

Gemini Decision 1: como usar a Decisions API, preço e limites

Gemini Decision 1 na Decisions API do Google: crie a chave grátis, rode o teste da caneca, veja os 3 tipos de resposta e quanto custa no plano grátis e no pago.

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

Uma IA do Google que não escreve texto. Ela só decide: sim ou não, uma opção de uma lista, ou uma nota.

O QUE VOCÊ GANHA

Você manda um conteúdo e perguntas fechadas. Ela devolve a decisão, a probabilidade e o quanto tem certeza.

QUANTO CUSTA

Tem plano grátis. No pago, US$ 0,10 por 1 milhão de tokens que você manda. A resposta não é cobrada.

Every question uses one of three question types:

Question type	Goal	Returns
Yes/no (`noul`)	Check whether a statement holds	`noul` (0.01 to 0.99), `confidence`
Choice (`choice`)	Pick one option from a fixed list	`choice`, `probabilities`, `confidence`
Score (`score`)	Rate input against ordered rubric levels	`score`, `probabilities`, `legend`, `confidence`

Os três tipos de pergunta e o que cada um devolve
Fonte: Google AI Studio, Decisions API

02 Faça agora

Se você nunca programou, faça só o Passo 1. Ele não tem código nenhum.

Os Passos 2 a 5 rodam o primeiro teste da documentação do Google, o da caneca do vídeo. Pra eles, você precisa de:

- [ ] **Uma conta Google**, para entrar no Google AI Studio.
- [ ] **Python no computador.**

> **Palavra difícil:** Python é uma linguagem de programação. O teste deste guia é um arquivo pequeno escrito nela.

PASSO 1 Teste sem código no Playground

O Google tem o Decisions Playground. Nele, você testa direto na tela os três tipos de pergunta: sim ou não, escolha e nota.

PASSO 2 Crie a sua chave

1. Abra a página API Keys do Google AI Studio.
2. Entre com a sua conta Google.
3. Clique em **Create API key**.
4. Copie a chave e guarde num lugar seguro.

> **Palavra difícil:** Chave de API é uma senha comprida. Ela diz ao Google que o pedido é seu.

PASSO 3 Guarde a chave no computador

No Mac, abra o arquivo `~/.zshrc` e coloque esta linha no final. Troque `sua_chave_aqui` pela sua chave:

```
export GEMINI_API_KEY=sua_chave_aqui
```

Depois feche e abra o terminal. Ou rode:

```
source ~/.zshrc
```

Assim, o Python acha a chave sozinho.

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

Atenção: A documentação pede para nunca colocar a chave dentro de um projeto que vai para o Git.

PASSO 4 Instale a biblioteca do Google

Rode no terminal. Se ela já estiver instalada, o comando atualiza:

```
pip install -U google-genai
```

Palavra difícil: Biblioteca é um pacote de código pronto. Esta é a do Google, e o teste usa ela.

PASSO 5

Rode o teste da caneca

É o primeiro exemplo da documentação do Google. Um cliente escreve que a caneca chegou trincada. A pergunta é: o produto chegou danificado?

1. Crie um arquivo chamado `teste.py` com o código do exemplo oficial:

```
from google import genai

client = genai.Client(http_options=
{"api_version": "v1beta"})

result = client.decisions.create(
    model="gemini-decision-1",
    input="The blue mug I ordered on Monday
arrived cracked on Wednesday.",
    questions={
        "is_damaged": {
            "type": "noul",
            "instructions": "Did the item
arrive damaged?",
        },
    },
)

answer = result.answers["is_damaged"]
print(f"P(yes)={answer.noul:.2f} confidence=
{answer.confidence:.2f}")
```

2. Rode no terminal:

```
python teste.py
```

3. A resposta mostra a probabilidade de "sim" e a confiança. No exemplo oficial, as duas vêm em 0,99. E o total de tokens de saída é 0.

The API returns calibrated probabilities and token usage:

JSON


```
{
  "answers": {
    "is_damaged": { "type": "noul", "noul": 0.99,
"confidence": 0.99 }
  },
  "usage": {
    "total_input_tokens": 100,
    "input_tokens_by_modality": [{ "modality":
"text", "tokens": 100 }],
    "total_output_tokens": 0,
```


A resposta da API, com "noul": 0.99 e zero tokens de saída
Fonte: Google AI Studio, Decisions API, Quickstart

03 O que o vídeo mostrou

Três tipos de resposta

- **Sim ou não:** devolve a chance de a resposta ser "sim".
- **Escolha:** escolhe uma opção de uma lista que você define.
- **Nota:** dá uma nota numa escala que você define.

Os limites de cada tipo estão em "Para quem quer ir além".

O nível de confiança

Toda resposta vem com um número que mostra o quanto a IA tem certeza.

A documentação sugere usar esse número pra decidir. A resposta segue sozinha, ou vai para uma pessoa revisar.

> **Atenção:** Em ações de alto risco, a documentação manda manter uma pessoa revisando.

Até mil perguntas sobre o mesmo arquivo

Num pedido só, você faz de 1 a 1.000 perguntas sobre o mesmo conteúdo. Pode misturar os três tipos.

O conteúdo é cobrado uma vez só, não importa quantas perguntas você faça. Cada pergunta a mais só soma o texto dela.

> Put independent questions about the same input in one request. Shared input counts once per request regardless of how many questions you ask (each question adds only its own instructions and criteria), and question types can be mixed. This example routes a support ticket using order data and the customer message.

