// DOCUMENTAÇÃO

Manual de Monitoramento da API Gemini

Versão 1.0 · Julho de 2026

Este manual tem como objetivo apresentar os principais recursos de monitoramento da API Gemini disponíveis no Google AI Studio, explicando como interpretar os indicadores de consumo, acompanhar os Rate Limits, identificar possíveis erros e compreender o funcionamento dos modelos utilizados pela aplicação.

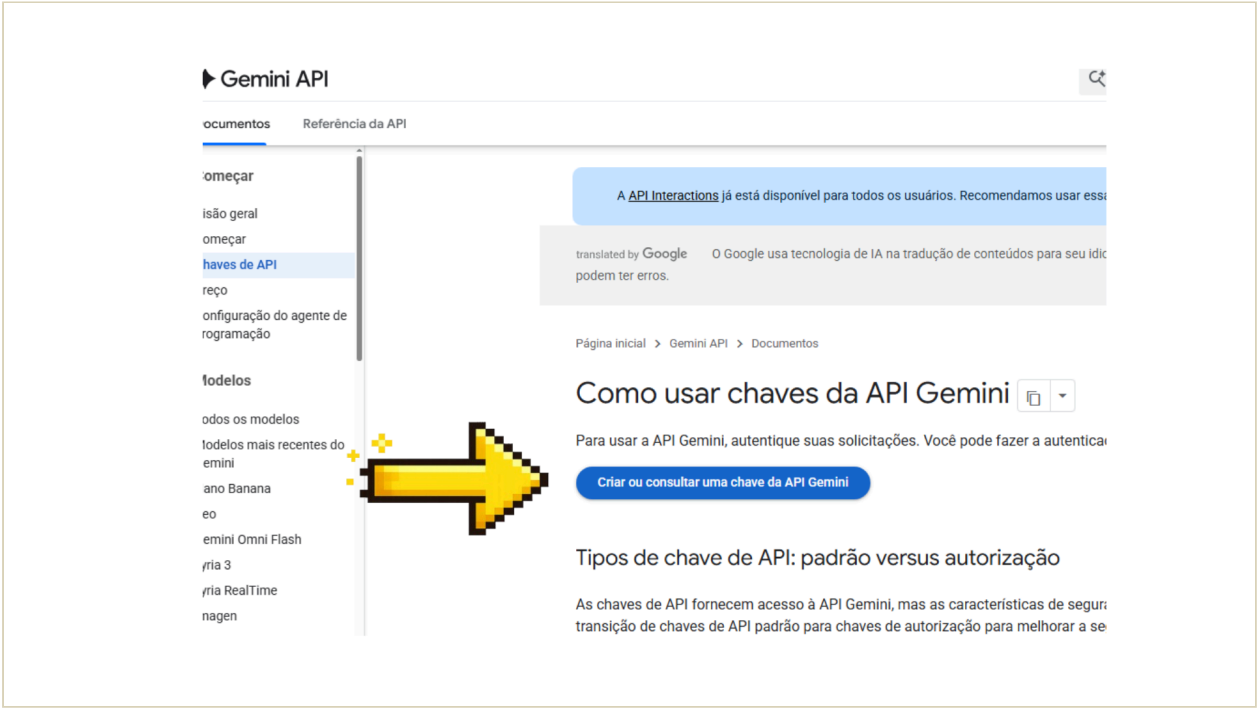
O conteúdo foi elaborado com base em um caso real de implementação e tem como finalidade facilitar futuras consultas, manutenções e a evolução de projetos que utilizam inteligência artificial.

Sumário

Acompanhando o consumo da IA (Gemini)

Para acompanhar o consumo da IA utilizada pelo seu projeto, acesse a página da API Gemini:

1. Abra o endereço da documentação da API Gemini.
2. No **menu lateral**, acesse **Chaves de API**.
3. Clique no botão **"Criar ou consultar uma chave da API Gemini"**.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

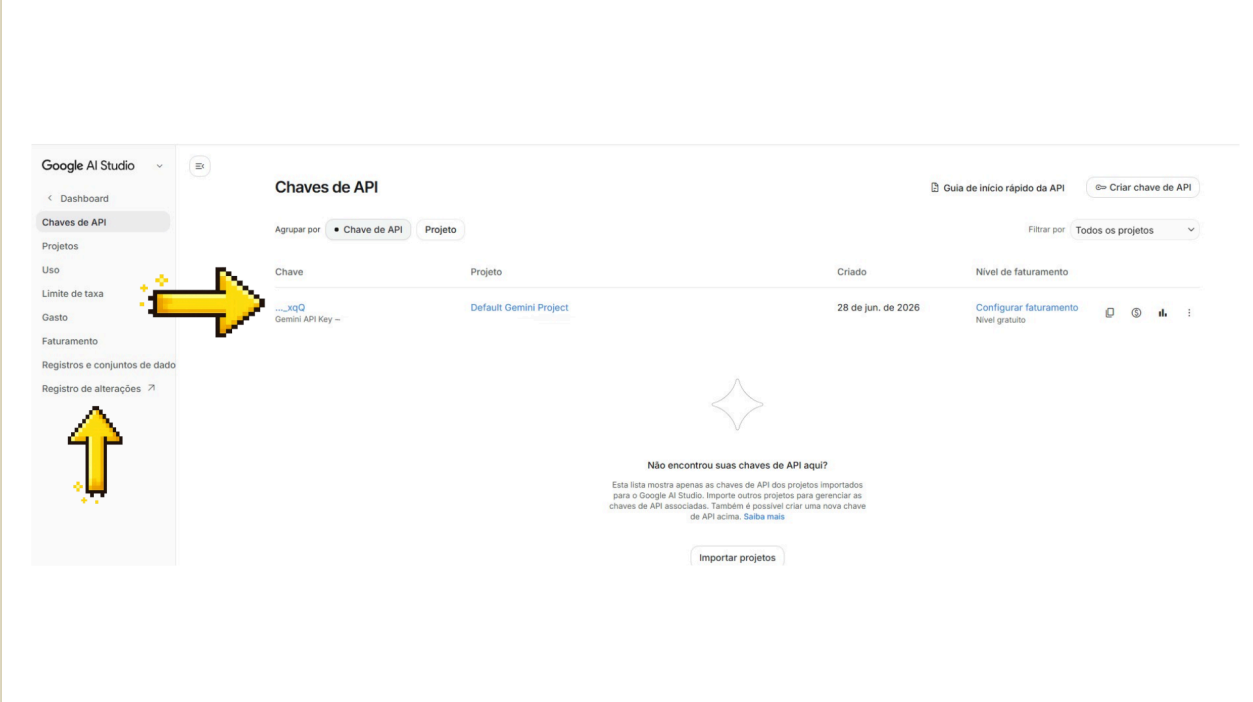
Esse botão abrirá a página do Google AI Studio, onde ficam cadastradas as chaves da API e onde também é possível acompanhar o consumo e os limites de utilização.

