

A BEGINNER'S MANUAL

GROK EDITION

Get the Most from Every AI Plan

Understand usage limits, stretch free and paid plans,
batch your work, and rotate between ChatGPT, Claude,
Gemini, Grok and Perplexity

Get the Most from Every AI Plan

Understand usage limits, stretch free and paid plans, batch your work, and rotate between ChatGPT, Claude, Gemini, Grok and Perplexity

GROK EDITION

Written by Grok Bot, with facts checked against official sources in October 2026.

Every AI chatbot has limits. Sometimes they are generous; sometimes they arrive in the middle of your most important task. Most people hit them sooner than they need to, because they spend their allowance on long, rambling chats, repeated uploads and messages that could have been one.

This manual explains, in plain language, **how usage limits actually work** on ChatGPT, Claude, Gemini, Grok and Perplexity, using each company's own help pages. Then it teaches the habits that make every message count: planning before you prompt, batching, keeping chats short, preparing files, choosing the right model and effort level, and rotating between providers so one limit never stops your work.

Limits and prices change often. Every figure here is dated, and every chapter tells you where to check the live number.

HOW TO READ THIS MANUAL

Tip — a shortcut or good habit.

Warning — easy to get wrong.

`$ command` — text you type, exactly as shown.

Note — useful background.

Danger — can lose data, money, or access.

`Settings > Privacy` — a menu path to click.

New words appear in **bold** the first time and are collected in the Glossary at the back.

Plan names, prices, models and limits were checked on official OpenAI, Anthropic, Google, xAI and Perplexity pages on October 16 and 17, 2026, and are U.S. prices before tax unless noted. These companies change plans frequently and sometimes show different details on different pages; this manual points out such differences where they were found. Always confirm the current details in the product or on the official pricing page before buying. Follow each service's terms of use; this manual does not recommend creating extra accounts or any other way of evading limits.

Contents

01	Why Limits Exist and How They Work	5
1.1	The common pattern	6
1.2	What makes a request expensive	6
1.3	Tokens: the unit behind it all	7
1.4	Context windows by plan	7
1.5	Resets: when you get usage back	8
02	ChatGPT: Plans and Limits	9
2.1	The plans at a glance	9
2.2	Thinking levels	10
2.3	Separate tool limits	10
2.4	Getting the most from ChatGPT	11
03	Claude: Plans and Limits	12
3.1	The plans at a glance	12
3.2	What uses up Claude's allowance	13
3.3	Claude's built-in savings: caching	13
3.4	Usage credits: keep working after a limit	13
3.5	Anthropic's own tips, summarized	14
04	Gemini: Plans and Limits	15
4.1	The plans at a glance	15
4.2	What Google says costs the most	16
4.3	When you hit a limit	16
05	Grok: Plans and Limits	17
5.1	The plans at a glance	17
5.2	How the weekly pool works	18
06	Perplexity: Plans and Limits	19
07	Side by Side: Choosing Where to Spend	20
7.1	Entry-level paid plans compared	20
7.2	Match the plan to your main job	21
7.3	When is a subscription worth it?	21
08	Plan Before You Prompt	22
8.1	The two-minute plan	22
8.2	The complete-brief template	23
09	Batching: More Done per Message	24

9.1	A batch request	24
9.2	Batching rules of thumb	25
10	Keep Chats Short and Clean	26
10.1	The handoff summary	26
10.2	Context hygiene habits	27
11	Prepare Files So They Cost Less	28
11.1	Trim before you upload	28
11.2	Projects: upload once, ask many times	29
12	Pick the Right Model and Effort	30
12.1	A simple ladder	30
12.2	Make expensive tools count	31
13	Control Output Length	32
13.1	Phrases that save usage	32
14	Rotate Between Providers	33
14.1	The portable handoff	33
14.2	A rotation plan	33
14.3	Keep a prompt library	34
15	When to Use an API Instead	35
16	Making Free Plans Go Further	36
16.1	A free-only toolkit	36
16.2	Free-plan habits	36
17	A Week of Smart Usage: Worked Example	38
18	Track Your Usage	39
18.1	Where to look	39
18.2	A simple usage log	39
19	Troubleshooting	40
	Quick Reference	41
	Glossary	42

Why Limits Exist and How They Work

- AI answers cost real computing power. Limits share that power among millions of users.
- Most services now measure **compute**, not just message counts. A long chat with big files costs far more than a short question.
- Limits usually **reset on a timer**: often a rolling few-hour window plus a weekly cap.

When you send a message to an AI chatbot, a data center does a surprising amount of work. The model reads your message, plus (usually) the **whole conversation so far**, plus any files, plus hidden instructions. It may search the web, run code or “think” for a while before answering. All of that costs money in chips and electricity.

So every provider limits use. Free plans get a little; paid plans get more; top plans get a lot. The details differ by company, but the shape is now remarkably similar.

1.1 The common pattern

IDEA	WHAT IT MEANS FOR YOU
Compute-based usage	Bigger, harder requests use more of your allowance than small ones
Short window	Many services reset part of your allowance every few hours (Claude and Gemini say every five hours)
Weekly cap	Paid plans often add a weekly limit on top, so you can't use a whole week's worth in one day
Per-feature limits	Image generation, video, deep research, voice and uploads often have their own, tighter limits
Fallback models	When you run out of the best model, some services switch you to a smaller one instead of stopping
Pay-as-you-go top-ups	Several services now let paid users buy extra usage after the allowance runs out

NOTE

Companies rarely publish a fixed number of messages any more. Anthropic says plainly that “there’s no fixed message count,” and Google calls Gemini’s limits “compute-based.” Learning **what makes a request expensive** is more useful than memorizing a number.

