



Your Claude bill follows how people use it. **Here's how to bring it down.**

10 habits for lawyers and 6 settings for organization admins.



— WHY THE BILL GROWS

On Enterprise you pay \$20 a seat, plus what each person uses at Anthropic's API rates

- **Claude re-reads the whole conversation** every time you send a message, so a long thread costs more with every question.

- **What Claude writes costs 5x what it reads.**

- **Opus costs 2.5x Sonnet** for the same work.

Usage is counted in tokens, small pieces of words. Everything Claude reads and writes adds to the count.

Source: claude.com/pricing and Anthropic's API pricing (Opus 5 \$5 in / \$25 out, Sonnet 5 \$2 in / \$10 out, per million tokens)



Part 1: **Lawyers**



01 MODEL

Default to Sonnet and switch to Opus on purpose

Opus costs 2.5x Sonnet for the same work, and Sonnet handles a lot of the routine legal work well. Your model choice stays set in later chats, so check the model picker before you start a big task.

Do this:

- **Sonnet:** client emails, letters, first drafts, a summary of a deposition or a set of records.
- **Opus:** work that needs careful reasoning across many facts, like a liability analysis, a brief, or case strategy.
- **Haiku:** simple bulk work, like sorting a folder of documents or listing what's in each one.



02 EFFORT

Turn effort down for routine work

Higher effort means Claude thinks longer, makes more tool calls, and checks more of its work before it answers, and all of that uses tokens. Lower effort means a quicker, more direct pass.

Do this:

- **Lower:** clear tasks you can check in one read, like turning a call transcript into three bullets or pulling names from ten emails.
- **Default:** finding patterns when everything Claude needs is in the documents you gave it.
- **Higher:** high-stakes work that is hard to verify, like a liability analysis or a brief.



03 NEW CHATS

Start a new chat when you switch topics

Every message re-sends the whole conversation, including instructions and tool results you never see in the chat window. A quick question at the end of a long thread can cost more than the same question in a fresh chat.

When a chat gets too long, Claude summarizes it and carries on from the summary. That uses more of your usage, and the summary can occasionally drop details from earlier in the conversation.

Do this: one chat per task. When a thread gets long, ask Claude for a short summary, check that the dates and facts you need are in it, and start a new chat with it.



04 PROJECTS

Put reference documents in a Project

Documents in a Project are cached and don't count against your limits when you reuse them. Uploading the same template into a new chat every time means paying for it every time.

When a Project gets too big for Claude to read in one go, it switches to searching it and reads only the parts that match your question, instead of loading every document each time.

Do this: set up one Project per case type, with the templates, playbooks and style guides for that type of case. Put large case files in a Project and ask specific questions about them.



05 CHAT OR COWORK

Use chat for chat work

Cowork handles a task in many small steps. It opens a file, reads it, decides what to do next, opens the next file, and checks its own work, and every one of those steps costs usage. The same question costs a lot more in Cowork than in chat.

Do this: If you need an answer, a draft, or a summary of one document, chat is enough. Cowork is worth it when the job means going through a folder of files or taking several actions in a row.

Research where Claude clicks through a website in your browser (Claude in Chrome) is billed the same way as Cowork.



05 COWORK SETTINGS

Cowork has three settings for how much it asks you before acting

- **Manual:** it asks you to approve every step.

- **Auto:** it works without asking, but checks each step first. For example, it looks for a document trying to slip it hidden instructions, or for data being sent somewhere it shouldn't go, and blocks anything risky. That extra checking costs extra usage.

- **Skip:** it doesn't ask and doesn't check.



06 CONNECTORS

Turn off connectors and web search you don't need

Tools and connectors are token-intensive. Every web search result, and everything a connector pulls in, like emails or matter records, lands in the conversation and gets re-sent with every message after it.

Do this: switch on the connector the task needs, and switch web search off when the answer is in your own documents. Ask a connector a narrow question, like one matter or one folder, instead of asking it to look through everything.



07 CONTEXT UP FRONT

Batch related questions and give full context up front

When Claude gets something wrong, it's usually because the context it had was missing, wrong or poor. Every correction after that is another round trip that re-reads the whole conversation. Three related questions in one message cost less than three separate messages.

Do this: put the facts, the actual document (not your description of it) and the format you want in your first message. Standing instructions, like your firm's letter format, go in the Project instructions so you don't have to retype them.



08 OUTPUT

Ask for shorter outputs

What Claude writes costs 5x what it reads. A long answer also stays in the conversation and gets re-sent with every message after it, so you keep paying for it. If answers get longer without getting better, your effort setting may be too high.

Do this: tell Claude the length and format you want, like *"five bullets"* or *"one paragraph."* For a summary, list the exact details you want pulled out, like dates, providers and treatment. When you edit a letter, ask for only the changed paragraph, not the whole letter again.



09 SKILLS

Turn repeat work into Skills

A Skill is a saved set of instructions Claude follows for one task. Until Claude uses it, a Skill costs about 100 tokens. In use, it costs under 5,000, and reference files inside it load only when the task needs them.

