interweb

Gerado a 28/11/2026 22:07

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

A resposta do agente com streaming: porque é que o primeiro pedaço pode demorar mais

Agentes pela API

CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-MZ-KB-3054	v2026.11.28	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/3054

A resposta do agente com streaming: porque é que o primeiro pedaço pode demorar mais

Agentes pela API · Criado em **28 de novembro de 2026** · 2 min de leitura



Com `stream: true`, a resposta chega aos pedaços, à medida que o modelo escreve. Num chat, o texto vai aparecendo e quem pergunta não fica a olhar para um ecrã parado. Com um agente funciona da mesma maneira, com uma diferença: o primeiro pedaço pode demorar mais.

Porque demora

Antes de escrever a resposta final, o agente pode usar ferramentas: ver a data, ler uma página, pesquisar na internet. Cada uso é uma volta ao modelo, e pode haver até 5 por pedido. Essas voltas não chegam pelo streaming. Só a resposta final chega aos pedaços, por isso, enquanto o agente trabalha, o pedido fica aberto e não chega nada.

Uma página lida com `web_fetch` pode levar até 10 segundos. Um agente sem ferramentas activas começa a escrever mais depressa.

Um exemplo em Python

Com o SDK da OpenAI e a chave na variável `INTERWEB_API_KEY`:

```
import os
from openai import OpenAI

client                =                OpenAI(base_url="https://api.interweb.co.mz/v1",
api_key=os.environ["INTERWEB_API_KEY"])

stream = client.chat.completions.create(
    model="agent/7c1e9a42-b35d8f06-a2c4e719-5d0b3f8e",
    messages=[{"role": "user", "content": "Estão abertos hoje?"}],
    max_tokens=400,
    stream=True,
    stream_options={"include_usage": True},
)
for chunk in stream:
```

```
    if chunk.choices and chunk.choices[0].delta.content:
        print(chunk.choices[0].delta.content, end="", flush=True)
    if chunk.usage:
        print("\nTokens:", chunk.usage.total_tokens)
```

O fluxo acaba com `data: [DONE]`. O SDK trata disso: o ciclo termina sozinho.

Não esqueça o include_usage

A contagem de tokens chega no último pedaço, se pedir `stream_options` com `include_usage`. Se não chegar, a reserva de créditos fica cobrada inteira. Peça-o sempre.

No seu chat

1. Mostre um sinal de espera, como «A escrever...», desde o envio até chegar o primeiro pedaço.
2. Dê tempo ao pedido. Se o agente tem ferramentas, não o corte ao fim de poucos segundos.
3. Use `max_tokens` para limitar o tamanho da resposta. O pedido pode baixar o máximo do agente, nunca subir.

Com uma chave `sandbox_`, o streaming também funciona: chega «Resposta de teste (sandbox).», com `usage` a zero. Serve para testar o código do chat sem gastar. O streaming em geral está em **mostrar a resposta à medida que é escrita**.

RECOMENDADO

Quer um agente a escrever em directo no seu chat? Veja o All in One AI.

Ver o All in One AI →

VEJA TAMBÉM

- **Mostrar a resposta da IA à medida que é escrita (streaming)** →
- **Usar o seu agente pela API: agent/ no campo model** →
- **As ferramentas do agente: ler páginas da internet e saber a data de hoje** →
- **Os limites do agente** →

All in One AI

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

RLX-MZ-KB-3054 · v2026.11.28 · Gerado a 28/11/2026 22:07

Versão online deste documento