SETE77 DIGITAL

IA QUE RESPONDE COM OS SEUS PREÇOS

Dify mais Ollama: atendimento com inteligência artificial rodando inteiro na sua infraestrutura, sem mensalidade.

Comece pela página 1. Ela lista os documentos que você precisa organizar antes, e é o trabalho que decide se isso vai funcionar ou não. A ferramenta é a parte fácil.

O TRABALHO DE VERDADE

ORGANIZAR ANTES DE INSTALAR

Se a sua tabela de preços não está escrita em lugar nenhum, nenhuma IA vai adivinhar. Faça essa parte primeiro, mesmo que você decida não instalar nada depois. Ela sozinha já melhora o seu atendimento.

OS DOCUMENTOS QUE VOCÊ PRECISA TER

1. Tabela de preços
   Com o que está incluso e o que não está.
   Se o preço varia, escreva de que depende.

2. As 20 perguntas que mais chegam
   Com a resposta de cada uma. Pegue do seu WhatsApp, não da sua cabeça: o que você acha que perguntam é diferente do que perguntam.

3. Prazos reais
   De orçamento, de execução, de entrega.
   Os reais, não os otimistas.

4. Política de troca, garantia e cancelamento
   Mesmo que hoje seja "a gente vê caso a caso".
   Escreva o caso a caso.

5. O que você NÃO faz
   Serviço fora do escopo, região que não atende, tipo de cliente que não serve. Isso evita a IA prometer o que você não entrega.

FORMATO

Arquivos de texto simples, PDF ou Word. Um assunto por arquivo. O Dify aceita todos e lê melhor arquivo curto e organizado que documentão único.

O item 5 é o mais esquecido e o que mais causa problema. Sem ele a IA responde "sim, atendemos" pra qualquer coisa, porque foi treinada pra ser prestativa. Escreva os limites.

INSTALAÇÃO

DIFY EM UM COMANDO

O Dify roda em Docker. Ele fica ligado o tempo todo esperando pergunta, então precisa de um servidor, não do seu notebook.

ONDE ESTÁ

dify.ai · github.com/langgenius/dify
153 mil estrelas

INSTALAR

```
git clone https://github.com/langgenius/dify
cd dify/docker
cp .env.example .env
docker compose up -d
```

Abre no navegador no endereço do servidor.
A primeira conta que você criar vira a de admin.

O QUE PRECISA

Servidor com 4 GB de RAM no mínimo, 8 GB confortável. Docker instalado. Custa a partir de uns US$ 10 por mês em hospedagem comum.

CRIAR A BASE DE CONHECIMENTO

1. Menu "Conhecimento", botão "Criar"
2. Suba os arquivos da página 1
3. Deixe a divisão automática ligada
4. Espere processar (leva alguns minutos)

CRIAR O ASSISTENTE

1. Menu "Estúdio", "Criar aplicativo"
2. Escolha o tipo "Chatflow" ou "Agente"
3. Ligue a base de conhecimento que você criou
4. Teste na própria tela antes de publicar

Dica de instrução: na configuração do assistente, escreva "Responda apenas com base nos documentos fornecidos. Se a informação não estiver lá, diga que vai verificar e encaminhe pro atendimento humano." É essa frase que impede a IA de inventar preço.

O PULO DO GATO

LIGAR NO OLLAMA E ZERAR O CUSTO

O Dify é grátis, mas precisa de uma IA por trás. Apontando pro ChatGPT você paga por uso e manda dado de cliente pra fora. Apontando pro Ollama, o custo vai a zero e nada sai da sua infraestrutura.

1. INSTALAR O OLLAMA NO MESMO SERVIDOR

```
curl -fsSL https://ollama.com/install.sh | sh
ollama pull llama3.2
ollama pull nomic-embed-text
```

O segundo é obrigatório: é ele que transforma os seus documentos em algo que o Dify consegue buscar.

2. DEIXAR O OLLAMA VISÍVEL PRO DOCKER

No arquivo de configuração do Ollama, defina:

```
OLLAMA_HOST=0.0.0.0
```

Sem isso o Dify não enxerga o Ollama, mesmo os dois estando na mesma máquina. É o erro mais comum dessa montagem.

3. CONFIGURAR NO DIFY

Configurações > Provedores de modelo > Ollama

```
Nome do modelo:  llama3.2
URL base:        http://host.docker.internal:11434
Tipo:            Chat
```

Depois adicione o segundo, do mesmo jeito:

```
Nome do modelo:  nomic-embed-text
Tipo:            Text Embedding
```

4. APONTAR A BASE PRO MODELO CERTO

Na base de conhecimento, escolha nomic-embed-text como modelo de busca. No assistente, escolha llama3.2 como modelo de resposta.

Se as respostas vierem ruins: o problema quase nunca é o modelo. É documento mal organizado ou instrução vaga. Antes de trocar por um modelo maior, releia a página 1.

ANTES DE COLOCAR NO AR

O TESTE QUE EVITA VERGONHA

TESTE COM 20 PERGUNTAS REAIS

Pegue 20 mensagens de verdade do seu WhatsApp e passe uma por uma. Conte quantas ele acertou. Menos de 18? Não coloque no ar ainda.

TESTE O QUE ELE NÃO SABE

Pergunte coisa que não está em documento nenhum. Ele tem que dizer que vai verificar, não inventar. Se inventou uma vez, vai inventar com cliente.

TESTE O CLIENTE CHATO

Peça desconto. Pergunte se faz por metade do preço. Diga que o concorrente cobra menos. Ele não pode negociar nada que você não autorizou por escrito.

COMECE POR DENTRO

Antes de colocar pro cliente, use internamente por duas semanas. Deixe sua equipe perguntar pra ele em vez de te interromper. Você descobre os buracos sem que ninguém de fora veja.

SEMPRE TENHA A SAÍDA HUMANA

Toda conversa precisa de um caminho fácil pra falar com gente. Cliente preso em robô que não entende é cliente perdido, e isso apaga qualquer economia.

A PARTE HONESTA

Isso não é projeto de uma tarde. Contando a organização dos documentos, o teste e o ajuste, conte com duas a três semanas de trabalho de verdade. Quem promete em uma tarde está vendendo demonstração, não implantação.

Onde isso rende mais rápido: não no atendimento ao cliente, mas na equipe. Funcionário novo perguntando pro sistema em vez de te interromper dá retorno na primeira semana e não tem risco nenhum de imagem.

SETE77 DIGITAL

Duas semanas ou uma reunião

Esse guia tem tudo o que você precisa pra fazer sozinho. Se preferir que alguém faça, é exatamente esse o nosso trabalho.

- 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