1.2 What makes a request expensive

Across the help pages of the major providers, the same cost factors come up again and again:

- 1. Conversation length.** Most chatbots re-read the earlier conversation every time you send a message. Message 40 in a chat costs much more than message 1.
- 2. Files and attachments.** A 100-page PDF must be read, and often re-read.
- 3. Model choice.** The biggest models cost more per message than small, fast ones.
- 4. Thinking or effort level.** “Thinking,” “reasoning,” “extended” and “high effort” modes do more work before they answer.
- 5. Tools.** Web search, deep research, code running, file creation and browsing all add work.
- 6. Media.** Images cost more than text; video costs the most. Google’s Gemini help page rates video generation as “high” usage.
- 7. Output length.** Asking for a 5,000-word answer costs more than asking for 500 words.

Google’s Gemini help page even publishes a simple table of relative costs: a simple chat message is “very low,” ongoing conversations and large file analysis are “medium,” and Deep Research and video generation are “high.”

1.3 Tokens: the unit behind it all

AI models read and write in **tokens**: chunks of text that are usually part of a word. OpenAI's commonly quoted rule of thumb is that one token is about four characters of English, or about three-quarters of a word. So 1,000 tokens is roughly 750 words.

You don't need to count tokens. Just remember:

- Everything in the conversation, including files and earlier messages, is made of tokens.
- Every model has a **context window**: the maximum number of tokens it can consider at once.
- Long chats and big files fill the context window and use up your allowance faster.

1.4 Context windows by plan

Context windows matter because when a conversation or file exceeds them, the model starts to forget or skip things. Two providers publish plan-by-plan figures:

SERVICE AND PLAN	CONTEXT WINDOW (AS PUBLISHED)
ChatGPT Free, Instant	27K tokens total
ChatGPT Go and Plus, Instant	54K tokens total
ChatGPT Pro, Instant	128K tokens total
ChatGPT Go and Plus, reasoning	256K tokens total
ChatGPT Pro, reasoning	400K tokens total
Gemini without a plan	32K tokens
Gemini with Google AI Plus	128K tokens
Gemini with Google AI Pro or Ultra	1 million tokens

OpenAI's pricing page explains that the part available for your own input is smaller than the total, because instructions, memories and the model's reasoning also take space. It estimates the input maximum at about 12 pages of text on Free Instant, about 40 pages on Go and Plus Instant, and about 320 pages on Go and Plus reasoning. Anthropic's pricing page lists Claude's context window as "up to 1M, varies by model" on every plan. Google says a 1-million-token window can hold about 1,500 pages of text or 30,000 lines of code.

TIP

If you're on a free plan with a small context window, the advice in this manual matters twice as much. A free ChatGPT chat with a long document can run out of room quickly; split the document and ask focused questions instead.

1.5 Resets: when you get usage back

SERVICE	HOW RESETS WORK (OFFICIAL DESCRIPTION)
ChatGPT	Shows the reset time when you hit a limit; Thinking limits vary by plan
Claude	Session limit resets every five hours; paid plans also have a weekly limit with a fixed reset time for your account
Gemini	Limit refreshes every five hours until you reach your weekly limit
Grok (paid)	One weekly usage pool; the reset date and time appear in Settings, Usage
Perplexity	No fixed published numbers; Pro has “extended” Pro Search access

☆ IMPORTANT

Support teams can’t reset your limits. OpenAI’s help center says so directly. When you hit a limit, your options are to wait, switch to another model or service, buy more usage where offered, or upgrade.

ChatGPT: Plans and Limits

- Free and Go users get unlimited everyday text chats, subject to abuse safeguards, but tighter limits on uploads, images, voice and data analysis.
- Plus costs \$20 a month and adds advanced reasoning, more uploads and expanded deep research.
- Pro comes in tiers (OpenAI's pages mention \$100 and \$200 tiers) with the largest context windows and Pro-level models.

All details in this chapter come from chatgpt.com/pricing and OpenAI's help center, checked October 17, 2026.

2.1 The plans at a glance

PLAN	PRICE (U.S.)	HIGHLIGHTS
Free	\$0	Unlimited everyday text chats; limited uploads, images, voice, deep research, data analysis; Projects; ads may appear in some countries
Go	Check the pricing page	More messages with tools, more uploads and images, longer memory; may include ads
Plus	\$20 a month, monthly only	Advanced reasoning models, expanded uploads and deep research, scheduled tasks, custom GPTs, Codex
Pro	Several usage tiers; OpenAI's help pages mention \$100 and \$200 tiers	Pro reasoning models, maximum deep research and context, unlimited image creation (subject to policies)
Business and Enterprise	Per user	Admin controls, compliance features, data excluded from training by default

i NOTE

OpenAI's pages were mid-rollout when this was written. The pricing page described Free as "unlimited text chats with GPT-6," while the help center said Free and Go use GPT-5.6 Luna by default. Your model picker shows what your account actually has.

