THE CRAIG WEISS
GROUP

NO TECHNICAL BACKGROUND NEEDED

# Understanding AI Token Costs

What tokens actually are, how vendors bill you for them, whether you can push back on the price, what quietly makes your bill bigger, and how to keep it in check. No jargon. No degree required.

INSIDE: What is a token? • How pricing works • Negotiating costs • Files, images & video • What drives costs up • How to reduce your bill • A final cheat sheet

# What Is a Token?

The basic unit AI vendors bill you for

A **token** is a small chunk of text. Roughly three-quarters of a word, if you're writing in English. A short word might be one token. A longer one might get chopped into two or three. Every message you send, every word the AI sends back — all of it gets broken into tokens and counted. Every single one.

Vendors bill on tokens. Not on "questions." Not on how many times you hit send. A short question with a long answer can cost more than a long question with a short answer. Nobody tells you that part up front.

**QUICK REFERENCE**

About 100 tokens ≈ 75 words. A typical two-page document runs 1,000-1,500 tokens. A short email reply? Maybe 50-100. Not much at all.

## How Vendors Actually Charge You

- **Priced per million tokens**, split into two rates: **input** (what you send) and **output** (what it writes back). Output usually costs more. Makes sense — generating new text takes more horsepower than just reading yours.
- **Different models, different price tags.** Smaller, faster models are cheap. The big, impressive ones cost more per token — though sometimes they get the job done in fewer tokens, which evens things out. Sometimes.
- **Non-text stuff can get billed separately, too.** Some vendors charge per image, per minute of audio, per second of video. On top of the token count, or instead of it, depending who you ask.

## Can You Negotiate Token Costs?

Short answer: yes. If you know who to ask.

- **Volume or enterprise deals.** Vendors will cut you a better per-token rate if you commit to real spend, monthly or annual. They want the guaranteed revenue. You want the discount. Everybody walks away happy.
- **Usage tiers.** A lot of platforms quietly drop the price per token once you cross certain thresholds. You don't even have to ask for that one.
- **Batch discounts.** Work that doesn't need an instant answer — batched overnight, processed in bulk — sometimes runs cheaper. If you can wait, you can save.
- **Flat-rate alternatives.** Ask if there's a subscription or fixed-price option instead of pure pay-per-token, especially if your usage is high and predictable. It's a fair question. Ask it.

**GOOD TO KNOW**

Negotiation is real when you're talking to an actual sales team about an actual account. Self-serve, individual plans? Fixed rates, published on a page somewhere. Still doesn't hurt to ask. It never does.

## Do Fees Go Up for PDFs, Slides, Images & Video?

Generally, yes. Here's why, plainly.

- **PDFs and PowerPoints** get turned into tokens, same as if you'd typed all that text yourself. A text-heavy document costs about what you'd expect — roughly the same as pasting it in.
- **Images** get converted based on size and detail. Bigger, sharper image, more tokens. Small and simple, fewer. Not complicated.
- **Video and audio** cost the most, by far. They get chewed through frame by frame, second by second. A few minutes of video adds up faster than you'd think.
- **Re-uploading the same file** over and over in one conversation multiplies the cost. It gets reprocessed every time it's part of the conversation, unless the vendor's smart enough to cache it. Not all of them are.

# What Tasks Cost More vs. Less

Side-by-side comparison

The total cost of anything you ask an AI to do comes down to one number: how many tokens it takes to read your request and write back an answer. That's it. Here's what tips the scale.

| COSTS MORE | COSTS LESS |
| --- | --- |
| Long documents or spreadsheets, read start to finish | Short, focused questions with short answers |
| Long, detailed answers you didn't really need | Trimmed source material instead of the whole document |
| Conversations that just keep going (every message drags the whole history with it) | A few well-planned follow-ups instead of a dozen small ones |
| Deep reasoning / "extended thinking" modes | Standard mode, for standard tasks |
| High-res images, video, or audio | Compressed or lower-res images when detail doesn't matter |
| Pulling out the most advanced model for a task that didn't need it | A smaller, faster model for the simple stuff |

**GOOD TO KNOW**

Conversation length matters more than people think. Most AI systems re-send the whole conversation every single time you type something new. An hour-long chat can cost more per message than a fresh one — even if the question itself is simple. Even if it's basically nothing.

