interweb

Generated on 15/11/2026 00:12

KNOWLEDGE BASE · PUBLIC DOCUMENT

Running AI models on your own VPS with Ollama: when it makes sense

Agents & AI on a VPS

CODE	VERSION	COUNTRY · LANGUAGE
RLX-AO-KB-932	v2026.10.10	Angola · English

Interweb Angola
RLX Angola, Lda. · Group RLX Holdings
Tax ID (NIF) 500 260 5724 · Commercial registration no. 26464-25/250724

https://www.meu.interweb.ao/knowledgebase/932

Running AI models on your own VPS with Ollama: when it makes sense

Agents & AI on a VPS Created on **October 10, 2026** 3 min read



Ollama is an open-source program that downloads open AI models and runs them on your own machine, without sending the text to a cloud provider. It makes sense when **data privacy or a predictable cost matters more than top quality**. It does not make sense if you want the best possible answer with the least work: for that, a cloud API is still the simplest road.

It is a third-party project, installed on a VPS with root access. Interweb provides the server; installing and maintaining Ollama and the models is up to you. The install steps and the available models are on the **official Ollama website**, which changes often.

Cloud or on your own server

	Cloud API	Ollama on your VPS
Where the text goes	To the model provider	Stays on your machine
Cost	Per use, grows with consumption	The server, plus the work of keeping it. The per-request charge disappears.
Quality	Access to the biggest and newest models	Limited to the models your machine can handle
Speed	Fast and steady	Depends on CPU and memory. Without a graphics card, it can be slow.
Upkeep	None on your side	Updating the program, the models and the system

When it is worth it

1. **Data that must not leave.** Internal documents, contracts, customer lists. If company policy or the law requires it, running on your own server avoids sending them to third parties. See **personal data protection**.

2. **High volume, simple tasks.** Sorting messages or summarising short texts, every day, at a fixed cost.

3. **Experimenting without paying per request.** For learning, a small model on a test VPS gives you the essentials.

4. **An agent that should depend on nobody.** OpenClaw can use local models, but that is where the strong machine is demanded. See **what it is and what you need**.

What decides whether your server copes

What rules is the **size of the model**. A bigger model needs more memory, and if it does not fit in the server's memory it will not run, or will run so slowly it is useless. Each model states its memory requirements on its page; read them before you download it. To see what your VPS has, use `free -h` and `nproc`. Plan specifications are on the **VPS**, **Cloud** and **Dedicated** pages; check there whether any includes a graphics card, because without one the models run on the CPU.

Do not open Ollama to the internet.
By default it usually listens only on the machine itself; confirm that in the documentation. If you expose it unprotected, anyone can use your server to run models. See **ports and firewall on a VPS**.

A local model makes things up too.
Running on your own server protects the data, it does not guarantee right answers. See **AI hallucinations**.

Start with a small model and measure: how long it takes, how much memory it uses, whether the quality is enough for the task. Move up in size only if the result justifies it.

RECOMMENDED

Want a server with more memory to try local models? Compare the plans.

See VPS plans →

SEE ALSO

- What OpenClaw is, and what you need to run it
- Hosting an AI agent: why it needs a VPS and not shared hosting
- Which operating system to choose for your VPS or server
- VPS plans

interweb

Interweb Angola

RLX Angola, Lda. · Group RLX Holdings

TAX ID (NIF)	500 260 5724
COMMERCIAL REGISTRATION NO.	26464-25/250724
REGISTERED OFFICE	Rua Direita do Patriota, Bairro Dangereux Belas, Luanda, Angola
CONTACT	info@interweb.ao · +244 942 011 191 · www.interweb.ao
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-AO-KB-932 · v2026.10.10 · Generated on 15/11/2026 00:12