# AI API bill checklist

Ten checks for a Claude or OpenAI API bill that's higher than you expected

**NeverPayTwice**
neverpaytwice.pages.dev

## Where the money goes

- [ ] **Thinking is billed as output, even when you don't see it.** On current Claude models thinking is on by default; lower the `effort` setting for simple tasks.
- [ ] **The prompt cache never hits.** If `cache_read_input_tokens` stays at 0 on requests that share a long beginning, something early in the prompt (a timestamp, an ID, reordered JSON) keeps changing it.
- [ ] **Every turn resends the whole conversation.** Turn 30 pays for turns 1–29 again as input. Cache the history, keep tool results short, start fresh when old context no longer matters.
- [ ] **Cache writes cost more than plain input** (1.25× for the 5-minute cache, 2× for the 1-hour cache on Claude). Writing often and reading rarely costs more than no cache.
- [ ] **Fast mode doubles the price** on Claude Opus 5.5. Check `usage.speed` on requests where nobody is waiting.
- [ ] **Work that could wait isn't batched.** Anthropic's Message Batches API is half price; most batches finish within an hour.
- [ ] **The model is bigger than the job.** Compare cost per finished task, not per request.

## Paying twice

- [ ] **The same request, sent again.** Double-clicks, refreshes, re-run jobs and agent steps each buy the answer again. Group requests by a fingerprint of their body to find them.
- [ ] **Retries after client timeouts.** Anthropic charges a request that was on track to succeed even if your client gave up, so a retry after a timeout can be paid twice. SDKs retry some failures automatically (twice by default on Anthropic's).
- [ ] **Log every request ID** ( `request-id` on Claude, `x-request-id` on OpenAI) with its time, model and `usage`. Without IDs, a billing question can't be checked.

**Think a charge is wrong?** Send the provider one message listing request IDs, times, models, token counts and the invoice line, and ask a specific question. Anthropic says failed requests aren't charged; credit purchases are non-refundable. Whether anything is refunded is the provider's decision. Guide and message template: neverpaytwice.pages.dev/dispute-ai-api-charge

---

Free scripts for each check (MIT): github.com/neverpaytwice/ai-token-cost-scripts · Free Claude Code token counter: neverpaytwice.pages.dev/count

NeverPayTwice ($9 desktop app) keeps a receipt for every AI request, flags repeat charges and keeps a refund ledger, on your own computer: neverpaytwice.pages.dev

Facts from Anthropic's and OpenAI's documentation and help pages, checked November 15, 2026. Prices and policies change; check the current pages. Share freely.