2.2 Thinking levels

In ChatGPT's newest model picker you choose a **thinking level**: Instant, Medium, High, Extra High or Pro. Higher levels think longer and use more of your allowance.

PLAN	MEDIUM AND HIGH	EXTRA HIGH	PRO
Free and Go	No ("Think" uses the smaller default model)	No	No
Plus	Yes	No	No
Pro	Yes	Yes	Yes
Business and Enterprise	Yes	Yes	Yes

When you reach a Thinking limit, OpenAI says ChatGPT "may continue with another available model," and it shows the reset time when it can.

2.3 Separate tool limits

Even when text chat is unlimited, OpenAI lists separate limits for:

- File and image uploads
- Image creation
- Voice
- Data analysis
- Deep research

The Free plan FAQ adds a useful fact: if you hit a rate limit on Free and then upgrade to Plus, Pro or Business, your usage resets.

2.4 Getting the most from ChatGPT

- ☐ Use **Instant** for quick tasks; save higher thinking levels for hard problems.
- ☐ Put recurring files in a **Project** instead of uploading them to every chat.
- ☐ Use **custom instructions** so you don't repeat your preferences in every message.
- ☐ Start a **new chat** when the topic changes.
- ☐ On Free, keep documents short; the Instant context window is about 12 pages of input.
- ☐ Check what's in **Memory** (Settings, Personalization) so stale facts don't waste space.

When upgrading helps most

Upgrading from Free to Plus makes the biggest difference if you:

- Regularly upload files or analyze data, which have tighter limits on Free
- Need longer documents in one chat (54K tokens on Plus Instant versus 27K on Free, and 256K with reasoning)
- Want advanced thinking levels for hard problems
- Use deep research, scheduled tasks or Codex often

If you mostly ask short text questions, Free's unlimited everyday text chats may be all you need.

TIP

Asking ChatGPT to “think harder” doesn't change the thinking level on Plus or Pro, according to OpenAI's help center. Use the picker instead.

Claude: Plans and Limits

- Every Claude plan has a usage limit that resets on a rolling five-hour session window. Paid plans add weekly limits.
- Pro is \$20 a month (or \$17 a month billed annually) and gives at least five times Free’s usage per session.
- Claude chat, desktop, mobile and Claude Code all draw from the **same pool**.

Details come from claude.com/pricing and the Claude Help Center, checked October 17, 2026.

3.1 The plans at a glance

PLAN	PRICE (U.S.)	USAGE	NOTABLE LIMITS
Free	\$0	Baseline	Sonnet and Haiku models; up to 5 projects; no Research or Claude Code
Pro	\$20 monthly, or \$17 a month billed annually (\$200 up front)	At least 5x Free per five-hour session	Adds Opus, Research, Claude Code, and Projects without the Free cap
Max 5x	From \$100 a month	5x Pro per session	Monthly billing; includes \$100 a month in API credits
Max 20x	Higher than Max 5x; see claude.com/pricing	20x Pro per session	Monthly billing; includes \$200 a month in API credits
Team and Enterprise	Per seat	Varies by seat type	Admin controls; Enterprise is billed annually

Anthropic adds that it “may limit your usage in other ways, such as weekly and monthly caps or model and feature usage.”

3.2 What uses up Claude's allowance

Anthropic's "Usage limit best practices" article lists the factors: message length, attachment size, conversation length, tools such as Research and web search, model choice, effort level, artifacts, and multi-step tasks like running code or browsing.

3.3 Claude's built-in savings: caching

Claude has one feature that rewards good habits directly. According to Anthropic:

- Content in **projects is cached** and counts less against your limits when reused.
- Caches expire after a period of inactivity, so the first message after a long break counts the content in full again.
- Projects also offer a retrieval (RAG) mode for larger knowledge bases.

TIP

If you'll ask several questions about the same documents, put them in a Claude project **once** and ask your questions in a focused session, instead of uploading the files to separate chats on separate days.

Watching your usage

On Pro, Max, Team and seat-based Enterprise plans, open **Settings, Usage**. You'll see progress bars for the current five-hour session (with time remaining) and the weekly limits, including when they reset.

3.4 Usage credits: keep working after a limit

Pro and Max subscribers can turn on **usage credits** in Settings, Usage. When you reach a limit, your requests switch to pay-as-you-go pricing at standard API rates instead of stopping. You prepay, can set a monthly spend limit, and can turn on auto-reload. These charges are separate from your subscription. If you subscribed in a mobile app, Anthropic says you can enable usage credits only on the web.

DANGER

Pay-as-you-go use can add up quickly with big models and long chats. Set a monthly spend limit before you turn credits on.

3.5 Anthropic's own tips, summarized

- ☐ Plan the conversation before you start.
- ☐ Be specific; avoid vague questions that need clarification.
- ☐ Combine related questions in one message.
- ☐ Review and edit your prompt before sending.
- ☐ Refer back to earlier parts of the chat instead of repeating them.
- ☐ Paid plans: let Claude search past chats or use memory instead of re-explaining.
- ☐ Send a whole document for editing at once, rather than in pieces.

Gemini: Plans and Limits

- Gemini’s limits are compute-based and refresh every five hours until you reach a weekly limit.
- Google AI Plus doubles the standard limits, AI Pro quadruples them, and AI Ultra gives 5x or 20x AI Pro.
- Context windows grow with the plan: 32K tokens without a plan, up to 1 million on Pro and Ultra.

