Gerado a 14/11/2026 22:19

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Rodar modelos de IA no seu VPS com o Ollama: quando faz sentido

Agentes e IA em um VPS

CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-BR-KB-2934	v2026.10.10	Brasil · Português (Brasil)

MozOut Brasil
RLX Holdings (Pty) Ltd. · Grupo RLX Holdings
N.º fiscal (SARS) 943 907 6234 · N.º de registo (CIPC) 2025/667953/07

https://www.my.mozout.com/knowledgebase/2934

Rodar modelos de IA no seu VPS com o Ollama: quando faz sentido

Agentes e IA em um VPS | Criado em **10 de outubro de 2026** | 3 min de leitura



O **Ollama** é um programa de código aberto que baixa modelos de IA abertos e os roda na sua própria máquina, sem enviar o texto a um fornecedor na nuvem. Faz sentido quando **a privacidade dos dados ou a previsibilidade do custo pesam mais do que a qualidade máxima**. Não faz sentido se o que quer é a melhor resposta possível com o mínimo de trabalho: para isso, uma API na nuvem continua sendo o caminho mais simples.

É um projeto de terceiros, que se instala num VPS com acesso root. A MozOut fornece o servidor; a instalação e a manutenção do Ollama e dos modelos são suas. Os passos de instalação e os modelos disponíveis estão no **site oficial do Ollama**, que muda com frequência.

Nuvem ou no seu servidor

	API na nuvem	Ollama no seu VPS
Para onde vai o texto	Para o fornecedor do modelo	Fica na sua máquina
Custo	Por uso, cresce com o consumo	O servidor, mais o trabalho de o manter. O gasto por pedido deixa de existir.
Qualidade	Acesso aos modelos maiores e mais recentes	Limitada aos modelos que a sua máquina aguenta
Velocidade	Rápida e constante	Depende do processador e da memória. Sem placa gráfica, pode ser lenta.
Manutenção	Nenhuma da sua parte	Atualizar o programa, os modelos e o sistema

Quando vale a pena

1. **Dados que não podem sair.** Documentos internos, contratos, listas de clientes. Se a regra da empresa ou a lei o impõe, rodar no seu servidor evita enviá-los a terceiros. Veja **proteção de dados pessoais**.

2. **Muito volume, tarefas simples.** Classificar mensagens ou resumir textos curtos, todos os dias, a um custo fixo.

3. **Experimentar sem gastar por pedido.** Para aprender, um modelo pequeno num VPS de ensaio dá-lhe o essencial.

4. **Um agente que não deve depender de ninguém.** O OpenClaw pode usar modelos locais, mas é aí que se exige a máquina forte. Veja **o que é e o que precisa**.

O que decide se o seu servidor aguenta

O que manda é o **tamanho do modelo**. Um modelo maior pede mais memória, e se não couber na memória do servidor não corre, ou corre tão devagar que não serve. Cada modelo indica na sua página os requisitos de memória; leia-os antes de o baixar. Para ver o que o seu VPS tem, use `free -h` e `nproc`. As especificações dos planos estão em **VPS**, **Cloud** e **Dedicado**; confirme aí se algum inclui placa gráfica, porque sem ela os modelos rodam no processador.

Não abra o Ollama à internet.
Por padrão costuma ouvir apenas na própria máquina; confirme na documentação. Se o expuser sem proteção, qualquer pessoa usa o seu servidor para rodar modelos. Veja **portas e firewall num VPS**.

Um modelo local também inventa.
Rodar no seu servidor protege os dados, não garante respostas certas. Veja **alucinações de IA**.

Comece por um modelo pequeno e meça: quanto demora, quanta memória usa, se a qualidade chega para a tarefa. Só suba de tamanho se o resultado o justificar.

RECOMENDADO

Quer um servidor com mais memória para experimentar modelos locais? Compare os planos.

Ver planos de VPS →

VEJA TAMBÉM

- O que é o OpenClaw e o que precisa para o rodar
- Um agente de IA precisa de um VPS, não de hospedagem compartilhada
- Que sistema operacional escolher para o VPS
- Planos de VPS

MozOut Brasil
RLX Holdings (Pty) Ltd. · Grupo RLX Holdings

N.º FISCAL (SARS)	943 907 6234
N.º DE REGISTO (CIPC)	2025/667953/07
SEDE	72 York Road, Kensington Joanesburgo 2094, Gauteng, África do Sul
CONTACTOS	info@mozout.com · +27 78 823 3325 · www.mozout.com
HORÁRIO	2ª. à 6ª. Feira - 08 à 16 Horas Sábado & Domingo - Fechado

Versão online deste documento

RLX-BR-KB-2934 · v2026.10.10 · Gerado a 14/11/2026 22:19