O trecho da documentação: o conteúdo conta uma vez por pedido
Fonte: Google AI Studio, Decisions API, Ask multiple questions

O que ela lê

- **Tipos:** texto, imagem, áudio, vídeo e PDF.
- **Tamanho:** até 100 MB por pedido. No caso de PDF, até 50 MB.

Onde o Google já usa

- **Gemini Notebook:** ela escolhe qual modelo responde cada pergunta. A espera nas perguntas comuns cai até 40%.
- **Stitch:** o app de design do Google usa a Decisions API. Ele monta o fluxo de telas de um aplicativo em segundos.



O Stitch montando as telas de um app (vídeo oficial, velocidade real)
Fonte: blog oficial do Google, 11/11/2026

Quanto custa

Você só paga pelo que manda. A resposta não gasta nenhum token de saída.

> **Palavra difícil:** Token é a unidade que a IA usa pra medir o tamanho do que você manda. A cobrança é por quantidade de tokens.

Plano	Entrada	Saída
Grátis (Free Tier)	De graça	Não cobra (0 tokens de saída)
Pago	US$ 0,10 por 1 milhão de tokens	Não cobra (0 tokens de saída)

> **Atenção:** No plano grátis, o Google usa o que você manda para melhorar os produtos dele. No pago, não.

Our multimodal decision model for binary, categorical, and rubric evaluations across text, images, video, audio, and PDFs using the Decisions API.

Each request bills shared input tokens once, regardless of question count. Each question adds tokens for its own instructions, criteria, and formatting, and generates zero output tokens. Batch API isn't supported.

Standard

Feature	Free Tier	Paid Tier, per 1M tokens in USD
Input price (text, image, video, audio, PDF)	Free of charge	$0.10
Output price	Not applicable (0 output tokens)	Not applicable (0 output tokens)
Context caching price	Not available	Not available
Used to improve our products	Yes	No

A tabela de preços oficial do Gemini Decision
Fonte: Google AI Studio, Pricing

O plano grátis tem limite de uso. Os números estão em "Para quem quer ir além".

04 Deu problema?

O que aconteceu	O que fazer
O botão de criar a chave não aparece	A mensagem "You do not have permission to create a key in this project" quer dizer que a sua conta não tem permissão naquele projeto. Use um projeto seu.
O teste dá erro com a biblioteca já instalada	A Decisions API exige a biblioteca na versão 3.2 ou mais nova. Rode `pip install -U google-genai` de novo.
Erro de conteúdo vazio	A API não aceita pedido sem conteúdo. Mande o texto, a imagem ou o arquivo junto.
Quero mandar muitos pedidos de uma vez pela Batch API	O Gemini Decision não aceita a Batch API. Os pedidos são feitos um a um.
A chave não é encontrada	Confira se a variável se chama `GEMINI_API_KEY`. E se você abriu um terminal novo depois de editar o `~/.zshrc`.

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

05 Para quem quer ir além

O modelo e a API

- **Nome:** Gemini Decision 1, modelo do Google lançado em 11/11/2026.

- **Onde roda:** na **Decisions API**, dentro do Google AI Studio.
- **Biblioteca:** a do Google, na versão 3.2 ou mais nova. A Decisions API não funciona nas versões antigas.
- **Outras linguagens:** a documentação também tem o exemplo em JavaScript e em REST.

Os três tipos, com os nomes do código

- **Sim ou não (`noul`):** devolve a probabilidade de "sim", sempre entre 0,01 e 0,99.
- **Escolha (`choice`):** escolhe uma opção de uma lista fixa, de 2 a 255 opções.
- **Nota (`score`):** escala de 2 a 255 níveis. A documentação recomenda de 2 a 10 para a maioria dos casos.
- **Confiança:** vem no campo `confidence` de toda resposta.

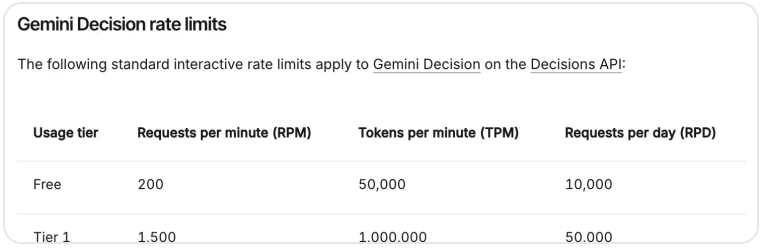
Perguntas no mesmo pedido

Cada pergunta é respondida sozinha: uma não enxerga a resposta da outra. Se uma decisão depende de outra, a documentação sugere fazer dois pedidos.

O pedido aceita até 262.144 tokens, somando o conteúdo e todas as perguntas.

Limites do plano grátis

- **Por minuto:** 200 pedidos e 50.000 tokens.
- **Por dia:** 10.000 pedidos.
- **Contagem:** os limites valem por projeto, não por chave.
- **Quando zera:** o contador diário zera à meia-noite no horário do Pacífico.



Gemini Decision rate limits

The following standard interactive rate limits apply to Gemini Decision on the Decisions API:

Usage tier	Requests per minute (RPM)	Tokens per minute (TPM)	Requests per day (RPD)
Free	200	50,000	10,000
Tier 1	1.500	1.000.000	50.000

Os limites de uso por plano
Fonte: Google AI Studio, Rate limits

06 Fontes

Conferido em 12/11/2026, nas páginas oficiais do Google:

- Decisions API, documentação
- Preços
- Limites de uso
- Chaves de API
- Anúncio no blog do Google, 11/11/2026

Tem guia novo a cada vídeo.
Todos em marraads.online · siga @marraads