In an LN Labs study of 40 legal tasks, Claude in Cowork with a legal plugin cost about 80 cents a task, against \$2.80 in plain chat, because a good setup reaches a good answer in fewer tries.

Do this: when the same request keeps coming back, like a demand letter, a medical chronology or a spoliation letter, write it up once as a Skill. Instructions you only need for one task go in a Skill, not in your Project instructions, which Claude reads with every message.



10 PDFS

Handle PDFs carefully in chat

When you attach a PDF to a chat, Claude reads every page as text and as an image, so it can see stamps, signatures and layout. Each page counts as 1,500 to 3,000 text tokens plus the image, and a chat holds 100 images, so one 90-page agreement uses up most of them.

A PDF in a Project's Files is read as text instead, without the page images. Cowork reads the text too.

Do this: put long contracts and record sets in the Project's Files, and don't also attach them to a chat in that Project, or they're read as images again. For a text-only document in a chat, upload a Word version. Keep the PDF when the stamps, signatures or layout matter.



Part 2: Organization admins



11

DEFAULT MODEL

Lock the org default model

In Organization settings > Models, admins choose the model every new conversation starts with, for the whole firm or per role. *"Use Anthropic's recommended default"* moves everyone to each new model automatically. *"Choose a specific model"* stays fixed until you change it. Members can still pick a different model for any conversation.

Do this: choose a specific model and start everyone on Sonnet. Set a default effort level on the same page.



11 NEW RELEASES

A new model can move your bill either way

Sonnet 5 came with a new tokenizer that can count the same text as up to 35% more tokens.

On legal agent work, Harvey found Opus 5 matched Opus 4.8 at its highest reasoning level while using 26% fewer tokens.

Do this: test a new model on your own work before you move the firm onto it. Instead of managing each release by hand, a role can sit on a preset: Frontier for the people who need the newest models, Basic for everyone else.



12 OPUS AND EFFORT

Restrict Opus and cap effort by role

In Organization settings > Models, admins choose which models the firm can use and the highest effort anyone can pick on each one. In Organization settings > Roles, a role can get fewer models and a lower cap. The limits apply in chat, Cowork and Claude for Microsoft 365, but not yet in Claude in Chrome.

Capping effort usually costs less quality than taking a model away. A frontier model at medium or low effort often does better than an older model at high or max effort.

Do this: when steering people isn't working, first cap effort at Medium or High instead of Max for people who aren't power users. Take Opus away from a role only if its use stays heavy after that.



13

SPEND LIMITS

Set spend limits per group, not only a firm-wide cap

A group limit gives every member the same monthly limit, and an individual limit always overrides it, higher or lower. If no firm-wide limit is set, it says unlimited and nothing is capped. Once one is set, it blocks everyone at once when it's hit, with alerts at 75% and 90%. People get a warning as they get close to their own limit and can click "*Request more usage*" to ask an admin. Limits reset on the 1st of each month.

Do this: start with low group limits, because raising a limit is easier than dealing with an overage. Organization settings > Usage shows who is close to their limit and who is blocked right now, so you can top someone up before they tell you. Raise limits before a training day, so nobody gets cut off halfway through.



14 ACCESS

Set access per role once you know how people work

Admins can switch Cowork, web search and connectors on or off per role.

Start with the same access for everyone and learn from the first weeks. The analytics show what each person spends and how much of it goes to chat and to Cowork.

Do this: once you have a good picture of how people use it, give each role the tools its work needs. If someone uses Cowork for simple questions, address it with training before you restrict them.



15 ORG INSTRUCTIONS

Add to your org instructions as you learn

Org instructions are standing instructions Claude follows in every conversation in the firm. They sit in Organization settings and take priority over each person's own instructions.

Set them up at the start, then add to them as you see how people work. When the first weeks show a pattern that costs more than it needs to, one instruction can address it for the whole firm.

Do this: review usage after the first few weeks and add what you find. For example, tell Claude to check with the user before building an HTML dashboard when a simpler format would do. Keep them short and specific. The limit is 3,000 characters, but Claude reads them with every message.



16 ANALYTICS

Check the analytics weekly at first

claude.ai/analytics shows spend per person and per model, the top 10 users by spend, and separate views for chat and Cowork. For each Skill, it shows how often it's used and what one use costs. You can ask it questions in plain language, and export a spend report as a spreadsheet to share with the partners.

Do this: check weekly for the first few weeks, then monthly. Heavy Opus use usually means people never got the model guidance, so fix the guidance before you raise a limit. For each Skill, compare what one run saves with what it costs.



16

SMART REPORTS

Let the reports come to you

Smart reports run on a schedule and tell you what people are actually working on, which connectors and Skills get used, who keeps hitting limits, and which sessions cost the most.

Do this: start with the most expensive sessions and look at what the work was. That tells you whether it was worth it, and what to teach next.



Your law firm's **Claude AI partner**

- A defensible Claude setup, with every setting explained to you
- Training for your team on your own matters
- Your firm's way of working, written into Claude
- A fractional AI partner who stays on after the setup

www.mpladvisors.com

Founded by Rok Popov Ledinski and Lotte Meijs

