Heavy AI User's
Cost-Discipline Playbook

Nine disciplines that cut typical AI bills 30–60% without using AI less. Use the checklist on Friday. Use the prompts and routing table on Monday.

☐	**1. Compress your context before pasting** Paste the relevant 3 pages, not the whole 30-page doc. If you need the full thing, summarize once and work from the summary.	**Saves 40–70% input tokens**
☐	**2. Paste text, not screenshots** A screenshot of an error / table / formula costs the AI ~1,000–2,000 tokens to process. The same as text is ~30. Type or copy the text instead.	**10–50× cheaper input**
☐	**3. Kill long-running chats. Start fresh.** Every turn re-sends the entire chat history. Switch topic = start a new chat.	**5–10× lower per-turn cost**
☐	**4. Default to the cheapest model that works** Haiku / Sonnet / mini handle most tasks. Reserve Opus / GPT-5 for long reasoning, complex code, edge-case judgment.	**5–30× cheaper per token**
☐	**5. Cap output length explicitly. Every time.** Add 'in under 100 words' or 'in 3 bullets' to every prompt that doesn't need to be long.	**Cuts output tokens 60–80%**
☐	**6. Fix the prompt, don't regenerate the answer** Bad answer? Identify what was missing and rewrite the prompt. Don't hit 'try again' 5 times.	**Saves 4× the regeneration cost**
☐	**7. Use prompt caching where supported** If you keep sending the same large brief, look up caching for your tool (Claude, ChatGPT, API). Cached input is billed at ~10% cost.	**Up to 90% off repeated context**
☐	**8. Never let an agent run unsupervised** Set hard step / time / dollar caps on any agent. Watch the first runs of any new task end-to-end.	**Prevents $50 surprise spends**
☐	**9. Run a 5-minute weekly cost audit** Friday afternoon. Open billing/usage. Compare to last week. Identify the discipline that slipped.	**Catches drift early**

Companion to the full article on aiforyourday.com — "How I Burned Through My $200/Month AI Budget in a Week."

Copy-Ready Prompts

Three prompt templates that implement disciplines 1, 2, and 4. Paste, fill in the brackets, run.

Context compression — summarize before you paste

When you've got a long doc but only need a slice. Summarize once, work from the summary.

```
Summarize the document below in under 300 words. Preserve anything that bears on [your topic]. Drop everything else. End with a one-line note on what was cut.

Document:
[paste here]
```

Summarize-then-drop — for long chats that drifted

Run at the bottom of a long chat to extract decisions. Paste output into a fresh chat.

```
Summarize this conversation into: (a) the 3-5 decisions we reached, (b) open questions, (c) what I should paste into a fresh chat if I want to continue tomorrow. Under 200 words.
```

Length cap — drop this into every prompt

Pick the cap that fits. The output is almost always tighter — and cheaper — with the cap on.

```
...your prompt here...

Keep the response under 100 words. No preamble. No caveats unless they materially change the answer.
```

Model-Routing Cheat Sheet

Default to the cheapest model that works. Switch up only when the task demands it.

Task	Use	Avoid
Drafting an email, summarizing a meeting, rewriting a paragraph	Haiku / GPT-mini / Flash	Opus / GPT-5
Long-document analysis, careful reasoning, nuanced writing	Sonnet / GPT-5	Haiku for accuracy-critical work
Complex code, multi-step reasoning, edge-case judgment	Opus / GPT-5 / DeepSeek R1	Smaller models — false economy
Quick factual lookup with current data	Perplexity (free)	Any model from training data alone

The 5-Minute Weekly Audit

Friday afternoon. Five steps. Pays for itself in the first week.

1	Open your AI tool's billing or usage page. (Copilot: admin center. ChatGPT: settings → data controls or admin. Claude: usage. API: console.)
2	Note this week's number. Compare to last week. Up? Down? Roughly flat?
3	If up: which of the 9 disciplines did you let slip? Most expensive habits are obvious the moment you look (the day you pasted that 60-page contract three times).
4	If down or flat: keep going. Note what worked so you keep doing it.
5	Once a quarter, ask: is my tier still right? If consistently under budget, downgrade. If over after discipline, upgrade.