Details come from Google’s “Gemini Apps limits & feature availability” help page and gemini.google/subscriptions (U.S.), checked October 17, 2026.

4.1 The plans at a glance

PLAN	PRICE (U.S.)	USAGE VS. NO PLAN	CONTEXT WINDOW
No plan (Free)	\$0	Standard; uses the Flash-Lite model	32K tokens
Google AI Plus	\$4.99 a month	2x	128K tokens
Google AI Pro	\$19.99 a month	4x	1 million tokens
Google AI Ultra	\$99.99 or \$199.99 a month	5x or 20x AI Pro	1 million tokens

Google AI plans are part of Google One and include other benefits such as cloud storage. Availability differs by country.

4.2 What Google says costs the most

REQUEST TYPE	RELATIVE USAGE COST
Simple chat message	Very low
Ongoing conversation	Medium
Large file analysis	Medium
Deep Research	High
Video generation	High

Google’s own tip on that page: **“Start a new chat when you move to a new topic,”** because Gemini won’t have to re-read a long history.

4.3 When you hit a limit

- With a Google AI subscription, you can keep chatting using the lighter Flash-Lite model.
- Otherwise, wait for the refresh or upgrade.
- Gemini notifies you as you approach the limit, then tells you when it will refresh.
- Check anytime at gemini.google.com: **Settings, Usage Limits**.

WARNING

Google warns that, for users without a plan, compute-heavy features like Deep Research “may be unavailable during periods of high demand,” and that free limits may be reduced before paid ones.

Grok: Plans and Limits

- Paid Grok plans use **one weekly usage pool** shared across Chat, Voice, Imagine, Build and other products.
- At the time of writing, x.ai/pricing listed Free, Plus (\$8), Premium (\$30), Super (\$100) and Ultra (\$200) a month.
- When the weekly pool runs out, paid features pause but the free-tier Chat and Voice limits still apply.

Details come from x.ai/pricing and xAI's documentation (the Grok FAQ and "Usage limits"), checked October 17, 2026.

5.1 The plans at a glance

PLAN	PRICE (U.S., BEFORE TAX)	USAGE
Free	\$0	Limited Chat and agent requests
Plus	\$8 a month	Shared AI usage pool
Premium	\$30 a month	4x Plus
Super	\$100 a month	15x Plus, priority access at peak times
Ultra	\$200 a month	50x Plus

NOTE

xAI's pricing page and its FAQ used different plan names when checked: the FAQ still described "SuperGrok" subscriptions, while x.ai/pricing listed "The X Subscription" tiers above. Check x.ai/pricing for the current lineup.

5.2 How the weekly pool works

According to xAI's FAQ, which describes a change rolling out in June 2026:

- One pool works across all Grok products. You can spend it all on one product or mix them.
- Products cost different amounts depending on compute. A chat message uses little; a high-quality video or a long coding task uses far more.
- **Settings, Usage** shows a progress bar, a breakdown by product (such as API, Build, Chat, Imagine and Voice), your weekly reset date and time, and any extra credit balance.
- When the pool is used up, paid features pause until the reset. You can buy Extra Usage Credits (on the web, from \$5; they expire a year after purchase), turn on Auto Top Up, or upgrade.

xAI notes that extra credits are priced at standard rates, which cost more per action than your included usage. If you regularly run out, it says upgrading is usually better value.

Grok file tips from xAI

- The web app accepts up to about 100 files per message (Android up to 20).
- Most files can be up to 150 MB.
- For PDFs over 100 pages, refer to specific pages by their page numbers.
- Break extremely large documents into smaller chunks if needed.

Perplexity: Plans and Limits

- Perplexity Free gives “practically unlimited” basic searches but only a very limited number of Pro Searches.
- Pro offers “extended” Pro Search, advanced models and more uploads; Max offers the highest access.
- Perplexity doesn’t publish fixed daily numbers, so watch the in-app indicators.

Details come from Perplexity’s Help Center (“Which Perplexity Subscription Plan is right for you?” and “What is Perplexity Pro?”), checked October 17, 2026.

PLAN	WHAT PERPLEXITY SAYS YOU GET
Standard (Free)	Practically unlimited basic searches; very limited Pro Searches; limited basic uploads; automatic model choice
Pro	Extended Pro Search; advanced models; image and video generation; more uploads; a limited number of “Create files and apps” queries every 30 days
Education Pro	Pro features for verified students and educators, \$10 a month
Max	Highest access to advanced models; extended limits for Create files and apps; early access to new products
Enterprise Pro and Max	Per seat (Enterprise Pro starts at \$40 a month or \$400 a year); data not used for training

Two useful details from the Pro article:

- **“Best” mode is available without quota limits.** It picks a model for you.
- Access to advanced models “may be limited during weeks of especially heavy usage.”

TIP

On Perplexity, use Best mode for everyday lookups and save Pro Search, Research and Create files and apps for questions that truly need them.

Side by Side: Choosing Where to Spend

- Each service has different strengths and different ways of counting usage.
- A sensible approach: one paid plan you use daily, plus free plans of the others for overflow and second opinions.
- Pick by what you actually do most, not by benchmark headlines.

