

AI Assistant App Widget Processing and Costs_v01

I have a couple different questions. Let's start with Claude Pro: 1) What are the parameters or boundaries that I can use my Claude Pro Account with? On almost all apps nowadays, there is a am AI Assistant Icon in the lower right corner. Can I do this same thing with my existing Claude Pro Account (if I have a consumer App?) And how about if I have a public website?

The short version: your Claude Pro account is for *you* to use Claude through Anthropic's own apps. It can't power a chat widget in your app or on your website. That requires a separate product, the Claude API.

What Pro covers

Claude paid plans give subscribers access to Claude on the web, desktop, and mobile, and offer enhanced features like more usage and priority access during high-traffic periods. Pro also covers tools like Claude Code, Claude in Chrome, and the Excel/PowerPoint add-ins. It's a single-person seat, though. Sharing your login or routing other people's questions through your account isn't allowed under the consumer terms. [claude](#)

Embedding an assistant in an app or on a website

This is what the API is for. The Claude Console is Anthropic's developer platform providing API keys and access to Claude models for building applications and integrations. A paid Claude subscription enhances your chat experience but doesn't include access to the Claude API or Console. If you want both, you'll need to keep the paid Claude plan and separately set up Console access for API usage. [claudeclaude](#)

It works like this:

- You create an account at [console.anthropic.com](#), add a payment method, and get an API key.
- Billing is pay-as-you-go by tokens (roughly, the amount of text going in and out). You aren't paying a flat monthly fee, so set a spending cap in the Console.
- The chat bubble in the lower-right corner is something you or a developer builds, or you get it from a third-party chat-widget service that lets you plug in your own Anthropic API key. Many do.

- The API key has to stay on your server, never in the website's front-end code. Otherwise anyone could take it and run up charges on your account.
- You'd write a system prompt that sets what the assistant knows, how it behaves, and what it won't discuss.

For a public site, you'd also want rate limiting so one visitor can't burn through your budget, and a privacy notice saying visitors are chatting with an AI.

This is probably the route you'd need for UrbanFamilyUSA, since giving clients AI assistant access is central to that project. Your Pro account stays your personal workbench, and the API handles anything the public touches.

What's your second question?

My second question is related to the first. I had assumed there would be a separate service charge for putting an AI widget on a public App because I couldn't imagine how this could happen without some kind of clock running. So my 2nd question is, what happens when the AI Assistant reaches the "Rate Limit? Does it shut down, throttle, ???

Searched the web

Your instinct was right. With the API, every conversation on your widget uses tokens, and you're billed per token. There's also more than one kind of "limit," and they behave differently.

In every case, Anthropic doesn't slow your widget down gradually. It turns away the individual request that crosses the line. What your visitor sees depends on how the widget was built.

1. Rate limits (the speed limit)

These cap how much traffic you can send per minute. They're measured in requests per minute, input tokens per minute, and output tokens per minute. If you exceed any of them, you get a 429 error describing which limit was exceeded, along with a retry-after header saying how long to wait. Capacity refills continuously rather than resetting on a schedule, so a rate-limit error usually clears within seconds. You'd mostly hit this during a burst, like a bunch of people chatting at once. A well-built widget waits and retries automatically, and the visitor just notices a slightly slower reply. A poorly built one shows an error or goes silent.

2. Spend limits (the monthly budget)

Each usage tier has a limit on how much you can spend on the API each calendar month. Once you reach it, you have to wait until the next month or qualify for the next tier before you can use the API again. You can also set your own lower cap in the Console, which I'd recommend for a public site. When usage reaches a spend limit you set, requests return an error that says when access resumes. You can raise or remove the limit to restore access sooner. This is the real shutoff. Until you raise the cap or the month rolls over, your widget stops answering.

3. Running out of prepaid credit

The API generally works on prepaid credits. If the balance hits zero, requests fail the same way. The Console has an auto-reload option if you'd rather it top up automatically.

4. Anthropic being overloaded

There's also a separate error that isn't a rate limit at all. The 529 overloaded error fires when Anthropic's own capacity is saturated, regardless of your tier. It's rare and temporary, and the fix is the same as for rate limits: wait and retry.

The practical takeaway

The shutoff behavior works in your favor as a safety net, because a runaway bot or a viral day can't generate a surprise five-figure bill if you've set a cap. The part you control is the visitor's experience. Whoever builds the widget should make it retry quietly on rate-limit errors. When the budget is exhausted, it should show a friendly "The assistant is unavailable right now" message with a fallback contact option instead of a broken bubble.

You can also limit how much each visitor can use per day on your end, so one person can't use up everyone's budget.

Good information. Let me mull this over for an hour or two....

Sounds good, take your time. When you come back, I can help you estimate what a widget might actually cost per month based on how many visitors you expect.