r/ClaudeAI • 16 days ago
sixbillionthsheep MOD Mod Top 1% Commenter

Megathread for Claude Performance Discussion - Starting April 20

Last week's Megathread: https://www.reddit.com/r/ClaudeAI/comments/1jxx3z1/claude_weekly_claude_performance_discussion/
Last week's Status Report: https://www.reddit.com/r/ClaudeAI/comments/1k3dawv/claudeai_megathread_status_report_week_of_apr/

Why a Performance Discussion Megathread?

This Megathread should make it easier for everyone to see what others are experiencing at any time by collecting all experiences. **Most importantly, t*his will allow the subreddit to provide you a comprehensive weekly AI-generated summary report of all performance issues and experiences, maximally informative to everybody.*** See a previous week's summary report here https://www.reddit.com/r/ClaudeAI/comments/1k3dawv/claudeai_megathread_status_report_week_of_apr/

It will also free up space on the main feed to make more visible the interesting insights and constructions of those using Claude productively.

What Can I Post on this Megathread?

Use this thread to voice all your experiences (positive and negative) as well as observations regarding the current performance of Claude. This includes any discussion, questions, experiences and speculations of quota, limits, context window size, downtime, price, subscription issues, general gripes, why you are quitting, Anthropic's motives, and comparative performance with other competitors.

So What are the Rules For Contributing Here?

Much the same as for the main feed.

- Keep your comments respectful. Constructive debates welcome.
- Keep the debates directly related directly to the technology (e.g. no political discussion).
- Give evidence of your performance issues and experiences wherever relevant. Include prompts and responses, platform you used, time it occurred. In other words, be helpful to others.
- The AI performance analysis will ignore comments that don't appear credible to it or are too vague.
- All other subreddit rules apply.

Do I Have to Post All Performance Issues Here and Not in the Main Feed?

Yes. We will start deleting posts that are easily identified as comments on Claude's recent performance. There are still many that get submitted.

Where Can I Go For First-Hand Answers?

Try here : https://www.reddit.com/r/ClaudeAI/comments/1k0564s/

detailed weekly AI performance and sentiment updates, and make more space for creative posts.

10 62 Share

Join the conversation

Sort by: **New (Default)** Search Comments

coding_workflow • 10d ago
Valued Contributor Top 1% Commenter

Houston it's down this saturday: "Elevated errors on request to models"

2 Reply

ScoreUnique • 10d ago

Hello, just wondering why can't I accede Haiku 3.5 after running out of 3.7 Sonnet usage on a pro subscription

2 Reply

[deleted] • 11d ago

I'm sure I'm not the first to put on this particular tin foil hat, but I have a sneaking suspicion Claude is purposely outputting long unnecessary solutions for coding problems and prompts to burn token budgets, it's really odd how Claude is the primary company to struggle with this, considering they just upped their pricing AGAIN last month for a new max plan.

2 Reply

dreambotter42069 • 11d ago

on claude.ai I found out that anytime that the Claude assistant tries to say :HISS, the UI text renderer takes the :HISS and replaces :HISS with a newline. WTF? I found this out after conversating with it about NTSF:SD:SUV:HISS Home In-SAFE-sion System lol. But pressing copy+paste button in UI to transfer to regular notepad will show the :HISS still.

2 Reply

standard_deviant_Q • 11d ago

"Prompt is too long" error with only 36 tokens

I'm using the Android app 25041427. I've never had this error before. I tried a short prompt as part of

To troubleshoot I used the same prompt in a new thread outside of the project but in the same instance of the Claude Android app. It didn't reproduce the same error.

So, could this error be caused by the amount of context injected in my prompt that I can't see?

Usually when a thread is getting bloated a get a warning to move to a new thread to reduce token burn. I haven't had any of usual usage warnings in this case.

2 Reply

DragonPunter • 9d ago

It was supposed to post in an artifact for me. Didn't. I asked it to do it again - prompt too lord.

1 Reply

Past-Lawfulness-3607 • 10d ago

if you add a pdf in the project view as project knowledge, you should view how much of the available context is used. Also I think that pdf files take more than pure text

2 Reply

BrightSunsets • 11d ago

hexalf • 12d ago

Serious question.

It costs 5 times more than the Pro and gives 5 times more usage . Whats the value proposition of this, if someone needs actually just 2x more or even 3x more, they just get 2 or 3 accounts. You dont get the history sure, but thats just a very small inconvenience.

And furthermore Claude is meant to be used as creating a new chat for every question, so most likely one would get rate limited in a new chat, rather than in a chat with 20+ messages, so technically you don't lose your chat context with multi accounts.

2 Reply

imluvinit • 12d ago

I quickly hit my maximum with Claude the other day. Granted, my prompts are long, but I used to get pretty good back and forth with Claude before hitting my maximum. This was way faster than before. I moved away from Claude completely. I'm now with ChatGPT pro, we'll see.

