RENATA SOFIA · AI AGENCY · GUIA GRATUITO

IA Open Source.

Quando faz sentido tirar a IA da cloud — privacidade, custo e controlo. E quando não faz. Um guia honesto sobre correr modelos na própria máquina, sem hype e sem prometer milagres.

Renata Sofia Barbosa
Consultora imobiliária a usar IA no terreno

v1.0 · 2026-05
ai.renatasofia.net

00 ANTES DE COMEÇAR

Open source não é a resposta — é uma resposta.

> IA open source são modelos que correm no computador de quem os usa, não na cloud. A pergunta certa não é "devo mudar?". É "no meu caso, o que ganho e o que perco se mudar?".

Vou ser directa desde já: a maior parte dos consultores não precisa de IA local. O Claude Pro na cloud chega para a esmagadora maioria dos casos — mais simples, mais rápido de começar, melhor qualidade de output. Quem está à procura de uma desculpa para complicar o setup, este guia não a vai dar.

Mas há três situações concretas em que vale a pena conhecer a alternativa: **privacidade extrema** (dados que não podem sair da máquina), **volume alto e regular** (centenas ou milhares de itens em lote), e **operação offline** (zonas sem cobertura). Este guia mostra exactamente quando o jogo muda — e quando não muda.

COMO USAR ESTE GUIA

Não é preciso instalar nada para ler. O objectivo é dar critério: ao chegar ao fim, qualquer consultor sabe responder com honestidade à pergunta "open source faz sentido para mim?". Para quem decidir avançar, os capítulos 5 e 6 mostram o caminho mais simples, passo a passo.

O que está dentro

01 CLOUD VS LOCAL 4 MIN

A comparação directa, sem rodeios.

> Antes de decidir o que quer que seja, convém ver os dois lados lado a lado. Cada coluna tem vantagens reais — a questão é qual delas pesa mais no caso de cada um.

ASPECTO	CLOUD (CLAUDE PRO / API)	OPEN SOURCE (LOCAL)
Qualidade	**Topo de gama**	Boa, mas um nível abaixo
Privacidade	Dados vão para o servidor	**Nunca saem da máquina**
Velocidade	Depende da internet	Muito rápida (depende do hardware)
Custo	~17€/mês (Pro)	**0€ depois de descarregar (hardware à parte)**
Setup	**5 min (conta já existe)**	1-3h (instalar, modelos, integração)
Manutenção	**Zero (a Anthropic actualiza)**	Manual (actualizar modelos)
Funciona offline	**Não**	**Sim**

A TRADUÇÃO PARA O DIA-A-DIA

Cloud é mais simples, mais rápido de começar e tem melhor qualidade. Open source é mais privado, mais barato em volume, e funciona sem internet. Não há "melhor" universal — há melhor **para uma situação concreta**. O resto deste guia ajuda a identificar essa situação.

Repare numa coisa importante: o custo da cloud não é o problema. Dezassete euros por mês é insignificante face ao tempo poupado. O que move a decisão para local é quase sempre a *privacidade* ou o *volume* — raramente o preço por si só.

02 QUANDO FAZ SENTIDO 5 MIN

Os 3 casos em que vale a pena.

> São três situações distintas. Basta uma delas pesar a sério no dia-a-dia para começar a justificar a complexidade extra do local.

Caso 1 · Dados pessoais sensíveis

Cenário: triagem de uma base com centenas ou milhares de contactos — nome, email, telefone, histórico de transacções, notas pessoais. A Anthropic é clara em não treinar nos dados de quem usa o serviço (convém confirmar sempre a política actual). Mas quem trabalha sob cláusulas de confidencialidade contratuais pode simplesmente preferir que nada saia da própria máquina.

```
Workflow típico — zero dados pessoais saem da máquina:
1. Base de contactos no Google Sheets
2. Exportar localmente (CSV)
3. Modelo local processa (scoring, classificação, segmentação)
4. Resultado volta para o Sheets
5. Nenhum dado pessoal tocou na cloud
```

Caso 2 · Processamento em lote

Cenário: processar 500 descrições de imóveis de uma vez, ou fazer scoring de 4.000 contactos todos os meses. Para volume baixo-médio, a poupança da cloud é irrelevante. Para volume alto e regular, começa a fazer diferença real.

VOLUME MENSAL	CLAUDE API	OPEN SOURCE LOCAL
500 descrições	~3-5€	0€
4.000 scorings	~10-15€	0€
50.000 itens (escala industrial)	100-200€	0€

Caso 3 · Trabalhar sem internet

Cenário: uma visita a um imóvel em zona com cobertura má. Tirar fotos e gerar uma descrição inicial ali mesmo. Um modelo local no portátil funciona offline; ao voltar a estar online, sincroniza-se com o vault.