Acessando o painel da API

Após clicar em **“Criar ou consultar uma chave da API Gemini”**, você será direcionado ao Google AI Studio, onde estarão listadas as suas chaves de API.

Nesta tela serão exibidas todas as chaves cadastradas. Localize a chave do seu projeto e clique no ícone de estatísticas (▯) ao lado dela, ou selecione as opções pelo menu lateral esquerdo.

Essa opção abrirá o painel de monitoramento da API, onde será possível acompanhar o consumo, os limites de utilização e possíveis erros registrados durante o uso da inteligência artificial.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

Tela Uso

No menu lateral, clique em **Uso**.

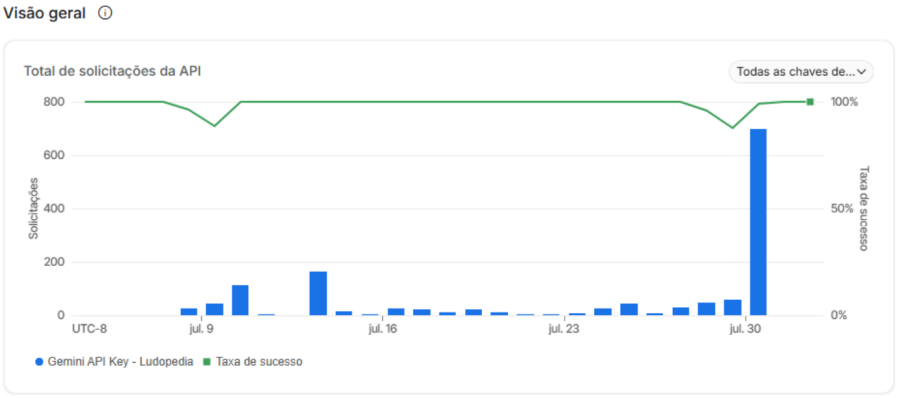
Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

Nesta tela é possível acompanhar o histórico de utilização da API, incluindo o consumo, a quantidade de solicitações realizadas e possíveis erros registrados. A seguir, veremos o significado de cada gráfico apresentado.

Total de solicitações da API

Este gráfico apresenta a quantidade de requisições realizadas para a API ao longo do período selecionado. A **linha verde** representa a taxa de sucesso das solicitações, enquanto as **barras azuis** indicam o número de chamadas realizadas em cada dia.

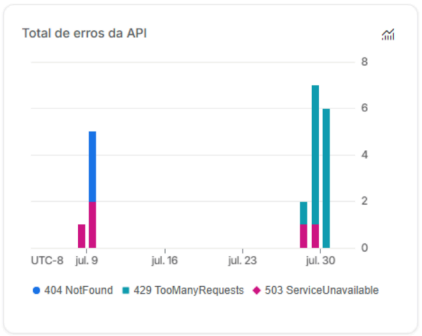
Por meio dele, é possível identificar dias de maior utilização e verificar se as requisições estão sendo processadas com sucesso.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

Total de erros da API

Este gráfico exibe os erros retornados pela API durante o período selecionado. Cada cor representa um tipo de erro diferente, permitindo identificar rapidamente possíveis problemas, como excesso de requisições, recursos não encontrados ou indisponibilidade temporária do serviço.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

A seguir, veremos o significado dos principais códigos de erro apresentados.

Códigos de erro da API

404 - Not Found

O recurso solicitado não foi encontrado. Esse erro pode ocorrer quando a aplicação tenta acessar um modelo, arquivo ou endereço inexistente ou incorreto.

Possíveis causas:

- Nome do modelo digitado incorretamente.
- Endpoint incorreto.
- Arquivo inexistente.

429 - Too Many Requests

O limite de utilização da API foi atingido. Esse é um dos erros mais comuns durante testes ou importações em massa e indica que a aplicação realizou mais solicitações do que o permitido pela cota atual (por minuto ou por dia).

Possíveis soluções:

- Aguardar alguns instantes antes de tentar novamente.
- Reduzir a quantidade de solicitações simultâneas.
- Habilitar o faturamento (Pay as You Go) para aumentar os limites da API.

503 - Service Unavailable

O serviço da API estava temporariamente indisponível. Geralmente é um problema momentâneo nos servidores da Google e costuma ser resolvido automaticamente após alguns instantes.

Possíveis soluções:

- Aguardar alguns segundos e realizar uma nova tentativa.
- Implementar novas tentativas automáticas (retry) na aplicação.

> **Importante:** a ocorrência desses erros não significa necessariamente que a aplicação parou de funcionar. Em muitos casos, uma nova tentativa é suficiente para que a solicitação seja processada com sucesso. Isso é especialmente comum nos erros 429 e 503.

API Live e gerar conteúdo

Esta seção apresenta o consumo da API pelos modelos utilizados, permitindo acompanhar a quantidade de informações processadas e o número de solicitações realizadas.

Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

Os gráficos mostram:

