

Claude Opus 4.6 Medium effort

June 17, 2026

/

Claude finished the response

I would like you to look at my chats with you from June 1st and June 2nd. On those days I used Opus 4.8 with a High effort setting. On both June 1st and June 2nd, I did several workflows in my Project on Fourteenth Century Transcription, multiple times getting a workflow of all three stages of my Project process. Today, June 17th, I did a chat using Opus 4.6 on the Medium effort setting and I exceeded my usage limit in the middle of Stage 2 of my Project process. Please explain why the usage limits are getting hit much more quickly today compared to what was happening on June 1st.

5:58 PM

Investigated model performance differences and usage limit variations

Investigated model performance differences and usage limit variations

Good question, Jonathan. I can see the pattern clearly from the chat histories. There are a few factors at play, and they compound:

Model tier matters for limits. Opus 4.8 and Opus 4.6 may carry different rate-limit allocations. Anthropic adjusts usage caps by model — higher-tier models don't necessarily have lower limits, but the token accounting can differ. The key variable, though, is likely the next point.

Your paleography workflow is exceptionally token-intensive. Looking at the June 1–2 sessions versus today's, the pattern is consistent: each session involves multiple rounds of Python/PIL image processing (bash tool calls), viewing dozens of cropped images (each image consumes a large number of tokens in the context window), and then producing long structured outputs. In today's June 17 session, I can see the Stage 2 reconciliation alone involved at least 10 targeted crop-and-view cycles for individual cruxes — each image view is a substantial token cost. The Stage 1 had already done a full set of line-by-line crops. That's 20+ image ingestions in a single conversation, plus all the text.

The most likely explanation is a change in usage limits between June 1 and June 17. Anthropic periodically adjusts rate limits for different plan tiers, and these can change

without much notice. What you could do on June 1 may simply reflect a more generous allocation at that time. Let me check if there's been any announced change.

Synthesized usage pattern discrepancies across model versions

Synthesized usage pattern discrepancies across model versions

Here's what's going on, Jonathan. The answer is straightforward once you see how the usage accounting works.

Usage limits are cost-based, not message-based. Anthropic meters your Pro plan by the dollar-equivalent cost of tokens consumed, not by counting messages. Every token — input and output — carries a price that varies by model. This is the critical detail.

Opus 4.8 API pricing is \$10 input / \$50 output per million tokens. Opus 4.6 is \$5 / \$25. So at the API-rate level, Opus 4.8 is twice as expensive per token as Opus 4.6. But here's the thing that seems counterintuitive: on June 1–2 you were using Opus 4.8 at High effort and getting *more* done before hitting limits, while today on Opus 4.6 at Medium effort you ran out faster. That doesn't track with pure per-token pricing — it suggests the explanation lies elsewhere.

The most likely culprit is your weekly cap. Your Pro plan has two overlapping limits: a 5-hour rolling window *and* a weekly cap. If your weekly limit reset happened to fall just before June 1, you would have had a full weekly budget available across those two days. Today, June 17, you may be working against a weekly budget that's already been partially consumed by other usage earlier this week (or by Claude Code / Cowork if you use those, since they share the same pool).

Image ingestion is a massive hidden cost in your workflow. Each manuscript crop you view gets tokenized as an image — at high resolution, a single image can cost thousands of tokens. Your June 17 session shows the classic pattern: full-page enhancement → 10 line-by-line crops in Stage 1 → then 10+ targeted crux crops in Stage 2. That's 20+ image ingestions in a single conversation, and each one is being sent back through the context window on every subsequent turn, compounding the cost. This pattern is consistent across your June 1–2 sessions too, but those sessions may have benefited from a freshly reset weekly budget.

What you can do practically:

Start a new conversation between stages rather than running all three in one chat — this prevents the accumulated image context from inflating every subsequent turn. You can also check your usage dashboard (Settings → Usage on claude.ai) to see where you stand in your weekly cycle and time your heavy transcription sessions accordingly.

Doesn't it seem like the "most likely culprit" is that Anthropic adjusted the rate limit for my plan tier between June 1st and June 17th? Are you allowed to inform me of when that rate limit changed and how it changed?

6:01 PM

Acknowledged knowledge gaps regarding internal rate limit changes

Acknowledged knowledge gaps regarding internal rate limit changes

You're right to push back on that — it's a reasonable hypothesis, and I shouldn't have buried it behind the weekly-cap explanation.

But the honest answer is: I don't know, and I don't have access to that information. I can't see Anthropic's internal rate-limit configurations, I don't receive changelogs about plan-tier adjustments, and Anthropic doesn't publish granular per-model token budgets for Pro subscribers. What I can tell you from the search results is that every *announced* change between May and mid-June moved limits in the generous direction — the May 6 doubling, a 50% promotional increase on May 13 (described as running through July 13). Nothing I found documents a *reduction* in that window.

That said, Anthropic has historically adjusted limits without public announcement, and the promotional increase could have expired or been modified quietly. I genuinely can't rule it out or confirm it.

What I'd suggest: check Settings → Usage on claude.ai, which should show your current limit status and reset times. If the numbers look different from what you experienced two weeks ago, that's your evidence. And if you want a definitive answer about whether the allocation changed, Anthropic's support team (via the help menu in claude.ai) would be the only source that could actually confirm it.

Want to be notified when Claude responds?

Claude Fable 5 is currently unavailable.

[Learn more\(opens in new tab\)](#)

[Claude is AI and can make mistakes. Please double-check responses.](#)