A ORDEM DE GRANDEZA HONESTA

Os valores da tabela são ordens de grandeza a Maio de 2026, para dar noção — não tabela de preços. Confirme sempre o custo actual antes de decidir. O ponto não é o número exacto: é perceber que abaixo de alguns milhares de itens/mês, o argumento do custo praticamente desaparece.

03 QUANDO NÃO FAZ SENTIDO 3 MIN

Os 4 sinais de que deve ficar na cloud.

> Esta é a parte que ninguém que vende "soluções de IA" quer dizer em voz alta. Para a maioria, a resposta certa é cloud — e está tudo bem com isso.

SINAL	PORQUE PESA
Sistema ainda por montar	Sem um vault organizado e instruções fixas, abrir uma camada extra de complexidade só dispersa. Primeiro o sistema sobre a cloud; depois, se justificar, o local.
Computador antigo	Modelos locais decentes precisam de hardware razoável. Se o portátil já aquece a abrir três separadores no browser, não está pronto.
A dor é qualidade, não privacidade	Os modelos open source são bons, mas estão um nível abaixo dos modelos de topo da cloud. Quem quer a melhor qualidade de output deve ficar na cloud.
Sem tempo para setup e manutenção	A cloud "funciona". O local "funciona se for cuidado" — 2-3h de instalação inicial mais manutenção pontual dos modelos.

A REGRA DE BOLSO

Se a única razão para querer open source é poupar dezassete euros por mês, não vale a pena. O custo do tempo de setup e manutenção supera de longe essa poupança, a não ser que o volume seja muito alto. Open source justifica-se por **privacidade** ou **escala** — não por uma factura mensal pequena.

04 O PANORAMA 4 MIN

Modelos e ferramentas úteis em 2026.

> A área evolui depressa: as escolhas de hoje podem mudar daqui a três meses. Esta é uma fotografia, não uma lista definitiva — convém verificar sempre o estado actual antes de decidir.

Modelos de texto

- **Llama (Meta)** — boa qualidade geral, vários tamanhos (7B, 13B, 70B parâmetros). Quanto maior, melhor o output, mas mais hardware é preciso.
- **Mistral** — leve, rápido, qualidade competitiva. Bom ponto de partida.
- **Qwen** — alternativa forte, sobretudo em multilingue (incluindo PT).
- **Gemma (Google)** — versão open weight do Gemini, em várias dimensões.

Modelos com visão (texto + imagem)

Úteis para analisar fotos de imóveis — sugerir descrição, detectar características:

- **Llama Vision · Pixtral (Mistral) · Qwen-VL**

Onde descarregar

FERRAMENTA	PARA QUE SERVE
Hugging Face huggingface.co	Repositório central, com modelos verificados pela comunidade. A fonte.
Ollama ollama.com	Interface simples (linha de comandos) para descarregar e correr modelos.
LM Studio lmstudio.ai	Interface gráfica tipo ChatGPT — o ponto de partida mais fácil para quem evita o terminal.

SOBRE OS "PARÂMETROS" (7B, 13B, 70B)

O "B" são milhares de milhões de parâmetros. Mais parâmetros, melhor o resultado — mas também mais memória e mais lentidão. Para começar e testar, um modelo de 3B a 7B chega bem. Os modelos grandes (70B) exigem hardware sério.

05 COMO COMEÇAR 4 MIN

O caminho mais simples, em 4 passos.

> Para quem decidiu que vale a pena testar, este é o percurso de menor atrito — via Ollama. Em menos de meia hora há um modelo a correr localmente.

1

Instalar o Ollama ~5 min

Ir a ollama.com, descarregar para o sistema (Mac, Windows ou Linux) e instalar como qualquer outro programa.

2

Descarregar um modelo 5-15 min

No Terminal (Mac/Linux) ou PowerShell (Windows), correr o comando abaixo. O download demora consoante a internet — o modelo tem 2-7 GB.

3

Testar ~1 min

Correr o modelo e escrever um pedido directo. A resposta sai em segundos, localmente, offline, a custo zero.

4

Integrar com o sistema avançado

Quando fizer sentido: o Ollama expõe um servidor local em `localhost:11434` que qualquer ferramenta consegue chamar (Claude Code, n8n, scripts próprios).

```
# Passo 2 — descarregar o modelo
ollama pull llama3.2

# Passo 3 — correr e testar
ollama run llama3.2

# Exemplo de pedido, já dentro do modelo:
> Descreve em 100 palavras um T2 em Cascais (80m2, varanda
  pequena, 2 quartos, renovado em 2020, condominio 60 euros/mes),
  para anuncio em Idealista. PT-PT. Sem buzzwords.
```

ALTERNATIVA SEM TERMINAL · LM STUDIO

Para quem prefere interface gráfica: descarregar o LM Studio (lmstudio.ai), instalar, ir a Discover, escolher um modelo (sugestão: Llama 3.2 3B ou Mistral 7B), descarregar e conversar — tipo ChatGPT, mas tudo local. Também expõe um servidor API local, tal como o Ollama. Recomendação: começar pelo LM Studio se a linha de comandos intimida; migrar para o Ollama mais tarde, se quiser.

