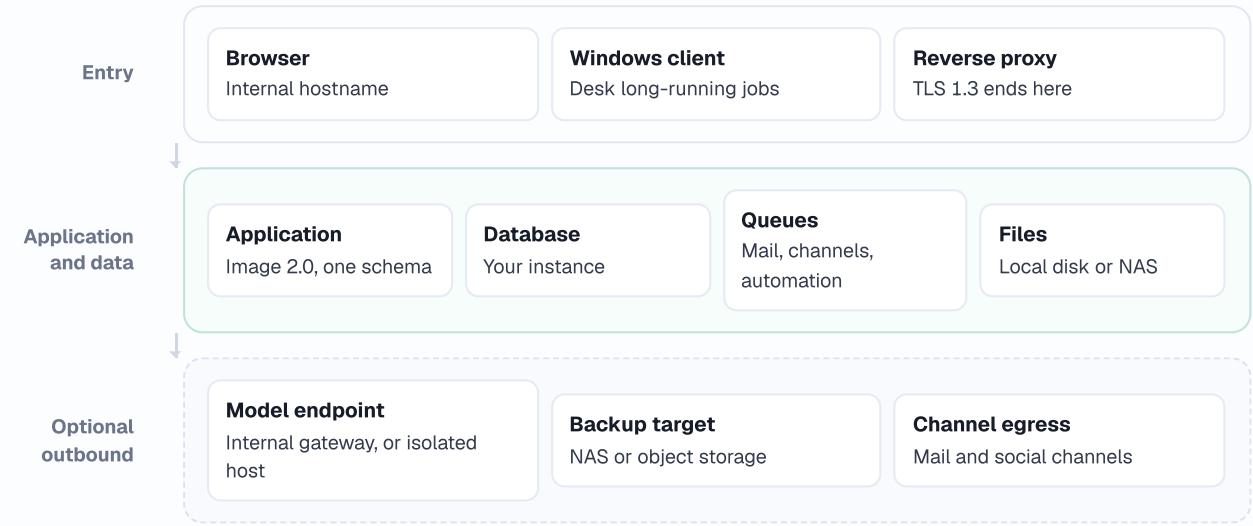


Figure 2-1 Four hops. The middle two never leave your network; the last row is optional — in a fully isolated environment the model endpoint can be internal and egress reduced to email.

2.2 Requirements

Item	Requirement
Hosts	Linux · Windows · macOS
Minimum	1 VM · 4 cores · 8 GB RAM · 3 Mbps
Network	Connected or fully offline; offline installs upgrade from offline packages
Model	Your API key, an internal gateway, or an isolated endpoint
Storage	Your disk; text is negligible, attachments dominate the footprint
Backup	Script included with the licence, run on your schedule

2.3 Offline and upgrades

- **Offline install.** Package and dependencies travel together; nothing is fetched at install time.
- **Offline upgrade.** New versions ship as an offline package and are applied from the UI — no activation call.
- **Same diagnostics.** Cloud and on-premise emit the same health checks and logs, so we debug from the same information.
- **We install it.** There is no self-serve download: a misconfigured install costs more than the install fee.

Sizing Under 50 seats and with attachments below a terabyte, a single 4-core, 8 GB machine is enough. Seats, concurrent sessions, and attachment volume are the three variables; tell us your expected scale during assessment and you get a purchase-ready parts list.

03

ROLLOUT

Four steps, each with a deliverable

The risk in a self-hosted project is not the software; it is nobody agreeing on what "installed" means. Each of the four steps therefore ends with a document: a parts list, a working environment, a connection test, and a pilot report you can take to a budget meeting.

3.1 The four steps

Step 01

Assess

Confirm architecture, residency, and isolation needs; produce a purchase-ready resource list.

Step 02

Install

A one-time $299 covers installation, configuration, and migration testing — about 24 hours.

Step 03

Connect

Add your model endpoint or key and point storage at your disk. Neither is charged for.

Step 04

Pilot

Run the full product for 60 days, then buy licences per seat. No annual commitment up front.

Figure 3-1 Steps 3 and 4 can swap order: some teams prefer to run the pilot before connecting a production model.

3.2 Deliverables and timing

Stage	Duration	You provide	We deliver
Assessment	1–3 business days	Network and isolation constraints, expected scale	Resource list, architecture confirmation, schedule
Install	About 24 hours	Server, hostname, certificate	Working environment, health-check report
Connect	Half a day	Model endpoint, key, storage path	Connection test, configuration record
Pilot	60 days	People and real work	Usage report, sizing advice, licence quote

3.3 What to watch in the pilot

- **Connect one channel first.** Website chat or the main mailbox — prove "customer in, someone answers, the record stays" before adding the second.
- **Run one real deal end to end.** Quote, order, contract, invoice; check that numbers, tax, and approvals survive a review.
- **Put a ceiling on AI.** Start with Copilot only; add Harness and Desk deliberately. Run a soft cap for two weeks before deciding on a hard one.

A pilot is not a load test Sixty days proves workflow and adoption, not throughput. If you expect high peaks — a promotion, a trade fair — say so during assessment and the sizing will account for it.

04

LICENSING

What it costs, and who owns what

The seat price includes the licence, software updates, and technical support. Model and storage are yours, so there is no AI fee and no storage fee on this route. Keeping those three things separate is what makes the arithmetic simple.

4.1 Price

$20 /seat/mo	**$2,400** floor	**$299** once	**$0** AI + storage
$240 per seat minimum 10 seats	10 seats, per year starting point	installation not credited	supplied by you we charge neither

Seats can be added at any time and are prorated to the renewal date. The installation fee is paid before the pilot begins, covers installation and migration, is invoiced separately, and is not credited against licence fees. Payment runs through 2Checkout (Verifone); invoices are available.

4.2 How it differs from the cloud

Item	Cloud	Self-hosted
Time to live	15 seconds, no install	About 24 hours, installed by us
Data location	Our infrastructure; region per plan	Your data centre or private cloud
Offline	Not possible	**Fully offline, offline upgrades**
Price	$15 / seat / month ($180/yr)	$20 / seat / month ($240/yr), 10 seats minimum
AI and storage	Points and storage as add-ons	Yours; we charge nothing for either
Operation	We run it	Install, updates, and support from us; hosts from you

4.3 Responsibility

Area	Customer	Onsites AI
Hosts and network	Servers, storage, network, operating system	Application install, configuration, upgrade support
Model and keys	Supplies endpoint and key; we never hold them	Integration and connectivity troubleshooting
Backup	Runs the schedule, stores the media	Provides the script and the restore procedure
Incidents	Host and network monitoring and response	Application-layer fixes and disclosure

Next step Describe the environment — Docker or VM, physically isolated, or a private-cloud VPC — and you get a decision-ready deployment answer within one business day. The pilot starts when the $299 installation fee is paid; licences come after sixty days.
onsitesai.com/self-host · support@onsitesai.com