

4 JEITOS DE RODAR IA NO *teu computador*

De graça, sem assinatura e, nos modelos locais, sem mandar as tuas conversas pra nenhum servidor. Escolhe um, instala e conversa.

Começa pelo que tu acha mais fácil: *aplicativo* (LM Studio ou Jan) ou terminal (Ollama).

01 / 04

OLLAMA · MIT · ★ 182 MIL

Ollama TERMINAL

O jeito mais usado de rodar modelo aberto no computador. Tu instala, digita um comando e conversa.

INSTALAR NO MAC (TERMINAL)

```
curl -fsSL https://ollama.com/install.sh | sh
```

INSTALAR NO WINDOWS (POWERSHELL)

```
irm https://ollama.com/install.ps1 | iex
```

Prefere baixar o instalador? ollama.com/download (Ollama.dmg no Mac, OllamaSetup.exe no Windows).

CONVERSAR COM O GEMMA 4, DO GOOGLE

```
ollama run gemma4
```

Cuidado: modelos com "cloud" no nome rodam nos servidores do Ollama. Pra ficar só local, liga o modo local: variável **OLLAMA_NO_CLOUD=1**.

Fontes: github.com/ollama/ollama · docs.ollama.com/faq · ollama.com/library/gemma4

02 / 04

LM STUDIO · APLICATIVO

LM Studio

SEM TERMINAL

Aplicativo pra Mac, Windows e Linux: tu baixa, escolhe o modelo numa lista e conversa. Grátis em casa e no trabalho desde 08/07/2025.

INSTALAR

lmstudio.ai/download, botão **Download LM Studio**. Abre o app, procura o modelo (ex.: Gemma 4 ou Qwen) e baixa.

O LM Studio não é código aberto, e cada modelo tem a sua própria licença.

Fontes: lmstudio.ai/download · lmstudio.ai/blog/free-for-work · lmstudio.ai/app-terms

03 / 04

JAN · APACHE-2.0 · ★ 44,9 MIL

Jan

SEM TERMINAL

Aplicativo de código aberto que parece o ChatGPT e roda Llama, Gemma, Qwen e o gpt-oss da OpenAI no teu computador.

INSTALAR

jan.ai, botão **Download** (jan.dmg no Mac, jan.exe no Windows). Abre, escolhe um modelo e conversa.

O Jan também liga em modelos online (OpenAI, Anthropic). Esses não são locais: as conversas saem do computador.

Fontes: github.com/janhq/jan · jan.ai

04 / 04

OPEN WEBUI · ★ 154 MIL

Open WebUI

POR CIMA DO OLLAMA

Uma tela parecida com a do ChatGPT por cima do Ollama, com histórico das conversas e vários modelos.

INSTALAR (TERMINAL, COM PYTHON 3.11)

```
pip install open-webui
open-webui serve
```

Depois abre **http://localhost:8080** no navegador.

De fábrica ele procura atualização e baixa modelos da internet. Pra cortar isso, liga **OFFLINE_MODE=true**. A licença não é 100% livre: acima de 50 usuários por mês tem regra de marca.

Fontes: github.com/open-webui/open-webui · docs.openwebui.com/reference/env-configuration

QUAL MODELO CABE *no teu* COMPUTADOR

A memória (RAM) é o que manda: o modelo precisa caber nela com folga. Pra ver a tua: no Mac, menu Apple > Sobre Este Mac; no Windows, Configurações > Sistema > Sobre.

TUA RAM	MODELO	EXEMPLO NO OLLAMA (TAMANHO DO DOWNLOAD)
8 GB	pequeno (uns 3B)	ollama run gemma4:e2b · 4,6 a 7,5 GB
16 GB	médio (uns 7B)	ollama run gemma4 · 6,6 a 9,5 GB ollama run gpt-oss:20b · 14 GB, roda a partir de 16 GB
32 GB	maior (uns 13B)	ollama run gemma4:26b · 16 a 19 GB

Fontes: README do Jan (8 GB p/ 3B, 16 GB p/ 7B, 32 GB p/ 13B) · ollama.com/library/gemma4 · ollama.com/library/gpt-oss

PRA COMEÇAR

MODELO

Prompt pra testar

LOCAL

"Me explica em 5 linhas o que é rodar uma IA no meu próprio computador e quando isso vale mais a pena do que usar o ChatGPT."

Desliga o wi-fi depois de baixar o modelo e pergunta de novo: ele continua respondendo.

Quer aprender a usar IA pra prosperar, em português?
Lista de espera do Prosperando com IA: aliceallgayer.com.br/lista?de=prosperando