# The Vague Question Problem

Why starting broad quietly costs more

Somebody opens with "help me with this" or "can you look at my project?" and here's what happens next: three, four, five rounds of back-and-forth before the AI even understands what's being asked. Every one of those messages drags the whole growing conversation along with it. Tokens add up. Nothing useful happened yet.

This is probably the single most common way people overspend. And it's invisible. It doesn't feel expensive in the moment. It just quietly stacks up, one clarifying question at a time.

## How to Fix It

- **Front-load everything in message one.** The goal, the audience, the format, the length — all of it, all at once. Not trickled in over four separate messages.
- **Use the formula:** WHO it's for + WHAT you want + FORMAT you need. One well-written prompt beats five vague ones. Every time.
- **Genuinely don't know what you need yet?** Fine. Say that. Ask it to ask you questions first — still cheaper than guessing your way there.

**EXAMPLE**

*"Can you help me with training?"* — then three more messages nailing down topic, audience, and length. Versus: *"Create a 5-question quiz on workplace safety basics for new warehouse employees, multiple choice, showing the score at the end."* One message instead of four. Not a hard choice.

# Ways to Reduce Your Token Costs

Practical habits that add up over time

- **Be specific from message one.** Covered this already — it's the biggest lever you've got.
- **New topic, new chat.** A conversation that keeps rolling gets more expensive with every single message, whether you notice it or not.
- **Upload only what you need.** The relevant pages, not the whole binder.
- **Ask for the length you want.** "Under 200 words" costs less than an open-ended request that turns into three pages you didn't ask for.
- **Reuse what you've already got.** Edit it. Don't re-explain the whole task from scratch every time.
- **Match the model to the job.** Save the expensive one for the hard stuff. Use the cheap, fast one for everything else.
- **Batch it.** Send related requests together instead of five separate messages spread across the day.
- **Ask about volume discounts or flat-rate plans** if you're a heavy, predictable user. Worst they say is no.

# Why Would a Vendor Not Charge Token Fees?

Flat-rate pricing is a choice, not the absence of cost

Some vendors — plenty of consumer subscriptions, for one — charge one flat fee a month instead of billing you per token. That doesn't mean the tokens are free. They're not. The vendor's still paying for every bit of compute behind the scenes. They've just decided to absorb it and average it out instead of itemizing your bill down to the token. Here's why:

- **Simplicity.** A flat fee is easy. Everybody understands "$20 a month." Almost nobody understands a variable, usage-based bill until they're staring at a surprise charge.
- **Killing the meter-running feeling.** Nothing kills exploration faster than watching a number tick up. Flat pricing removes that anxiety entirely.
- **Averaging across the whole user base.** The vendor's betting heavy users and light users cancel each other out. Usually a safe bet.
- **Bundling.** Fold the token cost into something bigger — a subscription, an enterprise deal, a platform fee — so nobody has to think about it line by line.

**GOOD TO KNOW**

Flat-rate doesn't mean unlimited, even if nobody says that part out loud. There's almost always a fair-use ceiling somewhere behind the scenes. Push past it hard enough, and don't be surprised if the vendor asks you to move to a metered plan.

# Cheat Sheet

Tips, hints & tricks to save on token fees

## BEFORE YOU START

- Write one specific, detailed first message — skip the vague warm-up.
- State the goal, audience, and format up front.
- Upload only the pages or sections you actually need.
- Ask for a plan before a big task, so you're not paying for a wrong-direction first draft.

## WHILE YOU WORK

- Keep conversations focused — start fresh for a new topic.
- Ask for concise answers when that's all you need.
- Batch related requests instead of many small ones.
- Reuse and edit prior outputs instead of restarting from scratch.

## WITH FILES & MEDIA

- Trim documents to the relevant sections before uploading.
- Compress or resize images when full detail isn't needed.
- Avoid re-uploading the same file into every message.
- Use video/audio only when text truly won't do the job.

## BIGGER-PICTURE MOVES

- Match the model to the task — save top-tier models for complex work.
- Ask about volume discounts if usage is high and steady.
- Ask if a flat-rate plan beats pay-per-token for your usage pattern.
- Revisit your usage monthly — habits drift as needs change.

### THE ONE HABIT THAT MATTERS MOST

Nearly every tip on this page comes back to the same one idea. Think for thirty seconds before you hit send. That's it. That's the whole system. A clear, complete first message is the biggest lever you have over your AI costs — and it doesn't cost you a thing to use it.