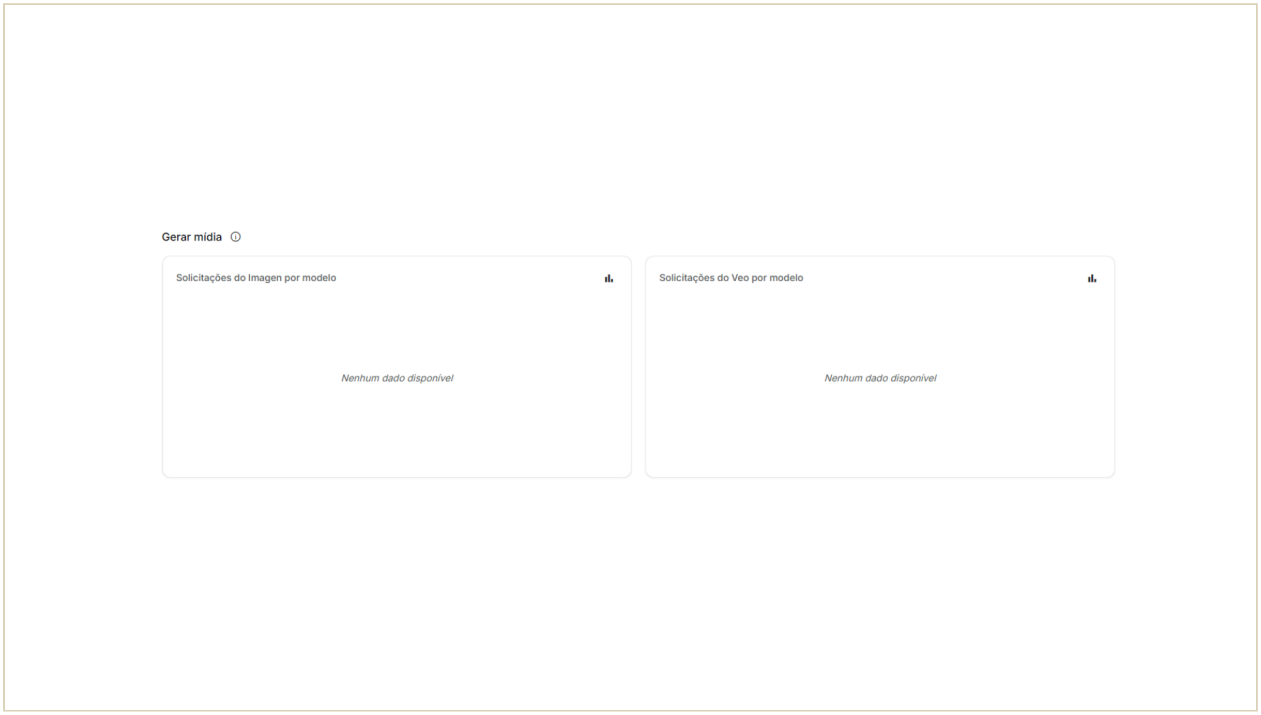
- **Tokens de entrada por modelo:** quantidade de informações enviadas para a IA.
- **Tokens de saída por modelo:** quantidade de informações geradas pela IA como resposta.
- **Solicitações por modelo:** número de chamadas realizadas para cada modelo durante o período selecionado.

Esses gráficos ajudam a identificar qual modelo está sendo mais utilizado e como está o consumo da API ao longo do tempo.

Gerar mídia

Esta seção apresenta o histórico de utilização dos modelos de geração de mídia da API Gemini.

Os gráficos exibem a quantidade de solicitações realizadas para geração de imagens e vídeos, permitindo acompanhar o consumo desses recursos quando utilizados pela aplicação.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

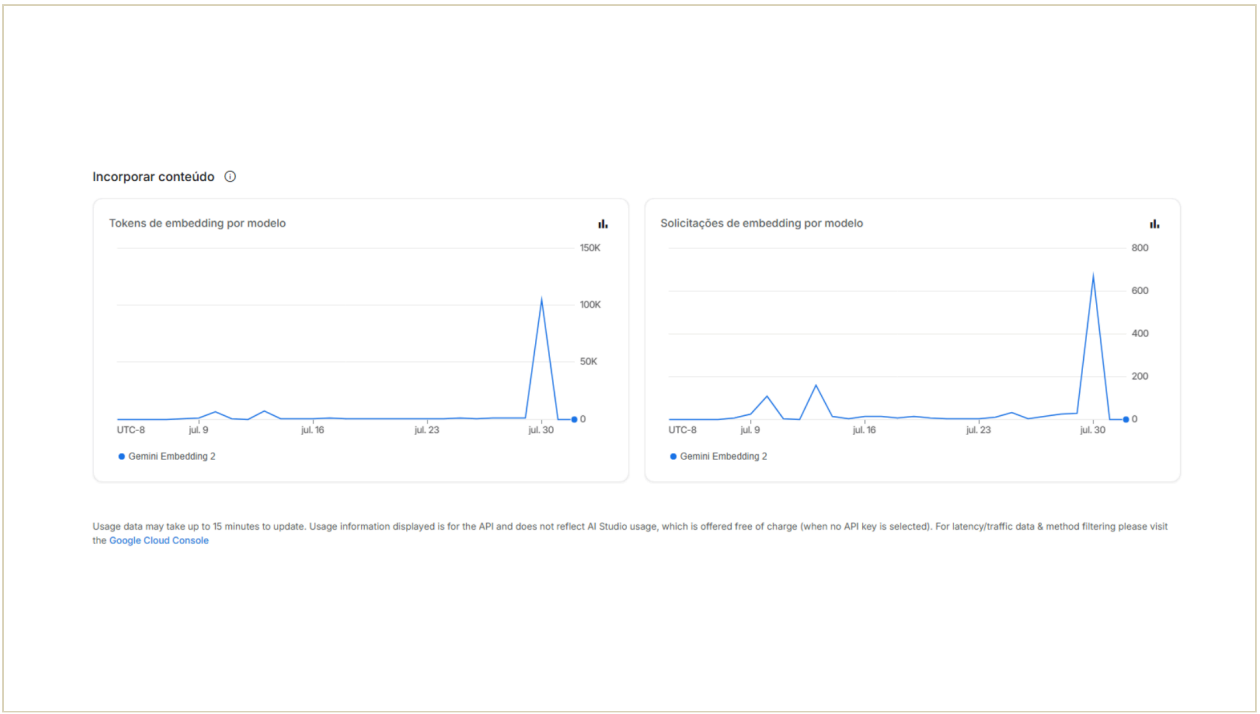
No momento, como esses recursos não foram utilizados, os gráficos permanecem sem informações.

- **Solicitações do Imagen por modelo:** exibe a quantidade de solicitações realizadas ao Imagen, modelo responsável pela geração de imagens por inteligência artificial.
- **Solicitações do Veo por modelo:** exibe a quantidade de solicitações realizadas ao Veo, modelo responsável pela geração de vídeos por inteligência artificial.

Incorporar conteúdo (Embeddings)

Esta seção apresenta o consumo relacionado à geração de **embeddings**, utilizados para transformar textos em representações numéricas que

permitem à inteligência artificial compreender o significado e a relação entre diferentes informações.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

Como exemplo, em uma locadora de jogos, ao gerar embeddings para a descrição, mecânicas e categorias dos jogos, a IA consegue identificar títulos semelhantes e realizar recomendações mais inteligentes, mesmo quando as palavras utilizadas são diferentes.