7.1 Entry-level paid plans compared

SERVICE	ENTRY PAID PLAN (U.S.)	HOW USAGE IS COUNTED	WHERE TO CHECK USAGE
ChatGPT	Plus, \$20 a month	Per-model and per-tool limits; reset times shown in product	Shown when you near or hit a limit
Claude	Pro, \$20 a month or \$17 a month annually	Five-hour session plus weekly limit, one pool across apps and Claude Code	Settings, Usage
Gemini	Google AI Plus, \$4.99 a month; AI Pro, \$19.99	Compute-based; five-hour refresh plus weekly limit	Settings, Usage Limits
Grok	Plus, \$8 a month; Premium, \$30	One weekly compute pool across products	Settings, Usage
Perplexity	Pro (check perplexity.ai for price)	Extended Pro Search; no fixed published counts	In-app indicators

7.2 Match the plan to your main job

IF YOU MOSTLY...	CONSIDER STARTING WITH...	WHY
Write and edit long documents	Claude or ChatGPT	Strong writing; projects for reference files
Research with sources	Perplexity, or Deep Research features in others	Citations are the core of the product
Work inside Gmail, Docs and Drive	Gemini	Built into Google apps on many plans
Follow news and posts on X	Grok	Connected to X
Code	Claude (Claude Code), ChatGPT (Codex) or Grok (Build)	Coding tools share the plan's allowance
Analyze very large files	Gemini AI Pro, ChatGPT Pro, or Claude	Largest published context windows
Make lots of images or video	Check media limits specifically	Media has the tightest limits everywhere

TIP

Before paying, spend one ordinary week using the free plans of two or three services for your real tasks. Note which one you reach for first. That's usually the one to pay for.

7.3 When is a subscription worth it?

Ask three questions:

- 1. Do I hit limits at least weekly on tasks that matter?** If not, free plans plus good habits may be enough.
- 2. Do I need a specific paid feature?** Examples: Claude Code, deep research volume, a 1-million-token context, scheduled tasks.
- 3. Would the time saved be worth more than the price?** If an hour of your time is worth more than a month's subscription, the answer is often yes.

Plan Before You Prompt

- The cheapest message is the one you didn't need to send.
- Spend two minutes planning: goal, audience, materials, format and limits.
- A clear first message often replaces five rounds of back-and-forth.

Most wasted usage comes from vague first prompts. The AI guesses, you correct, it guesses again. Each round re-reads the growing conversation.

8.1 The two-minute plan

Before you open the chat, jot down:

QUESTION	EXAMPLE
What do I want at the end?	A one-page cover letter
Who is it for?	A hiring manager at a small nonprofit
What materials does the AI need?	Job ad, my résumé, two achievements to highlight
What format?	Under 350 words, three paragraphs, plain text
What should it avoid?	Clichés like “passionate”; anything I haven’t actually done
How will I judge it?	Specific, warm, mentions the nonprofit’s mission

8.2 The complete-brief template

PROMPT

GOAL: [what you want at the end]
AUDIENCE: [who will read or use it]
CONTEXT: [background the AI needs, in a few lines]
MATERIALS: [pasted text or attached files, labeled]
FORMAT: [length, structure, tone, file type]
AVOID: [things you don't want]
QUALITY CHECK: Before answering, check your draft against my goal and format, and fix any problems. If something essential is missing, ask me up to three questions first instead of guessing.

NOTE

The last line matters. Inviting **up to three questions** prevents a wasted long answer built on a wrong guess, while capping the questions keeps the exchange short.

Edit before you send

Anthropic's own advice includes this simple step: "Take a moment to review your message for clarity and completeness." Read your prompt once. Did you include the file? The word limit? The audience?

Batching: More Done per Message

- Group related questions and tasks into one message.
- Number the tasks and ask for numbered answers.
- Don't batch unrelated topics; that bloats the context for everything after.

Both Anthropic and Google recommend combining related requests. Instead of six messages about six math problems, send all six at once.

9.1 A batch request

PROMPT

I have five related tasks about the attached meeting notes.
Answer each under its own numbered heading. Keep each answer under 100 words unless I say otherwise.

1. List every decision made, with who made it.
2. List every action item: owner, task, due date (write "none" if missing).
3. Draft a three-sentence summary for people who missed the meeting.
4. Identify any open questions nobody answered.
5. Suggest an agenda for the follow-up meeting (max 5 items).

9.2 Batching rules of thumb

- ☐ Batch tasks that use the **same materials** (one file, one topic).
- ☐ Number them so answers are easy to check and refer to later (“Redo #3 shorter”).
- ☐ Set length limits per item so one task doesn’t crowd out the others.
- ☐ Keep a batch to roughly five to ten items; very long batches can get sloppy.
- ☐ If one answer is wrong, fix only that item: “Only redo #4.”

Batch your edits too

When revising, collect all your changes into one message instead of drip-feeding them:

PROMPT

Revise the draft with these changes, all at once:

1. Cut the second paragraph.
2. Make the tone warmer in the opening.
3. Replace "utilize" with "use" everywhere.
4. Add a closing line inviting questions.

Return the full revised text only, with no commentary.

TIP

“Return the full revised text only, with no commentary” saves output tokens and saves you from scrolling past explanations.

Keep Chats Short and Clean

- Long chats get expensive and less reliable, because the model re-reads everything.
- Start a fresh chat for each new topic.
- When a chat gets long, ask for a **handoff summary** and continue in a new chat.

