

ARTIFICIAL INTELLIGENCE

Local AI with your company's context: Argos answers without your information leaving

Local AI

Argos

Documentation

16 October 2026 · 10 minute read

An AI assistant is only useful in network operations if it knows the network it operates, and if using it does not expose what it knows about that network. This article explains how Argos, Oktavia's assistant, answers with a model that runs on your own server, using vendor guides and your company's documents, and cites where every answer comes from.

A generic assistant does not know your network

Ask a general-purpose assistant how to restrict SSH on a router and you get an answer that is correct in the abstract: the one from an average manual, for any version, without knowing which access list your organisation uses for management, which networks it is administered from, or what your change procedure says. It sounds right and cannot be applied as is.

The obvious fix is to give it context: paste the configuration, the design, the inventory. That is where the second problem appears. A full configuration carries SNMP communities, tunnel keys, users and the whole network topology. Sending it to a cloud model means that information leaves your organisation, falls under a third party's retention policies and, in many sectors, runs straight into the security policy or the regulator.

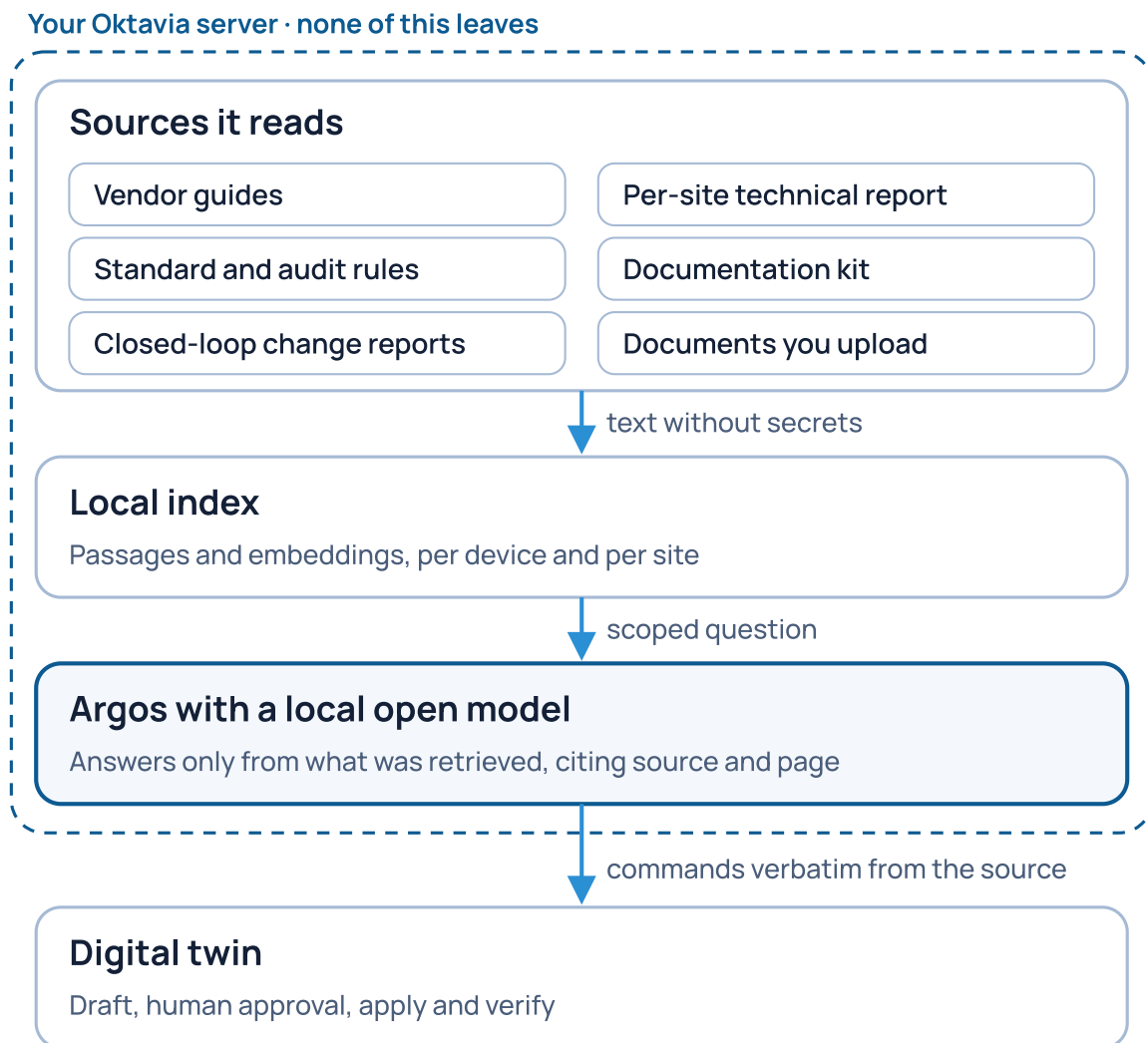
Either the assistant knows nothing about your network, or it knows too much and lives outside it. Neither works.