Os gráficos exibem:

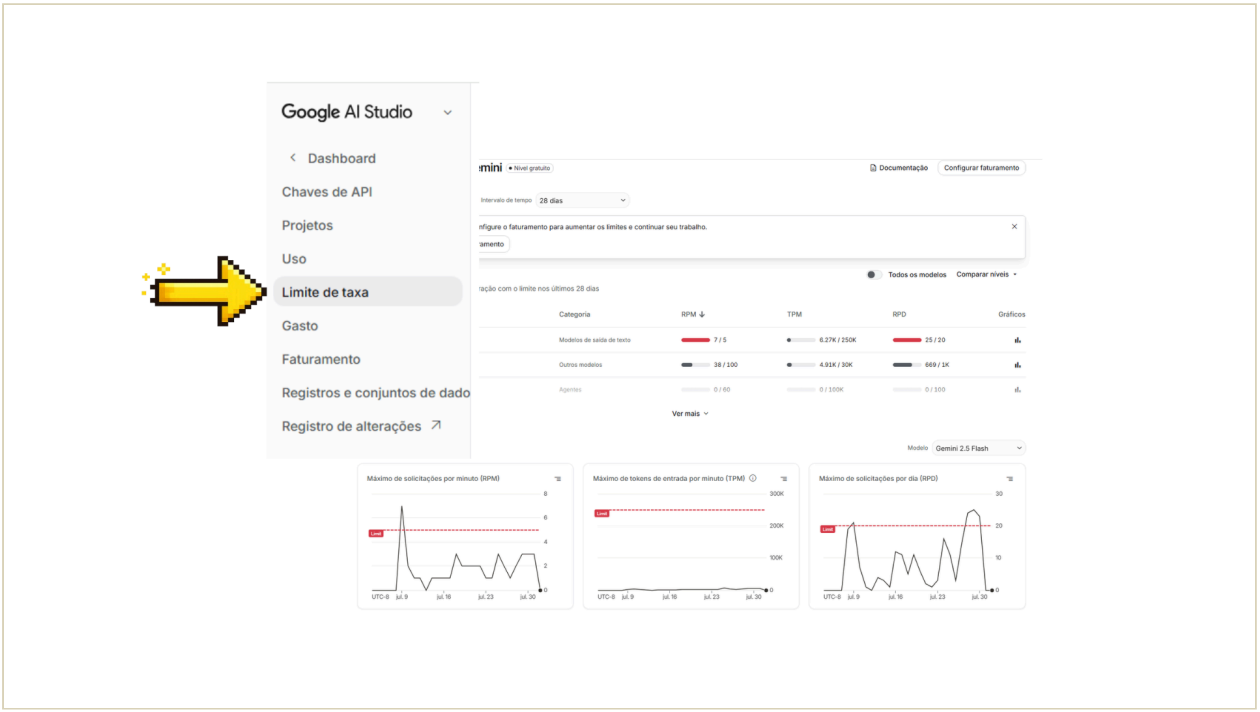
- **Tokens de embedding por modelo:** apresenta a quantidade de tokens processados pelo modelo durante a geração dos embeddings. Quanto maior o volume de informações convertidas, maior será o consumo de tokens.
- **Solicitações de embedding por modelo:** exibe a quantidade de chamadas realizadas ao modelo de embeddings durante o período selecionado. Esse gráfico é útil para acompanhar importações em massa ou atualizações do banco de dados.

Observação: os dados apresentados podem levar até 15 minutos para serem atualizados. Este painel exibe apenas o consumo realizado por meio da API Gemini, utilizando uma chave de API. Os testes realizados diretamente no Google AI Studio (Playground) não são contabilizados nesses gráficos. Para

análises mais detalhadas de latência e tráfego, é possível utilizar o Google Cloud Console.

Tela Limite de taxa

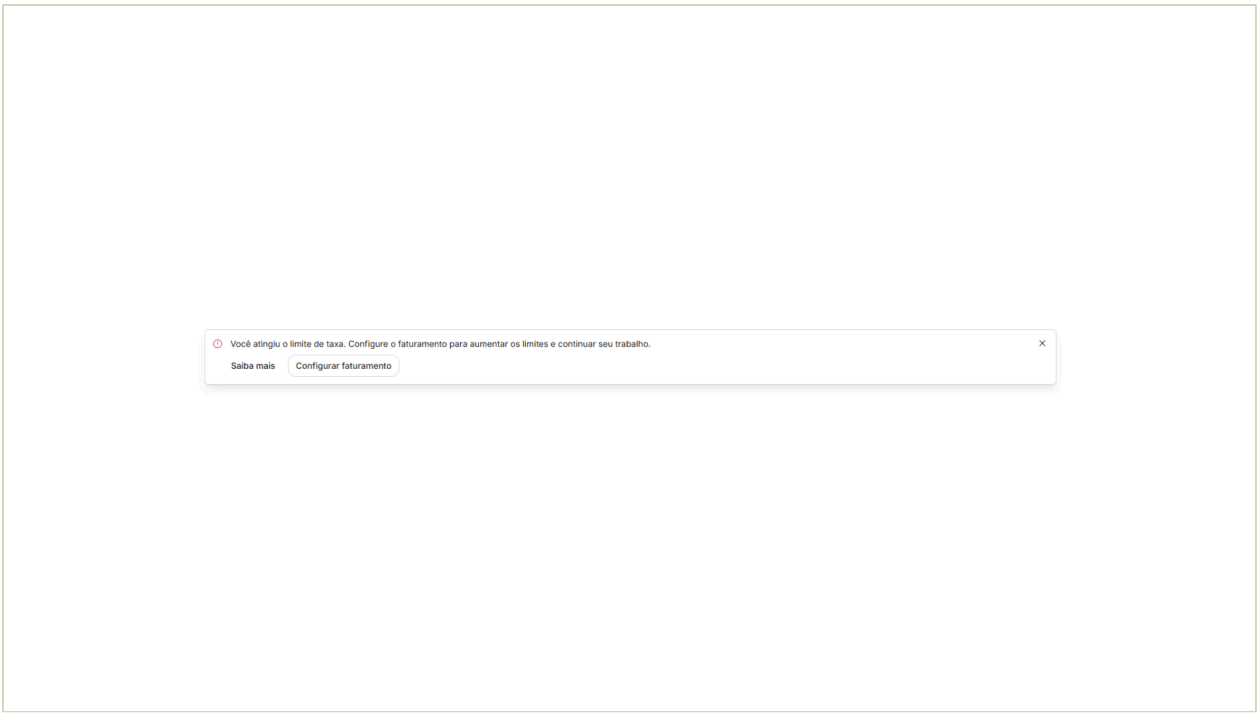
No menu lateral, clique em **Limite de taxa**.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

Esta tela apresenta os limites de utilização da API para cada modelo disponível. É por meio dela que é possível acompanhar quanto da cota está sendo utilizada e identificar quando algum limite é atingido.

Para facilitar o entendimento, este manual utiliza como exemplo uma conta no **nível gratuito**, na qual alguns limites foram propositalmente excedidos durante os testes. Assim, é possível visualizar como o Google sinaliza esses casos e compreender o impacto de cada limite.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

> **Importante:** atingir um limite não significa necessariamente que a aplicação deixará de funcionar completamente. Cada limite controla um aspecto diferente da utilização (como quantidade de solicitações por minuto ou por dia). Dependendo do limite atingido, apenas novas solicitações daquele modelo poderão ser temporariamente recusadas, enquanto as demais funcionalidades e modelos continuam operando normalmente.