Google's help page explains that ongoing conversations "require Gemini to re-read the previous conversation history." The same is broadly true of other chatbots. A long chat also fills the context window, after which the model may lose track of early details.

Signs a chat has gone on too long

- Answers ignore instructions you gave near the start.
- The AI contradicts something it said earlier.
- Responses slow down.
- You're warned that you're near a limit sooner than usual.

10.1 The handoff summary

PROMPT

We're going to continue this in a new chat. Write a handoff summary I can paste there. Include: the goal, key facts and decisions so far, my preferences and constraints, the current draft or status, and the next three steps. Keep it under 300 words.

Paste the summary into a new chat, attach only the files still needed, and carry on. You keep the important context while dropping hundreds of messages of history.

10.2 Context hygiene habits

- ☐ One topic per chat.
- ☐ Name chats clearly so you can find them later instead of starting over.
- ☐ Delete or archive dead-end chats so you don't accidentally continue them.
- ☐ Don't paste the same long text twice in one chat; refer to it ("the job ad above").
- ☐ Avoid "regenerate" loops; instead, say exactly what to change.
- ☐ Check saved **memory** occasionally and remove facts that are no longer true.

☆ IMPORTANT

Regenerating an answer repeatedly is one of the fastest ways to burn usage. Each regeneration is a full new answer. One specific correction ("make it 30% shorter and less formal") usually works better.

Prepare Files So They Cost Less

- Send only the part of a file the AI needs.
- Convert messy files (scans, slides, spreadsheets with many tabs) into clean text or a focused extract.
- Upload once, to a project, when you'll ask several questions about the same material.

Files are often the biggest hidden cost in a conversation. A 200-page PDF attached “just in case” can use more of your allowance than an hour of chatting.

11.1 Trim before you upload

INSTEAD OF...	TRY...
The whole 200-page report	The 12 pages you're asking about
A spreadsheet with 15 tabs	One tab exported as CSV
A slide deck full of images	The slide text pasted as plain text
A long email thread	The last few messages plus a two-line summary of the rest
A photo of a document	Copy the text, or use your phone's text-recognition feature first
A whole code repository	The two or three files involved in the bug

Point the AI at the right place

xAI's own advice for large PDFs applies everywhere: refer to **specific page numbers**. Better still, say what to look for:

PROMPT

In the attached contract, look only at the sections on termination and renewal (around pages 14 to 18). Summarize my obligations and deadlines in a table. Ignore the rest of the document.

11.2 Projects: upload once, ask many times

SERVICE	FEATURE	WHY IT SAVES USAGE
Claude	Projects	Anthropic says project content is cached and counts less when reused
ChatGPT	Projects	Files and instructions live with the project instead of being re-sent manually
Gemini	Notebooks and Gems	Keep sources and instructions together
Grok	Projects	Group files and chats (use grok.com; xAI says some other hosts lack Projects)
Perplexity	Projects (formerly Spaces)	Shared instructions and files for a topic

TIP

Group your questions about a project's files into one sitting. Claude's cache expires after inactivity, so ten questions spread over ten days cost more than ten questions in one hour.

Pick the Right Model and Effort

- Use the smallest, fastest option that does the job well.
- Save “thinking,” “Pro,” “Research” and video tools for tasks that truly need them.
- Many tasks (rewording, formatting, quick facts) don’t need the biggest model.

12.1 A simple ladder

TASK	USUALLY ENOUGH	STEP UP WHEN...
Fix grammar, reformat, translate a sentence	Fast default model (Instant, Haiku, Flash)	Never, usually
Summarize an article, draft an email	Default model	The writing is high-stakes
Plan a project, compare options	Default model or medium thinking	The trade-offs are complex
Hard math, tricky code bugs, legal-style analysis	Thinking or reasoning modes	The first answer is shaky
Multi-source report with citations	Deep Research or Perplexity Research	Always worth a careful prompt first
Images and video	Media tools	Plan the prompt carefully; media costs the most

NOTE

On Claude, the help center lists model choice and effort level among the factors that affect limits. On ChatGPT, Instant is unlimited for everyday text on Free and Go, while Thinking levels have their own limits.

12.2 Make expensive tools count

Deep Research, Pro reasoning and video generation have the tightest limits. Before using one:

- 1 Draft the prompt in a cheap chat first, and ask the AI to improve it.
- 2 State the scope: time period, region, sources to prefer, length.
- 3 Say what to exclude.
- 4 Specify the output format.
- 5 Run the expensive tool once, with the polished prompt.

PROMPT

I'm about to run a deep research task. Here is my draft request:
[draft]. Improve it so the research is focused and efficient: add a clear scope, sources to prefer, questions to answer, and an output format. Return only the improved prompt.

Control Output Length

- Longer answers use more allowance and take longer to read.
- Ask for the length and format you need, up front.
- Ask for “only the changes” when editing long documents.

13.1 Phrases that save usage

PHRASE	EFFECT
“Answer in under 100 words.”	Short answer
“Bullet points only, no introduction.”	Removes filler
“Return only the revised text.”	No commentary
“Show only the lines you changed.”	Avoids reprinting a whole document or file
“Give me 3 options, one sentence each.”	Small, comparable choices
“If you’re unsure, say so in one line instead of guessing.”	Prevents long, padded answers

Set it once with custom instructions

Most services let you save standing preferences, so you don’t repeat them:

PROMPT

Default style for all answers: concise and plain. Lead with the answer, then up to five bullet points of detail. No preambles, no repeating my question, no closing summaries unless I ask. Use tables for comparisons. Ask before writing anything over 500 words.

