Qwen 4 Max: API grátis, chave, preço e primeira chamada

Qwen 4 Max da Alibaba: como pegar 1 milhão de tokens grátis, criar a chave da API, quanto custa perto do Claude Opus 5.5 e a primeira chamada pronta.

1. Em 1 minuto
2. Faça agora
3. O que o vídeo mostrou
4. Deu problema?
5. Para quem quer ir além
6. Fontes

01 Em 1 minuto

O QUE É

O Qwen 4 Max é a IA mais forte da família Qwen, da Alibaba. Ela lê texto, imagem e vídeo.

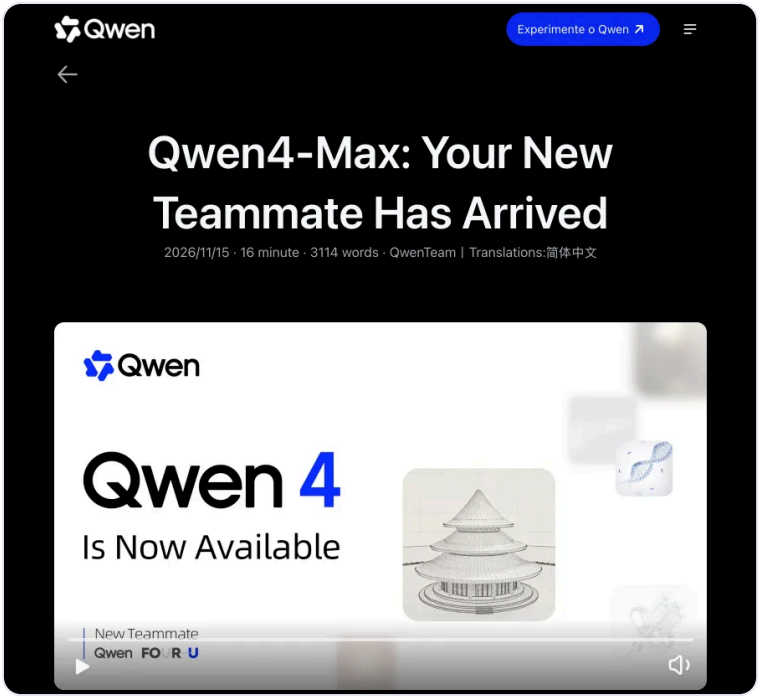
O QUE VOCÊ GANHA

Conta nova na região de Singapura ganha uma cota grátis pra testar. Pro Qwen 4 Max, a documentação fala em 1 milhão de tokens.

QUANTO CUSTA

Depois do grátis, US$ 1,50 por 1 milhão de tokens que você manda e US$ 4,50 por 1 milhão que ele escreve.

Palavra difícil: Token é a régua que a IA usa pra medir texto. Um texto grande vira muitos tokens.



A página oficial de lançamento do Qwen 4 Max, com o título e o vídeo de apresentação
Fonte: Qwen, blog de lançamento do Qwen4-Max

02 Faça agora

Se você só quer conversar com ele, sem código, use o Passo 1. Os Passos 2 a 5 são pra usar a API, que é onde vale a cota grátis.

> **Palavra difícil:** API é a porta que deixa um programa seu falar com a IA, sem abrir o site.

PASSO 1 Converse sem código

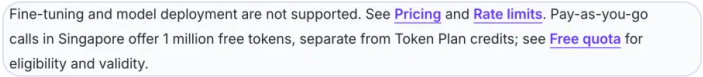
A Alibaba diz que dá pra testar o Qwen 4 Max no Qwen Studio, no QwenWork e no Qoder. É só entrar e escolher

o modelo.

PASSO 2 Crie a conta na região de Singapura

1. Abra o console da QwenCloud.
2. Crie a sua conta.
3. No canto de cima, à direita, deixe a região em **Singapore**.

> **Atenção:** A cota grátis só existe em Singapura. As outras regiões não têm.



O aviso oficial: 1 milhão de tokens grátis nas chamadas em Singapura
Fonte: QwenCloud, Latest model: Qwen4-Max

PASSO 3 Ative o grátis e trave a conta

1. Cadastre a forma de pagamento. A documentação pede isso pra liberar a cota grátis.
2. Abra a página **Free Tier** do console.
3. Ligue o modo que só usa a cota grátis. Ele também é chamado de "worry-free mode".

Com esse modo ligado, a chamada para quando o grátis acaba. Nada é cobrado.

Dica: A cota costuma valer 90 dias. Ela vence mesmo sem uso.

PASSO 4

Crie a sua chave

1. Abra a página API Keys, com a região em Singapore.
2. Clique em **Create API key**.
3. Escreva uma descrição e clique em **Generate Key**.
4. Copie a chave na hora e guarde. Ela só aparece inteira uma vez.