Nos próximos tópicos, veremos o significado de cada indicador apresentado nesta tela.

Curiosidade: o que é Rate Limit?

Ao utilizar serviços de inteligência artificial, você encontrará com frequência o termo **Rate Limit** (em português, limite de taxa).

O Rate Limit é um mecanismo utilizado pelos provedores de API para controlar a quantidade de solicitações que cada aplicação pode realizar em um determinado período de tempo.

Seu objetivo é garantir a estabilidade do serviço, evitar sobrecarga nos servidores e assegurar que todos os usuários tenham acesso aos

recursos disponíveis.

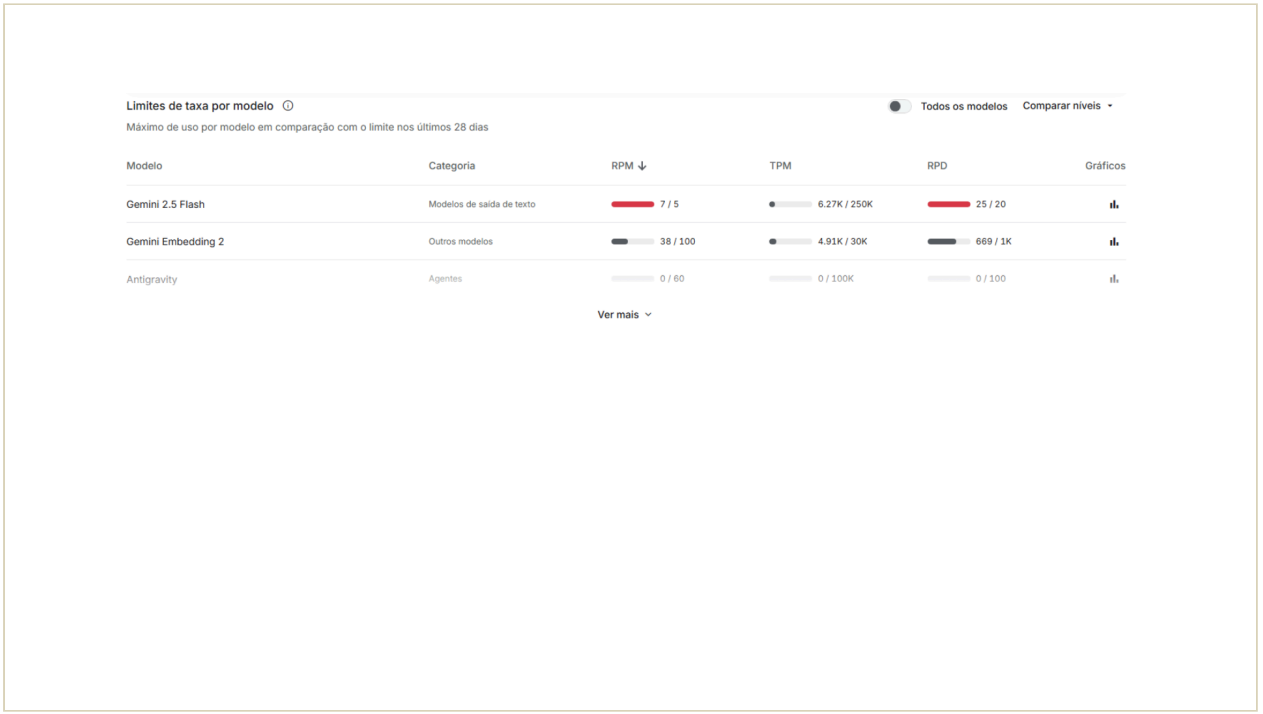
Por esse motivo, é comum encontrar limites como:

- **RPM (Requests Per Minute):** quantidade máxima de solicitações por minuto.
- **TPM (Tokens Per Minute):** quantidade máxima de tokens processados por minuto.
- **RPD (Requests Per Day):** quantidade máxima de solicitações por dia.

Ao longo deste manual, esses indicadores serão apresentados em detalhes. Familiarizar-se com o conceito de Rate Limit facilitará a interpretação dos painéis de monitoramento da API Gemini.

Limites de taxa por modelo

Nesta seção são exibidos os modelos utilizados pelo projeto e seus respectivos limites de utilização.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

No exemplo deste manual são apresentados os dois modelos utilizados pela aplicação:

- **Gemini 2.5 Flash:** responsável pela geração de respostas em texto.

- **Gemini Embedding 2:** responsável pela geração dos embeddings utilizados para busca por similaridade e recomendações.

Caso a aplicação passe a utilizar outros modelos da API Gemini, basta clicar em **“Ver mais”** para visualizar todos os modelos disponíveis e seus respectivos consumos.

Modelo

Indica o modelo da IA ao qual os limites se referem. Cada modelo possui cotas independentes. Assim, atingir o limite de um modelo não significa que os demais deixarão de funcionar.

Categoria

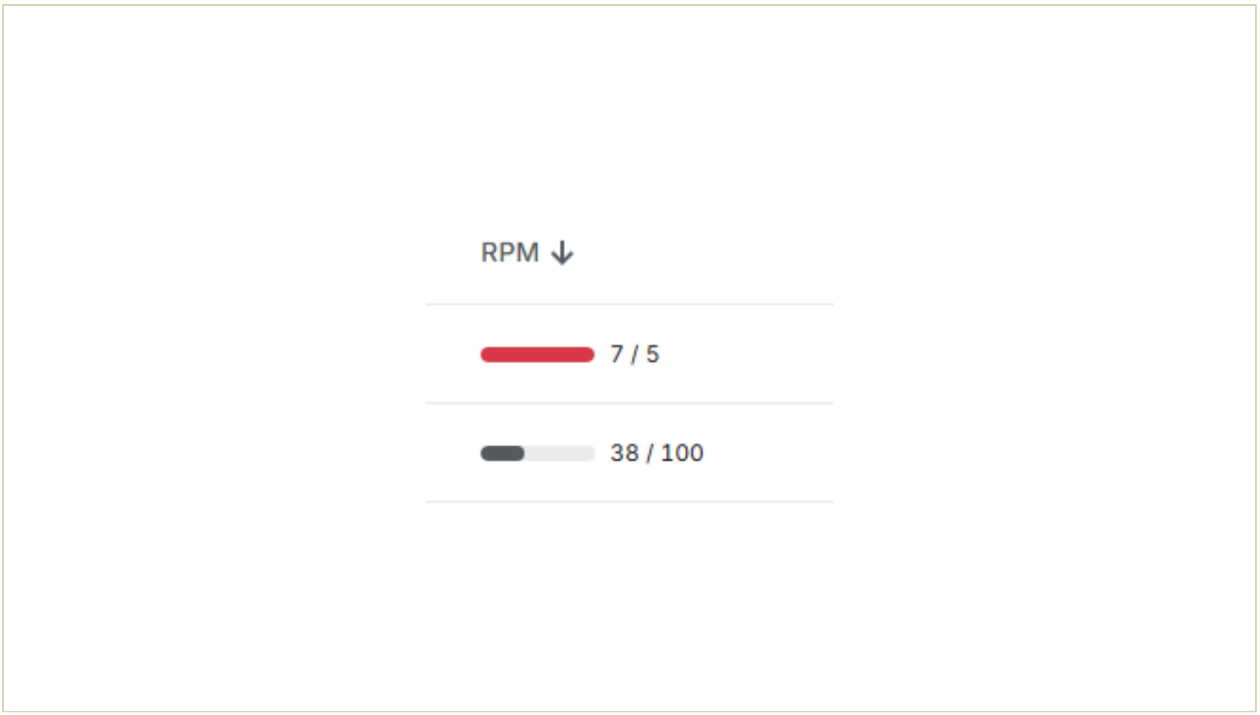
Informa a finalidade do modelo. No exemplo:

- **Modelos de saída de texto:** geração de respostas da IA.
- **Outros modelos:** modelos especializados, como geração de embeddings.

> **Importante:** os valores apresentados nesta tela representam o **maior consumo registrado** dentro do período selecionado (por exemplo, os últimos 28 dias), e não o consumo atual. Assim, mesmo que os limites já tenham sido renovados, os maiores picos continuarão sendo exibidos até que saiam do intervalo de tempo selecionado.

RPM (Requests Per Minute)

Representa o número máximo de solicitações permitidas por minuto.

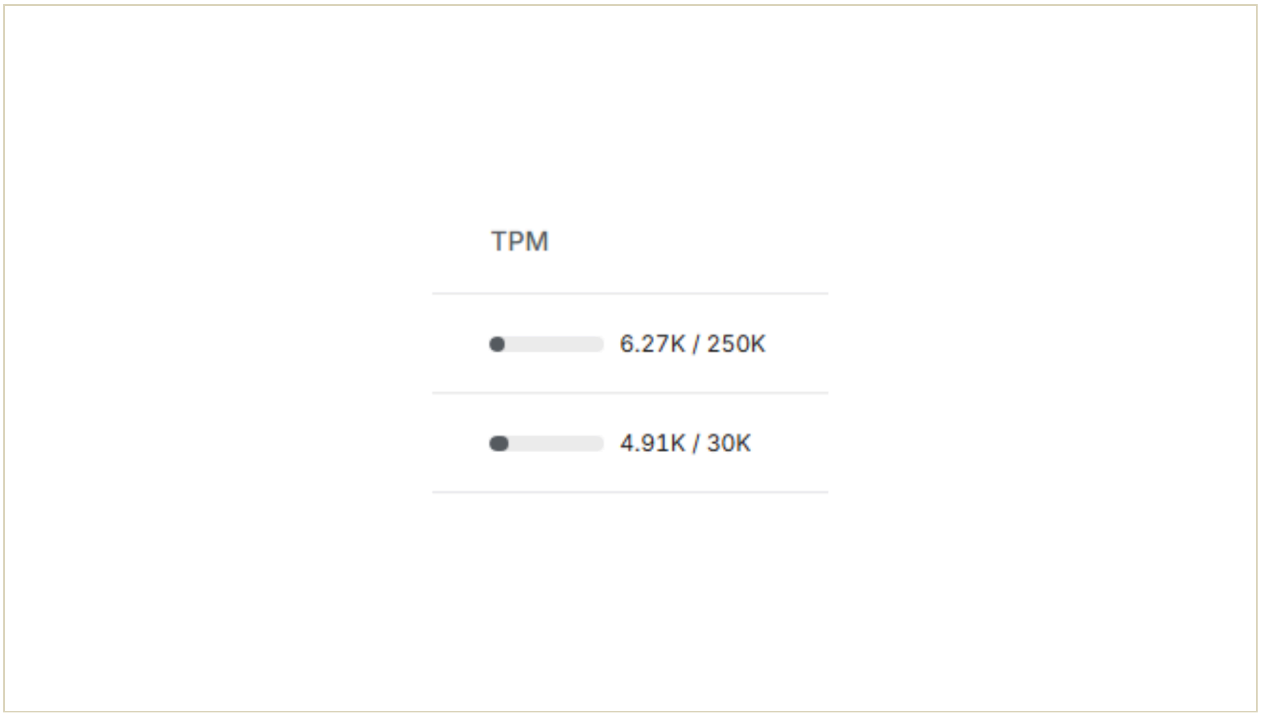


No exemplo, o Gemini 2.5 Flash apresenta **7 / 5**. Isso significa que, em determinado momento, foram realizadas 7 solicitações em um minuto, enquanto o limite da conta gratuita era 5 solicitações por minuto.

Ao atingir esse limite, novas solicitações podem receber o erro **429 (Too Many Requests)** até que o período de um minuto seja renovado.

TPM (Tokens Per Minute)

Representa a quantidade máxima de tokens que podem ser processados por minuto.



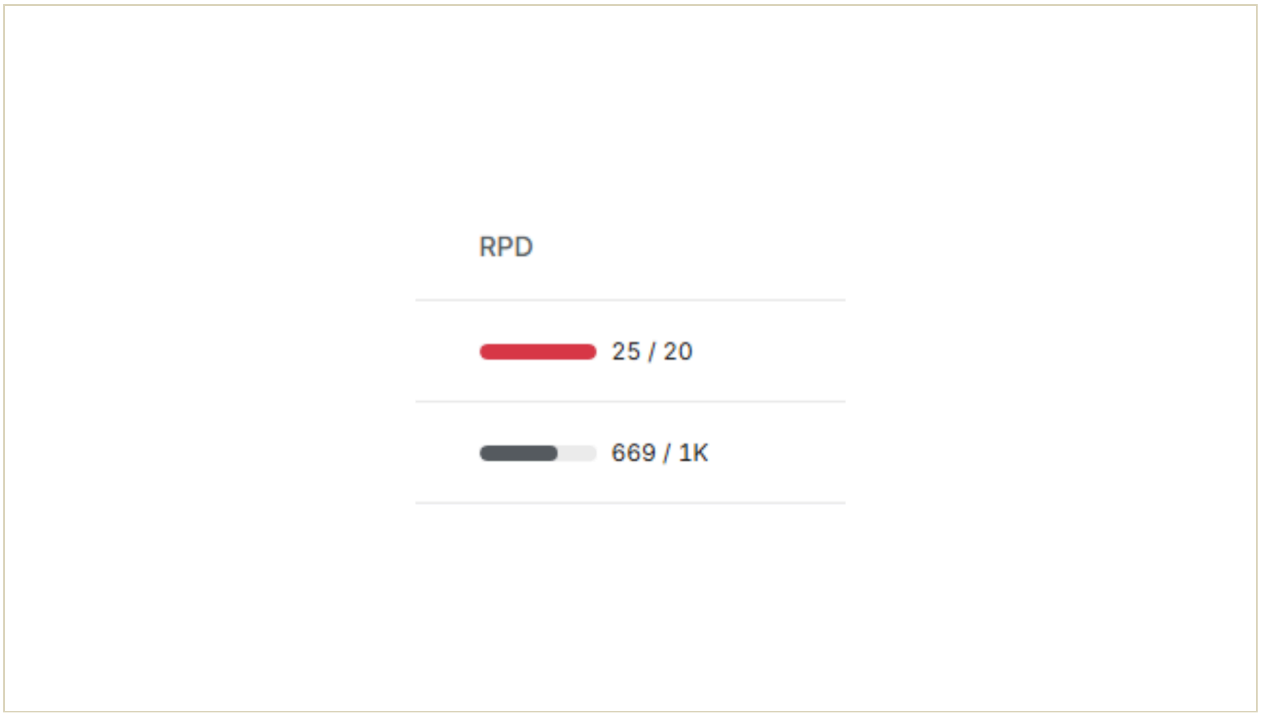
No exemplo:

