r/ClaudeAI • 2 days ago
sixbillionthsheep MOD Mod 🥽 Top 1% Commenter

Status Report - Claude Performance Megathread – Week of Apr 27– May 7, 2025

Status Report

Notable addition to report this week: Possible workarounds found in comments or online

Errata: *Title should be Week of Apr 27 - May 4, 2025*
Disclaimer: *This report is generated entirely by AI. It may contain hallucinations. Please report any to mods.*

This week's Performance Megathread here: https://www.reddit.com/r/ClaudeAI/comments/1keg4za/megathread_for_claude_performance_discussion/
Last week's Status Report is here: https://www.reddit.com/r/ClaudeAI/comments/1k8zsxl/status_report_claude_performance_megathread_week/

🔍Executive Summary

During Apr 27–May 4, Claude users reported a **sharp spike in premature "usage-limit reached" errors**, shorter "extended thinking", and reduced coding quality. **Negative comments outnumbered positive ~4:1**, with a dominant concern around **unexpected rate-limit behavior**. External sources confirm two brief **service incidents** and a major change to **cache-aware quota logic** that likely caused unintended throttling—especially for Pro users.

📊Key Performance Observations (From Reddit Comments)

Category	Main Observations
🧮**Usage-limit / Quota Issues**	Users on **Pro and Max hit limits after 1–3 prompts**, even with no tools used. Long cooldowns (5–10h), with **Sonnet/Haiku all locked**. Error text: *"Due to unexpected capacity constraints…"* appeared frequently.
🌐**Capacity / Availability**	94%+ failure rate for some EU users. Web/macOS login errors while iOS worked. **Status page remained "green"** during these failures.
⏳**Extended Thinking**	Multiple users observed **Claude thinking for <10s** vs >30s before. Shorter, less nuanced answers.
👨‍💻**Coding Accuracy & Tools**	Code snippets missing completions. Refusals to read uploaded files. Issues with new **artifact layout**. Pro users frustrated by the 500kB token cap.
👍**Positive Upticks (Minority)**	Some users said cache updates gave them **2–3× more usage**. Others praised **Claude's coding quality**. Max users happy with 19k-word outputs.

issue	possible **privacy leak**.

📉Overall Sentiment (From Comments)

- 🟥**Negative (~80%)**: Frustration, cancellation threats, "scam" accusations.
- 🟨**Neutral (~10%)**: Diagnostic discussion and cache behaviour analysis.
- 🟩**Positive (~10–20%)**: Mostly limited to Max-tier users and power users who adapted.

Tone evolved from **confusion → diagnosis → anger**. Most negativity peaked **May 1–3**, aligning with known outages and API changes.

📌Recurring Themes

1. **Quota opacity & early lockouts** (most common)
2. **"Capacity constraints" loop** — blocked access for hours
3. **Buggy coding / file handling**
4. **Sonnet / 3.7 perceived as degraded**
5. **Unclear caching & tool token effects**

🛠️Possible Workarounds

Problem	**Workaround**
Limit reached too fast	Use **project-level file cache**. Files inside a Claude "project" reportedly no longer count toward token limits.
Unknown quota usage	Use the **Claude Usage Tracker browser extension**.
Large file uploads too expensive	Split code into **smaller files** before uploading.
Capacity error loop	Switch to **Bedrock Claude endpoint** or fallback to Gemini 2.5 temporarily.
High tool token cost	Add header: `token-efficient-tools-2025-02-19` to Claude API calls.

✨Notable Positive Feedback

> “Lately Claude is far superior to ChatGPT for vibe-coding… All in all I am **very happy** with Claude (for the moment).”
>
> “Cache change gives me 2–3x more usage on long conversations.”

> "Two prompts in a new chat, no context… **rate limited**. Can't even use Haiku."
>
> "Answers are now much shorter, and Claude gives up after one attempt."
>
> "Pro user, and I'm locked out after three messages. What's going on?"

🌐External Context & Confirmations

Source	Summary	Link to Reddit Complaints
🛠️**Anthropic Status (Apr 29 & May 1**)	Sonnet 3.7 had **elevated error rates** (Apr 29), followed by **site-wide access issues** (May 1).	Matches capacity error loop reported Apr 29–May 2.
🧮**API Release Notes (May 1**)	Introduced **cache-aware rate limits**, and **separate input/output TPMs**.	Matches sudden change in token behavior and premature lockouts.
📝**Anthropic Blog (Apr**)	Introduced **"token-efficient" tool handling**, cache-aware logic, and guidance for reducing token burn.	Matches positive reports from users who adapted.
💰 **TechCrunch (Apr 9**)	Launch of **Claude Max ($100–$200/ month)** tiers.	Timing fueled user suspicion that Pro degradation was deliberate. No evidence this is true.
📄**Help Center (Updated May 3**)	Pro usage limits described as "**variable**".	Confirms system is dynamic, not fixed. Supports misconfigured quota theory.

⚠️**Note:** No official acknowledgment yet of the **possible API prompt leak**. Not found in the status page or public announcements.

🧩Emerging Issue to Watch

- **Privacy Bug?** One user saw **other users' prompts** in their Claude output via API. No confirmation yet.
- **Shared quota across models?** Users report Sonnet and Haiku lock simultaneously — **not documented** anywhere official.

✅Bottom Line

- The most likely cause of recent issues is **misconfigured cache-aware limits** rolled out Apr 29–May 1.
- **No evidence** that Claude Pro was intentionally degraded, but **poor communication** and **opaque**

fully solve the unpredictability.

- Further updates from Anthropic are needed, especially regarding the **prompt leak report** and **shared model quotas**.

24 5 Share

Join the conversation

Sort by: **Best** Search Comments

sixbillionthsheep MOD • 2d ago •
Mod Top 1% Commenter

Please direct any discussion of your performance experiences/observations to the Megathread here: https://www.reddit.com/r/ClaudeAI/comments/1keg4za/megathread_for_claude_performance_discussion/

Vote

djc0 • 2d ago
Valued Contributor Top 1% Commenter

I like this kind of reporting

7 Reply

inventor_black • 2d ago
Intermediate AI

Great way to not have to search to know how everyone's experience is going.

LLM performance is the most inconsistent thing introduced to the realm of computing. (indeterministic + they adjust the model)

4 Reply

ctrl-brk • 2d ago
Valued Contributor Top 1% Commenter

Bluesky please, not Twitter. Reddit, not Discord.

This is an interesting use of Claude and unique, haven't seen any other sub doing this.

3 Reply

Lost_Cyborg • 2d ago

nobody uses bluesky