Mozout

Gerado a 14/11/2026 22:19

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

IA que responde a partir dos seus documentos: o RAG explicado

Como funciona a IA

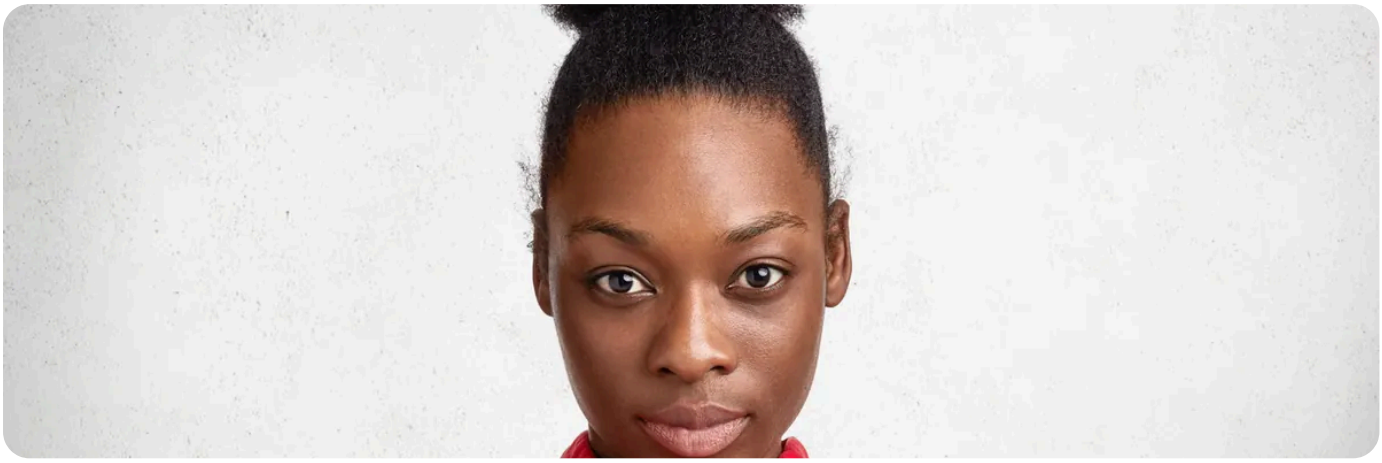
CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-BR-KB-2976	v2026.10.18	Brasil · Português (Brasil)

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/2976

IA que responde a partir dos seus documentos: o RAG explicado

Como funciona a IA Criado em **10 de outubro de 2026** Atualizado em **18 de outubro de 2026** 3 min de leitura



RAG (geração com recuperação, em inglês *retrieval-augmented generation*) é a técnica por trás de quase todos os assistentes que «respondem com base nos nossos documentos». Em vez de treinar um modelo com os seus textos, o sistema **procura os excertos relevantes e entrega-os ao modelo junto com a pergunta**. O modelo responde a partir deles. Reduz as respostas inventadas, mas não as exclui.

Como funciona, em cinco passos

1. **Reúnem-se os documentos:** as perguntas frequentes, as páginas do site, os manuais, os contratos-tipo.
2. **Cortam-se em pedaços** pequenos, de modo que cada um trate de uma coisa.
3. **Cada pedaço é transformado em números** (um «*embedding*») que representam o seu sentido, por um modelo próprio para isso. Textos com significado parecido ficam com números parecidos.
4. **Guardam-se num banco de dados próprio para pesquisa por sentido** (uma base vetorial). Algumas bases comuns têm uma extensão para isto.
5. **Quando alguém pergunta,** a pergunta também vira números, o sistema vai buscar os pedaços mais próximos e entrega-os ao modelo com a instrução «responde só com base nisto».

RAG ou outra coisa

Abordagem	Serve para	Limite
Colar o texto na conversa	Um documento, uma vez	Não escala, e paga-se o texto inteiro em cada pedido

Abordagem	Serve para	Limite
RAG	Muitos documentos, que mudam, e perguntas variadas	Só é tão bom como os documentos e a forma de os cortar
Afinar um modelo (*fine-tuning*)	Ensinar um estilo ou um formato	Não é uma boa maneira de lhe ensinar fatos que mudam

O que precisa para o montar

1. **Os documentos, limpos e atuais.** Se a política mudou e o documento antigo continua lá, o assistente cita-o.

2. **Um modelo de embeddings e um modelo de resposta,** de um fornecedor ou corridos por si. Veja **obter uma chave de API**.

3. **Um lugar onde guardar os vetores.** Um serviço pronto, ou um banco de dados seu com a extensão adequada. Se a instalar num VPS, veja **instalar o PostgreSQL num VPS** e confirme se a versão suporta a extensão que escolher.

4. **A cola entre tudo.** Uma aplicação sua ou um fluxo numa ferramenta como o n8n. Veja **o que é o n8n**.

Cuidado com quem pode ver o quê.

Se pôr no mesmo índice documentos públicos e internos, o assistente pode citar os internos a qualquer pessoa. Separe os índices ou controle o acesso por usuário.

Ainda pode inventar.

Se os excertos não trazem a resposta, um modelo mal instruído completa com o que «sabe». Mande-o dizer «não encontrei» e mostre sempre de que documento veio a resposta. Veja **alucinações de IA**.

Comece com poucos documentos e uma lista de perguntas de teste cuja resposta conhece. Se não acerta nessas, não está pronto para os clientes.

RECOMENDADO

Quer pôr um assistente destes num VPS seu? Veja os planos e escolha o tamanho.

Ver planos de VPS →

VEJA TAMBÉM

- Pôr um assistente de IA a conversar no seu site
- Tokens explicados: como se contam as APIs de IA
- Instalar o PostgreSQL num VPS
- Um agente de IA precisa de um VPS, não de hospedagem compartilhada

Mozout

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

RLX-BR-KB-2976 · v2026.10.18 · Gerado a 14/11/2026 22:19

Versão online deste documento