- **Gemini 2.5 Flash:** 6,27K / 250K
- **Gemini Embedding 2:** 4,91K / 30K

Isso demonstra que, embora o limite de solicitações tenha sido atingido em alguns momentos, o consumo de tokens permaneceu muito abaixo da cota disponível.

RPD (Requests Per Day)

Representa a quantidade máxima de solicitações permitidas por dia.



No exemplo, o Gemini 2.5 Flash apresenta **25 / 20**. Isso indica que foram realizadas 25 solicitações em um dia cuja cota gratuita permitia 20.

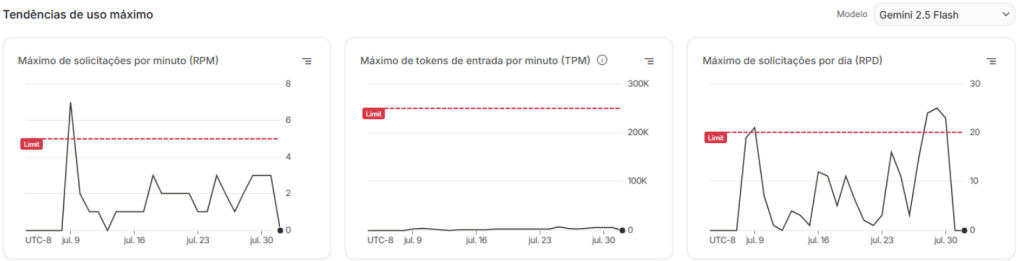
Ao atingir esse limite, novas chamadas ao modelo poderão retornar o erro **429 (Too Many Requests)** até que a cota diária seja renovada.

Gráficos

O ícone de gráfico permite visualizar o histórico detalhado de utilização daquele modelo específico, facilitando a análise do consumo e a identificação de períodos de maior utilização.

Tendências de uso máximo

Esta seção apresenta os maiores picos de utilização registrados para o modelo selecionado durante o período analisado. Os gráficos permitem identificar em quais momentos a aplicação se aproximou ou ultrapassou os Rate Limits definidos para a API.

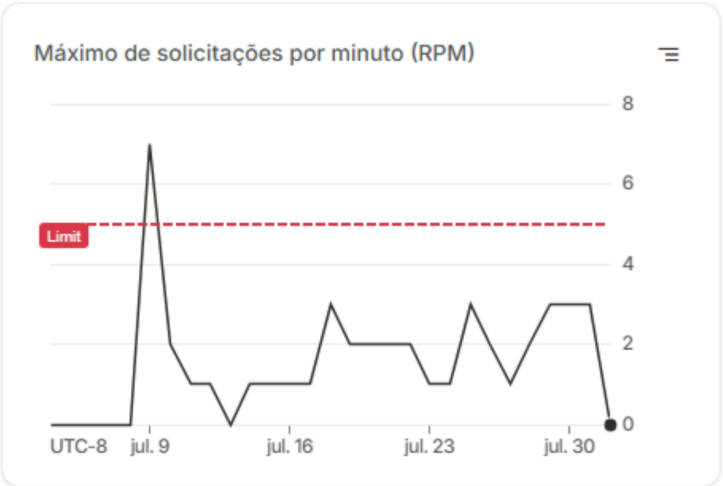


Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

A **linha tracejada em vermelho** representa o limite da conta, enquanto a linha do gráfico mostra o consumo registrado ao longo do período.

Máximo de solicitações por minuto (RPM)

Este gráfico apresenta o maior número de solicitações realizadas em um único minuto durante o período selecionado.



No exemplo apresentado, o pico de utilização ocorreu durante a fase de desenvolvimento da aplicação, quando diversos testes foram realizados em sequência. Em um dos momentos, foram efetuadas 7 solicitações em um único minuto, ultrapassando o limite gratuito de 5 solicitações por minuto.

Apesar desse pico, a API gratuita demonstrou um bom desempenho durante todo o desenvolvimento, atendendo normalmente às necessidades do projeto. O limite foi ultrapassado apenas em momentos específicos de testes intensivos, situação pouco comum durante o uso normal da aplicação.

Máximo de tokens de entrada por minuto (TPM)

Este gráfico apresenta a maior quantidade de tokens de entrada enviados para a IA em um único minuto durante o período selecionado.

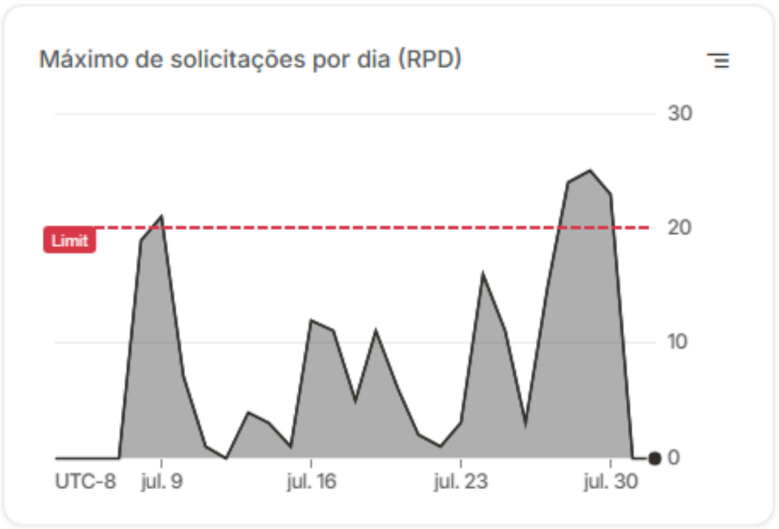
Tokens representam as informações enviadas ao modelo, como perguntas, instruções, descrições e demais conteúdos processados pela inteligência artificial.