Put it in ChatGPT’s custom instructions, Claude’s profile preferences, Gemini’s personalization settings or a Gem, or a Grok or Perplexity project’s instructions.

Rotate Between Providers

- When one service hits a limit, continue in another with a handoff summary.
- Use different services for what each does best, and for second opinions.
- Keep a private “prompt library” so any chat can pick up your work.

Having free accounts with two or three services is the simplest way to keep working through a limit. It also gives you second opinions on important answers.

14.1 The portable handoff

The handoff summary from the “Keep Chats Short” chapter works across companies too. Ask the first AI for a summary, open another service, paste it with any needed files, and continue.

PROMPT

Here is a handoff from another AI assistant. Read it, confirm in two sentences what we're doing and what's next, then continue with step 1.

[paste handoff summary]

14.2 A rotation plan

SITUATION	MOVE TO...
Hit a limit mid-draft	Another chatbot, with the handoff summary
Need sources for a claim	Perplexity, or a search-enabled chat
Want a second opinion	A different company's model (they make different mistakes)
Working in Google Docs or Gmail	Gemini
Hit a weekly limit	Another service, or wait for the reset date you saw in Settings

14.3 Keep a prompt library

Store your best prompts, custom instructions and handoff summaries in a notes app or document. Then any service, on any plan, can start strong. Include:

- Your custom-instruction text
- Templates you use often (the complete brief, batch requests, handoffs)
- Background about recurring projects (a short “about this project” paragraph)

WARNING

Rotating between your own accounts on **different services** is normal. Creating multiple accounts on the **same service** to get around its limits may break its terms and can lead to suspension. Read each service’s terms.

Privacy when rotating

Each service has its own privacy settings. Before pasting sensitive material into a new service, check whether it uses chats for training and how to turn that off. Volume 18, *AI Privacy & Security*, covers this in detail.

When to Use an API Instead

- APIs bill per token instead of a monthly subscription.
- They suit automation and occasional heavy use, but need setup and spending caps.
- At OpenAI and Anthropic, subscriptions and APIs are billed separately.

OpenAI and Anthropic both say API usage is separate from their consumer subscriptions (OpenAI says API usage “is separate and billed independently” from Plus; Anthropic says Pro “does not include API usage”). xAI is an exception worth knowing: its current plans list Grok API access as part of the shared usage pool. Pay-per-token APIs can be cheaper for light, occasional use and essential for automation, but you need an API key, a little setup, and spending limits. Volume 15, *AI APIs & Costs*, explains how to estimate and cap API spending.

CHOOSE A SUBSCRIPTION WHEN...	CHOOSE AN API WHEN...
You chat interactively every day	You want to automate a repeated task
You want apps, voice, images and research tools	You're building something for others
You prefer a predictable monthly price	Your usage is light or bursty
You don't want to manage keys	You're comfortable with setup and caps

Making Free Plans Go Further

- Free plans are surprisingly capable if you use them deliberately.
- Combine several free plans, each for what it does best.
- Spend scarce features (uploads, images, research) only on what matters.

16.1 A free-only toolkit

NEED	FREE OPTION (AS DESCRIBED BY EACH COMPANY)
Everyday questions and drafting	ChatGPT Free (unlimited everyday text chats, subject to safeguards)
Writing with a few reference files	Claude Free (up to 5 projects)
Searching with citations	Perplexity Free ("practically unlimited" basic searches)
Google apps and quick help	Gemini without a plan
News and posts on X	Grok Free (limited Chat requests)

16.2 Free-plan habits

- ☐ Use text chat for most work; save uploads and images for when you really need them.
- ☐ Paste the relevant excerpt instead of uploading a whole file.
- ☐ Keep each free chat short; free context windows are small (27K tokens on ChatGPT Free Instant, 32K on Gemini without a plan).
- ☐ Do heavy tasks early in your reset window, so a limit doesn't land mid-task.
- ☐ When one service says "limit reached," rotate with a handoff summary.
- ☐ On Gemini without a plan, expect Deep Research to be unavailable at busy times.

 TIP

Note the reset time when a limit message appears. Put it in your calendar if the task matters; you'll know exactly when to come back.

Spend scarce features in the right order

On a free plan, treat uploads, images and research like a small monthly budget:

1. **Text first.** Try a plain text question; many tasks never need more.
2. **Excerpt second.** If the AI needs a document, paste the relevant part.
3. **Upload third.** Upload only when layout, tables or images matter.
4. **Research last.** Use deep research for questions that need many sources, once, with a polished prompt.
5. **Images when final.** Generate images after the wording and idea are settled, not while you're still brainstorming.

A Week of Smart Usage: Worked Example

Meet Priya, a freelance marketer with one paid plan (Claude Pro) and free accounts with ChatGPT, Gemini and Perplexity. Here's how she spreads a busy week.