06 HARDWARE + DECISÃO 3 MIN

O que é preciso, e como decidir.

> Modelos locais não correm bem em qualquer máquina. Antes de instalar o que quer que seja, vale a pena confirmar se o hardware aguenta.

COMPONENTE	MÍNIMO	RECOMENDADO
CPU	i5 / Ryzen 5 (2018+)	M-series (Mac) ou i7 / Ryzen 7 (2022+)
RAM	8 GB	16-32 GB
GPU	(opcional)	NVIDIA RTX (8 GB+ VRAM) ou Mac com chip M
Disco livre	20 GB	50-100 GB (para vários modelos)

Os **Macs com Apple Silicon (M1, M2, M3, M4)** são particularmente bons para correr modelos locais — a memória unificada ajuda muito. Quem pondera trocar de portátil e sabe que vai usar IA local pode ter isto em conta. A regra honesta: se o portátil é de 2017 ou anterior, ficar na cloud — não vale o investimento.

ÁRVORE DE DECISÃO · 4 PERGUNTAS

1. Trabalha com dados pessoais que NÃO podem sair da máquina? 2. Processa centenas/milhares de itens em lote, com regularidade? 3. Tem um computador potente (M1+ ou PC com GPU)? 4. Está disponível para 2-3h de setup mais manutenção pontual?

O VEREDICTO

3 ou mais "sim" → vale a pena experimentar local. **0 a 2 "sim"** → a cloud é a escolha certa por agora. Simples assim. E não há vergonha nenhuma em ficar na cloud — é onde a maioria deve estar.

07 O PADRÃO HÍBRIDO 3 MIN

O melhor dos dois mundos.

> Vou ser transparente: na maior parte do meu dia-a-dia, uso Claude Pro na cloud. A qualidade é melhor, a integração com o vault é mais directa, e o custo é insignificante face ao tempo poupado. O local entra em pontos muito específicos.

Onde uso modelo local

- **Triagem em lote da base de contactos** — milhares de nomes acumulados ao longo de anos. Em vez de enviar para a cloud, processa-se localmente. Privacidade máxima.
- **Pré-classificação de PDFs sensíveis** — por exemplo, cadernetas prediais com NIFs visíveis, antes de processar apenas as partes necessárias no Claude.
- **Geração de embeddings** para pesquisa semântica no vault — corre em segundo plano, sem tocar na cloud.

A REGRA DO PADRÃO HÍBRIDO

O modelo local trata da **camada de dados sensíveis**. O Claude trata da **camada de raciocínio e geração final**. Os dois lados articulam-se via vault e ficheiros intermédios — e nunca trocam dados pessoais directamente. É o desenho que dá privacidade onde é precisa, sem abdicar da qualidade onde ela conta.

Um caminho faseado, sem pressa

FASE	O QUE FAZER
Mês 1-3	100% cloud. Montar o sistema, construir os fluxos, perceber onde está o valor.
Mês 4-6	Avaliar com a árvore de decisão. Se a base cresceu e quer scoring regular sem cloud, ou se aparecem PDFs sensíveis com frequência — explorar o Ollama.
Mês 6+	Padrão híbrido, se aplicável. Local na camada sensível, Claude na camada de raciocínio.

A grande maioria nunca chega a precisar de IA local — e isso é um bom sinal, não um falhanço. O que quase toda a gente precisa é de *sistema* à volta da IA cloud. O local é uma ferramenta de precisão para casos concretos, não um objectivo em si.

RENATA SOFIA · AI AGENCY

Já tem o critério.
Falta aplicá-lo ao seu caso.

Este guia dá a teoria honesta: quando local faz sentido, quando não faz. Mas a decisão real depende do stack actual, dos dados que trata e do hardware que tem. É aí que uma avaliação técnica concreta vale mais do que qualquer guia genérico.

OFERTA DE ENTRADA

Sessão de descoberta · 20 minutos

Conversa de calibração, sem custo, sem compromisso. Olhamos juntos para o setup actual e a postura de privacidade e custo — e digo abertamente se faz sentido (ou não) avançar para uma Auditoria AI 1:1, onde avalio o stack a fundo e desenho o caminho certo (cloud, local ou híbrido).

→ ai.renatasofia.net

SE PREFERIR CONVERSAR PRIMEIRO

DM directa

Escreva-me em DM o que está a tentar fazer com IA — e que dados o preocupam. Respondo. Não há funil escondido.

→ Instagram @renatasofia.re · Facebook Renata Sofia Barbosa

DA MINHA MESA EM CASCAIS

Renata Sofia Barbosa

@renatasofia.re · ai.renatasofia.net