What Oktavia does: an AI agent that runs on your server

Oktavia includes a local AI agent: an open language model that runs on the same Oktavia server you install, together with an index of your documentation. When someone asks a question, Argos searches that index for the passages on the topic, hands them to the model and asks it to answer from them only. The technique is known as retrieval-augmented generation (RAG): the model does not need to "know" your network, it looks it up every time.

- **Nothing leaves the server.** Documents, index, question and answer stay on your infrastructure.
- **No per-question cost.** There are no tokens to pay a provider for each query; the cost is the server you already run.

- **It does not train on your data.** The model is not retrained with your information. Documents are looked up, not memorised, and removing them from the index is enough to stop them being used.



What goes into the index, how Argos answers, and the path a change takes before it touches the network.

What it reads: your company's context

A local model is only as good as what it can look up. The value is in the corpus, which is why Oktavia builds much of it for you. Every answer cites the document and page it came from.

- **Official vendor guides** for the platforms in your inventory: the Cisco IOS XE hardening guide, those for FortiOS, Junos and Meraki, and general guides such as the NSA network infrastructure security guide. They are downloaded automatically when the vendor allows it; otherwise they are uploaded as PDF, HTML or Markdown.
- **Your network's technical report**, per site, which Oktavia generates from the real discovered topology: devices, links, routing neighbours, services. It is the design document almost nobody keeps current, and here it comes from the network, not from someone's memory.

- **Your configuration standard:** the intent declared in the Configurator and what each compliance rule audits. Oktavia generates it, with secrets masked.
- **Your team's documents,** from a documentation kit with templates: design standard (HLD), troubleshooting runbooks, change procedure, security policy, and circuits and providers. Mandatory sections are validated on upload, and each template comes with a filled example.
- **Every closed-loop change report:** what was requested, what was applied, and the before and after. That way Argos learns from previous changes on your own network, not someone else's.

Where it is used

The local agent is not a separate chat window. It shows up where the operator is already working:

- **Diagnosis per incident.** In each incident Argos builds, the "What the documentation says" section summarises what to check, in which order and with which verification commands, based on your runbooks and the vendor guides for that device.
- **CVE remediation.** Argos reads the vendor's security advisory and only accepts commands that appear verbatim in it. With that it prepares urgent changes as drafts in the Digital Twin, or proposes them to the standard in the Configurator so new devices are born fixed.
- **Questions about your network.** "According to the documentation, how do I restrict SSH on this device?" is answered with the vendor guide for that platform, the management access list from your standard, and a citation for each part.

The guardrails

A language model can make things up with great confidence. In network operations that is not a matter of style: an invented command is a truck roll. So the agent works under rules that do not depend on the model behaving well:

- **It answers only from what was retrieved.** If the passages do not cover the question, it says so, and every statement carries its citation.
- **Verbatim commands only.** A command is accepted only if it appears exactly in a cited source. Anything the model "fills in" on its own is discarded.
- **Secrets are never indexed.** Lines with communities, keys or passwords are dropped before entering the index. What is not in the index cannot appear in an answer.
- **Scope per device and per site.** Retrieval is limited to the device or site in question and its organisation; an answer never mixes one customer's network with another's.
- **Changes still go through the loop.** Nothing is applied from an answer. Every change goes to the Digital Twin, a person approves the exact diff, and the closed loop verifies before and after.

AI proposes and cites. People approve. The network confirms.

✓ Good practices

- **Curate the corpus.** Official guides, current versions. An old document answers with old syntax.
- **Prefer text PDFs.** A scanned PDF is an image and needs character recognition before it is useful.
- **Fill in the kit's mandatory sections.** They are there because they are what the diagnosis looks up.
- **Name runbooks after the symptom.** "Tunnel down between sites" gets found; "Procedure 14" does not.
- **Regenerate Oktavia's part after a big change.** A new site or a migration changes the technical report and the standard.
- **Size the server.** On CPU an answer takes about a minute; with a GPU, seconds.
- **Escalate to a cloud model only explicitly.** If you ever need it, let it be at a person's request and with a bounded summary of the device or site, never with full configurations.

🔄 From incident to resolution

The local agent is one piece of the same path described in [the article on closing the loop](#): detect, propose, approve, apply and verify. And it follows the same principles as [security in Oktavia and Argos](#): no secrets to AI, and a tool that adds no attack surface.

What would an assistant tell you today about your own network?

If the answer is "nothing I could apply" or "we don't let it see the configuration", it is worth seeing with your own documentation. To quote, we only need a rough idea of how many devices you manage.

About Oktanet

Oktanet is a Mexican network automation company focused on controlled, traceable, and scalable operations for enterprise environments across Latin America.

[LinkedIn](#) · [YouTube](#)

© 2026 Oktanet · Oktavia · Public article, with no confidential customer information.