No exemplo apresentado, o consumo permaneceu muito abaixo do Rate Limit disponível durante todo o período, demonstrando que, mesmo durante os testes intensivos da aplicação, o volume de informações enviadas à IA foi pequeno em relação à capacidade oferecida pela API gratuita.

Isso indica que, para este projeto, o limite de TPM não representa uma preocupação, sendo muito mais provável atingir os limites relacionados à quantidade de solicitações (RPM e RPD) do que ao volume de tokens processados.

Máximo de solicitações por dia (RPD)

Este gráfico apresenta o maior número de solicitações realizadas em um único dia durante o período selecionado.



No exemplo apresentado, houve alguns dias em que o número de solicitações ultrapassou o limite diário da conta gratuita. Esses picos ocorreram durante a fase de desenvolvimento da aplicação, quando diversos testes e validações foram realizados para garantir o correto funcionamento da inteligência artificial.

Em um cenário de utilização normal, com usuários realizando consultas ao sistema, a tendência é que o consumo diário seja significativamente menor do que o registrado durante o desenvolvimento. Caso o volume de acessos aumente ou a aplicação passe a atender um grande número de usuários simultaneamente, pode ser interessante habilitar o **Pay as You Go** para ampliar os Rate Limits disponíveis.

Ferramentas

Esta seção apresenta o consumo de recursos adicionais disponibilizados pela API Gemini, como **Embasamento da pesquisa (Grounding)** e **Embasamento do mapa**.

Esses recursos permitem que a IA consulte fontes externas para fundamentar suas respostas, aumentando a precisão em determinados tipos de consultas.

No momento, esses recursos não são utilizados pela aplicação, por isso os gráficos permanecem sem dados.

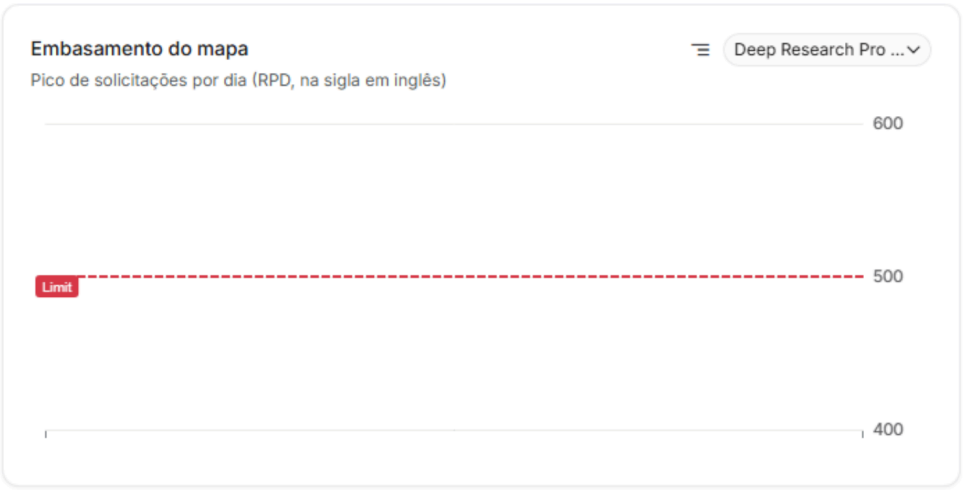
Embasamento da pesquisa

Apresenta o histórico de utilização do recurso de Grounding com Pesquisa, que permite à IA consultar informações na web para complementar suas respostas.

Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

Embasamento do mapa

Apresenta o histórico de utilização do recurso de Grounding com Mapas, utilizado quando a IA consulta informações relacionadas a locais, estabelecimentos e dados geográficos.



Captura de tela do site oficial GEMINI API feita no dia 31/07/2026

> **Observação:** caso esses recursos sejam utilizados futuramente, os gráficos passarão a exibir o consumo e permitirão acompanhar seus respectivos Rate Limits.

Considerações finais

Esta documentação foi elaborada com base nos recursos atualmente utilizados pela aplicação. Por esse motivo, alguns painéis do Google AI Studio podem não apresentar informações, seja porque o recurso ainda não foi utilizado ou porque ele não faz parte da arquitetura atual do projeto.

O Google atualiza constantemente a API Gemini, disponibilizando novos modelos, funcionalidades e ferramentas. Da mesma forma, novas necessidades da aplicação podem surgir ao longo do tempo, tornando necessário o uso de recursos que hoje não estão presentes neste manual.

À medida que novos recursos forem incorporados ao projeto, novas documentações serão produzidas, explicando seu funcionamento, finalidade e

forma de monitoramento, mantendo este material sempre atualizado e útil para futuras consultas.

Consulta Rápida

Termo	Significado
Rate Limit	Limite de utilização imposto pela API para controlar o consumo de recursos.
RPM (Requests Per Minute)	Quantidade máxima de solicitações permitidas por minuto.
TPM (Tokens Per Minute)	Quantidade máxima de tokens processados por minuto.
RPD (Requests Per Day)	Quantidade máxima de solicitações permitidas por dia.
Token	Unidade utilizada pela IA para medir o volume de texto enviado e recebido.
Prompt	Instrução ou pergunta enviada para a inteligência artificial.
Embedding	Representação numérica de um texto, utilizada para busca por similaridade e recomendações inteligentes.
Gemini 2.5 Flash	Modelo utilizado para gerar respostas em texto.
Gemini Embedding 2	Modelo utilizado para gerar embeddings.
Grounding	Recurso que permite à IA consultar fontes externas (como pesquisa ou mapas) para fundamentar suas respostas.
API Key	Chave utilizada para autenticar a aplicação junto à API Gemini.
Playground (AI Studio)	Ambiente do Google AI Studio utilizado para testes manuais com os modelos de IA.

Termo	Significado
Pay as You Go	Modalidade de cobrança baseada no consumo da API, que amplia os Rate Limits da conta.
404 - Not Found	O recurso solicitado não foi encontrado.
429 - Too Many Requests	Algum Rate Limit foi atingido.
503 - Service Unavailable	A API estava temporariamente indisponível.
Imagen	Modelo da Google para geração de imagens por IA.
Veo	Modelo da Google para geração de vídeos por IA.
Dashboard	Painel do Google AI Studio para acompanhamento do consumo, Rate Limits e estatísticas da API.

Versão 1.0 • Julho de 2026 — Laura Lopes | CoreDot