interweb

Gerado a 15/11/2026 01:59

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Tokens explicados: como se contam as APIs de modelos de IA

Como funciona a IA

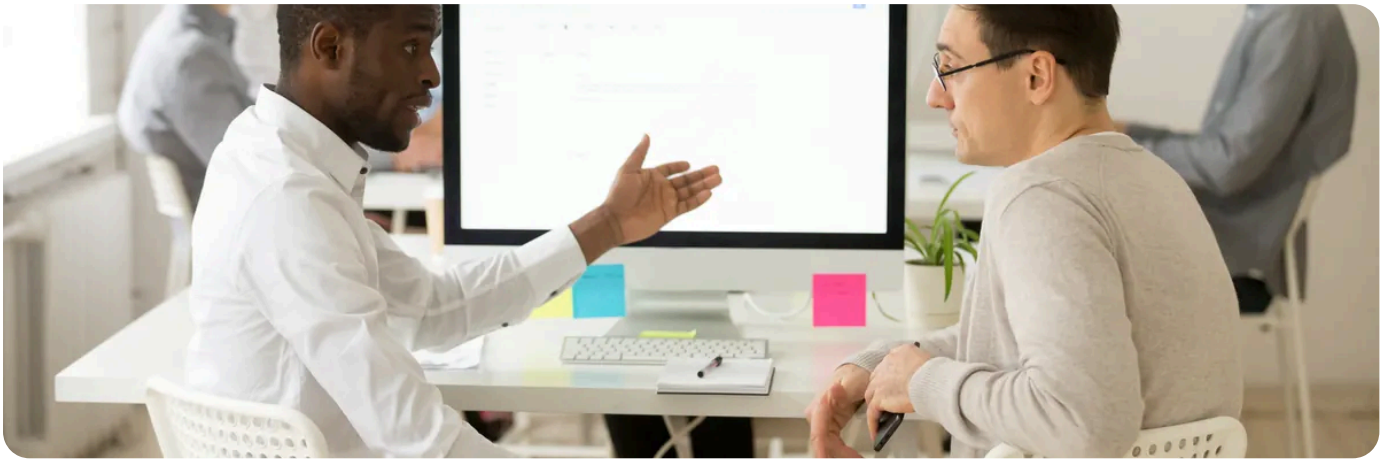
CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-MZ-KB-954	v2026.10.10	Moçambique · Português

Interweb Moçambique
RLX Holdings, S.A. · Grupo RLX Holdings
NUIT 401 191 607 · NUEL 105 045 974 · CAE 62022

https://www.meu.interweb.co.mz/knowledgebase/954

Tokens explicados: como se contam as APIs de modelos de IA

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



Um **token** é um pedaço de texto: pode ser uma palavra curta, parte de uma palavra comprida, um sinal de pontuação ou um espaço. As APIs de IA contam tokens, e não palavras nem caracteres, para medir **o que o modelo leu e o que escreveu**, e é por esses tokens que cobram. Perceber isto é perceber porque é que uma conversa comprida custa mais do que parecia.

Não escrevemos aqui o preço de nenhum modelo, porque muda. O preço por token está na **página de preços do fornecedor**, e o que cada modelo consumiu aparece no painel dele.

O que entra na conta

O que conta	Exemplo	Nota
O que envia (entrada)	A sua pergunta, as instruções, os documentos que anexa	Inclui as instruções de sistema que a aplicação acrescenta por si.
O que o modelo escreve (saída)	A resposta	Em regra, os fornecedores cobram a saída de modo diferente da entrada. Confirme na página de preços.
O histórico da conversa	As mensagens anteriores	Costuma ser reenviado a cada pedido, por isso o custo de uma conversa cresce.
Os resultados de ferramentas	Uma página lida, uma pesquisa, uma base de dados consultada	Tudo o que o modelo «lê» para responder entra como entrada.

Os três factos que explicam quase tudo

1. **Cada modelo conta à sua maneira.** O mesmo texto pode dar números de tokens diferentes em modelos de fornecedores diferentes. Use a ferramenta de contagem do fornecedor, se a houver, ou o campo de uso que a resposta da API devolve.

2. **Outras línguas costumam custar mais.** Em geral, o português gasta mais tokens do que o inglês para dizer o mesmo. A diferença depende do modelo, por isso meça com o seu texto.

3. **Há um tecto de contexto.** Cada modelo só lê um certo volume de tokens de uma vez (a «janela de contexto»). Se o histórico, os documentos e a resposta não couberem, o modelo corta ou falha. O valor vem na documentação do modelo.

Como gastar menos, sem perder qualidade

1. **Recomece conversas longas** quando o assunto mudar, em vez de arrastar um histórico que já não serve.

2. **Envie só o que interessa.** Não anexe um documento inteiro para perguntar sobre um parágrafo.

3. **Peça respostas curtas** quando bastam. A saída também se paga.

4. **Use o modelo adequado.** Tarefas simples não pedem o maior. Veja **qual ferramenta serve para cada tarefa**.

5. **Meça antes de automatizar.** Corra a tarefa umas quantas vezes à mão, veja o consumo no painel e só então a agende.

Um agente a trabalhar sozinho gasta tokens mesmo quando não olha.
Ponha um limite de gasto no fornecedor antes de agendar tarefas. Veja **o custo de um agente**.

Uma assinatura de uma aplicação de conversa e a chave de API são coisas diferentes, com contas diferentes. Pagar uma não dá saldo na outra.

PRECISA DE AJUDA?

Tem um servidor ou um site onde quer ligar a IA e não sabe por onde começar? Conte-nos o que pretende.

Abrir um pedido de suporte →

VEJA TAMBÉM

- **Obter uma chave de API de um modelo e guardá-la**
- **Quanto custa o OpenClaw: as chaves de API**
- **ChatGPT, Claude ou Gemini: qual serve para cada tarefa**

interweb

Interweb Moçambique

RLX Holdings, S.A. · Grupo RLX Holdings

NUIT	401 191 607
NUEL	105 045 974
CAE	62022
SEDE	Bairro 7 de Setembro Vilankulo 1304 · CEP 0411-02, Inhambane, Moçambique
CONTACTOS	info@interweb.co.mz · +258 84 635 1111 · +258 86 635 1111 · www.interweb.co.mz
HORÁRIO	2ª. à 6ª. Feira - 08 à 16 Horas Sábado & Domingo - Fechado

Versão online deste documento

RLX-MZ-KB-954 · v2026.10.10 · Gerado a 15/11/2026 01:59