DAY	TASK	WHERE	USAGE-SAVING MOVE
Monday	Set up a client project	Claude project	Uploads the brand guide and past campaigns once
Monday	Draft five social posts	Same Claude project	One batch request with numbered posts and word limits
Tuesday	Research competitors	Perplexity Free	Basic searches for facts; one Pro Search for the hardest question
Wednesday	Write a long blog post	Claude project	Complete brief first; edits batched into one message
Wednesday	Claude session limit reached	ChatGPT Free	Handoff summary pasted into a new chat to finish the conclusion
Thursday	Edit a Google Doc with a client	Gemini	Works where the document lives
Friday	Proofread everything	ChatGPT Free, Instant	Fast model; "Show only the lines you changed"
Friday	Check weekly usage	Claude Settings, Usage	Sees the weekly reset time; plans Monday's heavy work after it

Result: she rarely hits a weekly limit, and when she hits a session limit, work continues elsewhere within a minute.

Track Your Usage

- Check each service’s usage screen weekly for a month.
- Notice which tasks drain your allowance.
- Adjust habits, or the plan, based on what you see.

18.1 Where to look

SERVICE	LOCATION
ChatGPT	Limit and reset messages in the chat; plan details in Settings
Claude	Settings, Usage (session and weekly bars, reset times, usage credits)
Gemini	gemini.google.com, Settings, Usage Limits
Grok	Settings, Usage (percentage, product breakdown, reset date, credits)
Perplexity	In-app indicators and account settings

18.2 A simple usage log

Keep a note for two weeks:

DATE	SERVICE	TASK	HIT A LIMIT?	WHAT USED THE MOST?
Mon	Claude	Blog draft	No	Long chat with three files
Tue	ChatGPT	Images for slides	Yes, image limit	Many regenerations

After two weeks, the pattern is usually obvious: one kind of task (often images, research or very long chats) causes most limit hits. Fix that habit first.

Troubleshooting

PROBLEM	LIKELY CAUSE	WHAT TO DO
Hit the limit after only a few messages	Long chat, big files or a heavy model or tool	New chat with a handoff summary; trim files; use a lighter model
AI forgets early instructions	Context window filling up	Restate key instructions, or move to a new chat with a summary
Plus or Pro answers feel the same as before	Using the default level	Pick a higher thinking level in the model picker
Paid for a plan but still on Free	Signed in with a different account	Check the account email and sign-in method; restore purchases in the app
Weekly limit gone by Wednesday	Heavy media, research or coding early in the week	Check the usage breakdown; spread heavy tasks; consider top-ups or a higher tier
Gemini Deep Research unavailable	High demand on accounts without a plan	Try later, or use another service's research tool
Usage credits charged unexpectedly	Credits enabled with no spend limit	Set a monthly spend limit or turn off auto-reload
Answers too long	No length instruction	Add length limits, or set them in custom instructions
Different pages show different prices	Rollouts and regional pricing	Trust the checkout page and your account's billing settings

Quick Reference

- **Costs most:** long chats, big files, thinking modes, research, images, video
- **Plan first:** goal, audience, materials, format, avoid, quality check
- **Batch:** number related tasks; set a length per item; redo one item, not all
- **New topic, new chat;** long chat, handoff summary
- **Files:** send the excerpt, give page numbers, upload once to a project
- **Model ladder:** fast model first; step up only when needed

- **ChatGPT:** Free unlimited everyday text; Plus \$20; Thinking levels by plan
- **Claude:** five-hour session plus weekly; Pro \$20 or \$17 annually; Settings, Usage
- **Gemini:** five-hour refresh plus weekly; Plus 2x, Pro 4x, Ultra 5x or 20x Pro
- **Grok:** one weekly pool; Plus \$8, Premium \$30, Super \$100, Ultra \$200
- **Perplexity:** Best mode without quota limits; Pro Search extended on Pro
- **Rotate:** handoff summary to another service; don't evade limits with extra accounts

Sources (checked October 16 and 17, 2026)

- OpenAI: [ChatGPT pricing](#); [GPT-6 and other models in ChatGPT](#); [ChatGPT Free Tier FAQ](#); [What is ChatGPT Plus?](#)
- Anthropic: [Claude pricing](#); [Usage limit best practices](#); [What is the Pro plan?](#); [Manage usage credits](#)
- Google: [Gemini Apps limits and feature availability](#); [Google AI subscriptions](#)
- xAI: [Pricing](#); [Grok FAQ](#); [Usage limits](#)
- Perplexity: [Which plan is right for you?](#); [What is Perplexity Pro?](#)

Glossary

- **Batching:** combining several related tasks into one message.

- **Cache:** stored content the service can reuse more cheaply, such as Claude project files.

- **Context window:** the most text (in tokens) a model can consider at once.

- **Deep Research:** a tool that runs many searches and reasoning steps to write a report; among the most expensive features.

- **Fallback model:** a smaller model a service switches you to after a limit.

- **Handoff summary:** a short summary of a chat's goal, decisions and next steps, used to continue in a new chat or service.

- **Session limit:** a usage limit that resets after a short window, such as five hours.

- **Thinking level (effort):** how much reasoning a model does before answering; higher levels use more allowance.

- **Token:** a chunk of text, roughly three-quarters of an English word on average.

- **Top-up or usage credits:** paid extra usage after your plan's allowance runs out.

- **Usage pool:** one shared allowance spent across several products, as on Grok.

- **Weekly limit:** a cap on total usage per week, on top of shorter session limits.