Palavra difícil: Chave de API é uma senha comprida. Ela diz à QwenCloud que o pedido é seu.

Atenção: Nunca compartilhe a chave. Quem tiver ela gasta no seu nome.

PASSO 5

Faça a primeira chamada

Esse é o exemplo da documentação oficial. Você precisa de Python no computador.

1. Guarde a chave no terminal. Troque `sua_chave_aqui` pela sua chave:

```
export DASHSCOPE_API_KEY=sua_chave_aqui
```

2. Instale a biblioteca:

```
pip install openai
```

3. Crie um arquivo `teste.py` com este código:

```
import os
from openai import OpenAI

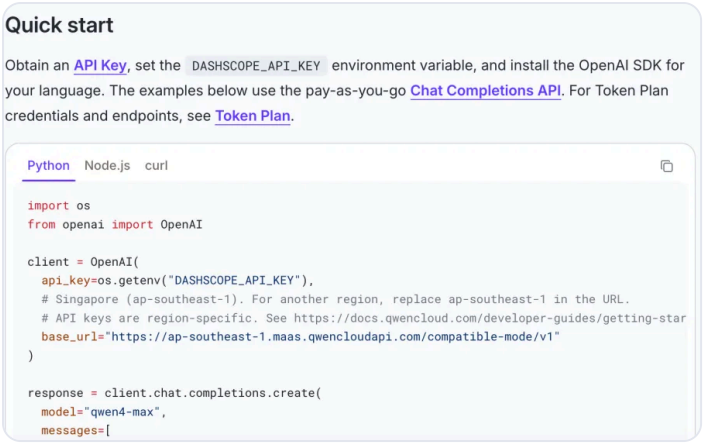
client = OpenAI(
    api_key=os.getenv("DASHSCOPE_API_KEY"),
    base_url="https://ap-southeast-
1.maas.qwencloudapi.com/compatible-mode/v1",
)
completion = client.chat.completions.create(
    model="qwen4-max",
    messages=[{"role": "user", "content":
"Summarize the benefits of solar energy in
three bullet points."}],
)
print(completion.choices[0].message.content)
```

4. Rode no terminal:

```
python teste.py
```

A resposta do Qwen 4 Max aparece na tela.

Palavra difícil: Terminal é o programa onde você digita comandos no computador.

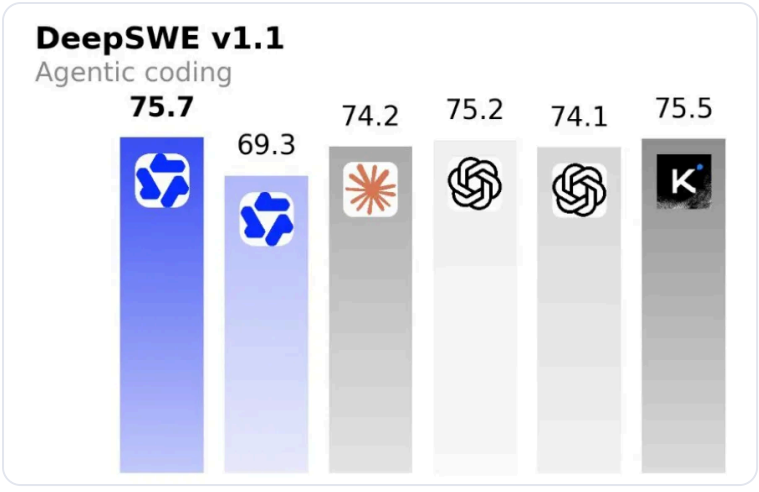


O início rápido oficial, com o modelo qwen4-max e o endereço de Singapura
Fonte: QwenCloud, Latest model: Qwen4-Max

03 O que o vídeo mostrou

O teste de programação

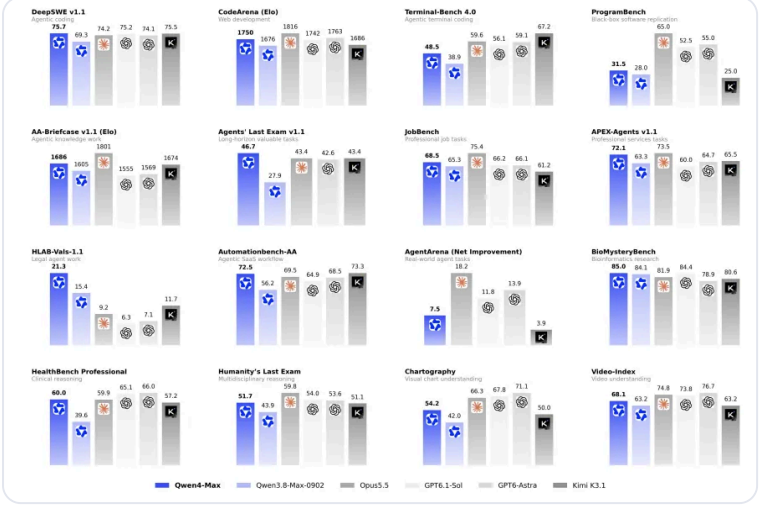
No DeepSWE, um teste de programação, o gráfico oficial da Alibaba dá 75,7 pro Qwen 4 Max e 74,2 pro Claude Opus 5.5. Os números são da própria Alibaba.



