SETE77 DIGITAL

SUA PRÓPRIA IA RODANDO SEM INTERNET

O passo a passo do Ollama, do zero até a primeira conversa. Sem mensalidade e sem mandar nada pra servidor de terceiro.

Guia escrito pra quem nunca abriu um terminal. Se travar em algum ponto, a página 5 tem os erros mais comuns e o que fazer em cada um.

ANTES DE BAIXAR NADA

QUANTO DE MEMÓRIA VOCÊ TEM

É esse número que decide o que vai rodar na sua máquina. Pular essa conferida é o motivo número um de gente instalar, achar lento e concluir que "não funciona".

NO WINDOWS

Aperte Ctrl + Shift + Esc
Vá na aba "Desempenho"
Clique em "Memória"
O número grande no topo é o que interessa

NO MAC

Menu da maçã (canto superior esquerdo)
Clique em "Sobre Este Mac"
A memória aparece na primeira tela

O QUE O NÚMERO SIGNIFICA

Menos de 8 GB	não vale a pena tentar
8 GB	roda modelo pequeno, e resolve
16 GB	confortável para o dia a dia
32 GB ou mais	roda modelo grande, perto do pago

Se você tem menos de 8 GB: não insista. A versão gratuita do ChatGPT no navegador continua sendo seu melhor caminho, e isso não é derrota, é a ferramenta certa pra máquina que você tem.

PASSO 01

INSTALAR O OLLAMA

É instalador normal, igual a qualquer programa. Não pede cadastro, não pede cartão e não pede email.

ONDE BAIXAR

ollama.com
(o projeto fica em github.com/ollama/ollama)

COMO

1. Entre no site e clique em Download
2. Escolha Windows, Mac ou Linux
3. Abra o arquivo baixado
4. Próximo, próximo, concluir

DEPOIS DE INSTALAR

Não vai abrir janela nenhuma. É assim mesmo.
O Ollama fica rodando quieto no fundo,
esperando você chamar. No Windows aparece um
ícone perto do relógio; no Mac, na barra de cima.

Se o antivírus reclamar: é comum com programa que fica rodando em segundo plano. O Ollama é um projeto público com código aberto e centenas de milhares de usuários. Se preferir conferir antes, o código inteiro está no GitHub.

PASSO 02

ESCOLHER O MODELO CERTO

A IA vem em versões de tamanhos diferentes, chamadas modelos. Escolha pela memória que você conferiu na página 1, não pelo nome mais bonito.

```
TABELA DE ESCOLHA
8 GB de RAM
   ollama run llama3.2
   O menor. Escreve, resume e organiza bem.

16 GB de RAM
   ollama run mistral
   ollama run gemma2
   Respostas mais longas e mais consistentes.

32 GB ou mais
   ollama run llama3.1:70b
   Chega perto da qualidade do ChatGPT pago.
   Baixa uns 40 GB, então tenha espaço em disco.

ERRO CLÁSSICO
Baixar o modelo gigante numa máquina modesta.
Ele instala, mas trava tudo e demora um minuto
pra responder. Comece pequeno. Você sempre pode
baixar outro depois: eles convivem sem conflito.
```

Espaço em disco: cada modelo ocupa de 2 a 40 GB. Pra apagar um que você não usa mais: **ollama rm nome-do-modelo**. Pra ver o que está instalado: **ollama list**.

PASSO 03

RODAR E CONVERSAR

Aqui é onde as pessoas travam, então vai bem devagar. Você precisa abrir uma janela preta chamada terminal. Ela assusta e não faz nada de errado sozinha.

ABRIR O TERMINAL

Windows: tecla Windows, digite "cmd", Enter
Mac: Command + espaço, digite "terminal", Enter

COLAR ESTA LINHA E APERTAR ENTER

```
ollama run llama3.2
```

O QUE VAI ACONTECER

Na primeira vez ele baixa a IA. Leva alguns minutos e mostra uma barra de progresso.
Quando terminar, aparece uma setinha piscando.
Pode escrever ali e apertar Enter.

O TESTE QUE CONVENCE

Desligue o wi-fi e pergunte outra coisa.
Continua respondendo. Está tudo rodando dentro do seu computador.

PRA SAIR

Digite /bye e aperte Enter.

Não gosta de terminal? Instale o **LM Studio** (lmstudio.ai) no lugar. É o mesmo princípio com uma tela pra clicar, também é grátis e não exige uma linha de comando. Quem prefere tela costuma desistir menos.

QUANDO DER ERRADO

OS PROBLEMAS MAIS COMUNS

"ollama não é reconhecido como comando"
O terminal foi aberto antes da instalação terminar. Feche e abra de novo.

DEMORA MUITO PRA RESPONDER
O modelo é grande demais pra sua máquina. Volte pra página 3 e baixe um menor.

O COMPUTADOR TRAVA INTEIRO
Mesma causa. Além disso, feche o navegador enquanto usa: aba de navegador come memória igual IA come memória.

RESPOSTA VEIO EM INGLÊS
Escreva no começo do pedido: "Responda sempre em português do Brasil." Modelo pequeno esquece o idioma se você não falar.

RESPOSTA VEIO FRACA OU GENÉRICA
Não é a máquina, é o pedido. Modelo local precisa de instrução mais detalhada que o ChatGPT. Use a estrutura da nossa série de prompts: contexto, tarefa numerada, formato e uma lista do que não fazer.

O que esperar de verdade: modelo local não busca na internet, não sabe notícia de hoje e não gera imagem. Ele substitui o pago no que você usa todo dia, que costuma ser escrever, resumir e organizar.

SETE77 DIGITAL

E se isso rodasse dentro do seu processo?

Instalar é o fácil. O que muda o jogo é a IA ligada no que o negócio já faz: no atendimento, no orçamento, na cobrança. Sem ninguém precisar lembrar de abrir nada.

- Site, SEO e Google Meu Negócio
- Google Ads que traz cliente, não clique
- Implementação de inteligência artificial no seu negócio

Quer ver o que dá pra automatizar no seu negócio? Chama a gente.

WHATSAPP
(11) 92523-0404

SITE
sete77digital.com.br

INSTAGRAM
@sete77digital