O gráfico do DeepSWE no anúncio oficial do Qwen 4 Max
Fonte: Qwen no X, anúncio de 16/11/2026

Ele não ganha em tudo

No mesmo gráfico, são 16 testes. O Opus 5.5 fica na frente do Qwen 4 Max em 10 deles.



Os 16 testes do gráfico oficial de lançamento
Fonte: Qwen no X, anúncio de 16/11/2026

O preço, lado a lado

	Qwen 4 Max	Claude Opus 5.5
O que você manda (1 milhão de tokens)	US$ 1,50	US$ 4
O que ele escreve (1 milhão de tokens)	US$ 4,50	US$ 20

Preços de Singapura no Qwen e do site da Anthropic no Opus, conferidos em 26/11/2026.

Qwen4-Max prices

USD per million tokens. Thinking and non-thinking modes use the same price, with no input-length tiers.

Billing item	Singapore (International)	Beijing (China) / Global
Input	$1.5	$1.413
Output (thinking + answer)	$4.5	$4.24
Cache write (5m)	$1.875	$1.767
Cache write (1h)	$3	$2.827
Explicit / implicit cache read	$0.075	$0.071

A tabela oficial de preço do Qwen4-Max
Fonte: QwenCloud, Pricing

Model	Base tokens	
Name	Input	Output
Claude Opus 5.5 For long-running agentic coding and knowledge work	$4 / MTok	$20 / MTok

A linha do Claude Opus 5.5 na tabela oficial da Anthropic
Fonte: Claude Platform Docs, Pricing

As três coisas

- **Lê texto, imagem e vídeo.**
- **Aguenta um material enorme numa conversa só:** até 1 milhão de tokens.

- **Chama ferramentas e devolve no formato que o seu sistema espera.**



Capability	Support	Capability	Support
Text generation	✓	Thinking	✓
Function calling	✓	Structured output	✓
Image understanding	✓	Video understanding	✓
Explicit cache (Qwen4)	✓	CoT Summary	✓
Web search	✓	PDF understanding	✓

A tabela oficial de recursos do Qwen4-Max
Fonte: QwenCloud, Latest model: Qwen4-Max

Pesos fechados

O Max não tem os pesos abertos. Ou seja, não dá pra baixar e rodar no seu computador. A Alibaba diz que os modelos menores do Qwen 4 vão sair abertos.

04 Deu problema?

O que aconteceu	O que fazer
A chave dá erro de autenticação	A chave só funciona no endereço da região onde foi criada. Confira se as duas são de Singapore.
Não aparece cota grátis	A cota só existe em Singapura e pede a forma de pagamento cadastrada. Veja na página **Free Tier**, aba **Model Invocation Quota**.
Tenho medo de ser cobrado	Ligue o modo que só usa a cota grátis, na página **Free Tier**. A chamada para quando o grátis acaba.
O saldo não mudou depois do teste	O console pode levar alguns minutos pra atualizar.

Pronto, o principal acabou aqui. Daqui pra baixo é a parte técnica, pra quem quer entender os detalhes.

05 Para quem quer ir além

- **Nome do modelo na API:** `qwen4-max`.
- **Endereço de Singapura:** `https://ap-southeast-1.maas.qwencloudapi.com/compatible-mode/v1`.
- **Tamanho:** 2,4 trilhões de parâmetros no total, com 71 bilhões ativos por token, segundo a Alibaba.
- **Conversa:** até 1.000.000 de tokens. Saída de até 131.072 tokens sem raciocínio.
- **Cache:** o que se repete entre chamadas sai por US$ 0,075 o milhão em Singapura.
- **Outras APIs aceitas:** Responses, Anthropic Messages e DashScope.
- **Recursos marcados na doc:** chamada de função, saída estruturada, raciocínio, busca na web e leitura de PDF.
- **Fora da região de Singapura:** US$ 1,413 na entrada e US$ 4,24 na saída, sem cota grátis.

06 Fontes

Conferido em 26/11/2026.

- Qwen, blog de lançamento do Qwen4-Max
- Alibaba Cloud, anúncio do Qwen4-Max (16/11/2026)
- Qwen no X, anúncio com o gráfico e o preço
- QwenCloud, Latest model: Qwen4-Max
- QwenCloud, Pricing

- QwenCloud, Free quota
- QwenCloud, Get an API key
- Claude Platform Docs, Pricing

Tem guia novo a cada vídeo.

Todos em marraads.online · siga @marraads