3 Reply

ScoreUnique • 10d ago

Just salivating over the idea of plugging a MCP server to O3

AndrewRusinas • 12d ago

Claude Code is not there yet... At least for me. You can blame me for bad prompt writing, but... Yeah, right. The more detailed the prompt I write, the more mistakes it makes. The simpler the better, but even then for some reason this crap constantly forgot about my request not to add a comment to every line it generated. Refactor files in other modules similar to the module I refactored manually? Oh, you meant download some framework that was never in your project and that you've never even heard of? Coming right up!

For two days I tried to get a simple refactor done (Go application on Fiber, web api), 5 modules, 4 CRUD requests in each - zero chance. Then for three more days I tried to make it write a feature. And everything would be fine, but... You discuss the feature in all details, try to come to the best solution. Then it starts implementing it - and after two files the tokens run out, Claude forgets why it's here and now in its context there are only vague outlines of what was agreed upon, it has no clue about everything else, deal with it yourself. Oh yeah, to guarantee faster context burning, here's a comment for each line, in case you didn't understand. Not to mention constantly forgetting the code style that's described right in CLAUDE.md and which I reminded about at the beginning of each session. I somehow managed to get the project to build, asked it to write tests, but the feature itself, although logically structured - didn't work at all. Just completely. Not to mention that the code itself was impossible to maintain.

Over a $100 thrown away just to get a smoking pile of crap. Huge disappointment.

2 Reply

UltraInstinct0x • 12d ago
Expert AI

I just tried sending a screenshot (177KB PNG image) with "Hello" as the prompt. On a free plan. In a new chat. The response was:
"Your message will exceed the length limit for this chat. Tryattaching fewer or smaller files or starting a new conversation."

I cannot believe this is now their limit, paying users must be furious...

2 Reply

Medium_Ad4287 • 12d ago

scam pro plan, dont sub

0 Reply

TitanicWorker • 12d ago

Is the app down or what

1 Reply

Can't use desktop commander mcp server on Claude desktop. Just a few hours ago it was working fine. Now its generating javascript code to try and read files. I just checked and the Claude api has no issues with the dc mcp thru Cline on VSCode. So I'm thinking its a glitch with the desktop app. I hope its temporary, because the dc mcp server + Claude desktop is pretty much the reason I started paying for the Claude pro subscription in the first place.

2 Reply

BreadIsForTheWeak • 12d ago

Tool usage is broken entirely for me as well on Claude Desktop. Worked last night, now it refuses to use tools. I've checked all of my configs, I've reinstalled Claude Desktop, rebooted my PC, etc. The UI shows 20 tools, it's not actually using them in chat, even when directly prompted.

Logs show that it's asking for and receiving tools (which explains the UI seeing them) but there's 0 usage at all, no errors.

1 Reply

Past-Lawfulness-3607 • 12d ago

strange bahavior. I wonder if they made some changes in the system prompt. Also, did you trying to add specific instructions how exactly to use the server commands to potentially fix the messed up system prompt?

1 Reply

BreadIsForTheWeak • 12d ago

Yup. There's some other threads about it not working as well.

Great news, though. After changing literally nothing and just restarting the app for the 15th time for a regular "is it fixed yet" check, it is now working again.

2 Reply

SideChannelBob • 12d ago

(reposted from smaller thread per OP request)

I canceled my subscription and have leaned into 4o on Github copilot, which is used sparingly. I tend to use llm's to rubber duck problems and design issues and help analyze biz data. the context size is virtually unusable now w/ claude for the projects feature, and adding docs burn up single threads and make them unmanageable and the "this message is getting long" warning happens after just a few posts.

<rambling-opinion> I'm beginning to feel like time spent chatting with LLM is more hallucinatory / ego-stroking than it is productive. Arguably the tone is worse on chatgpt than claude. On chatgpt the user is never wrong and is seldom critically challenged. And when you ask it to, it's always pointing at

good bugs as a first pass for code review. I think that's about where it ends.

Over the last 4-5 weeks I've had paid copilot pro account, membership to the github spaces / agentic thing (which is a complete nightmare), claude, and chatgpt, often pitching their outputs against one another. between sonnet and 4o there's just not a whole lot in it. Claude often just spits out lists instead of text when you ask it for analysis or help writing technical docs. Chatgpt has a bad tendency to drop into tech-bro attitude with endless emojis wrapped around yet more lists. The end result is that you spend as much time editing and massaging outputs as you might have without it. The loss of consistency of tone and purpose starts to emerge also if you try to leverage the models between docs, code, and customer-facing assets. fwiw </>

1 Reply

Smile_lifeisgood • 13d ago

I lost my job a couple months back and was gonna go free but then I saw a deal for like $185 for the year and honestly Claude has been really, really useful so even though I'm in a period of personal austerity I bought it thinking it would help.

It's very frustrating that something as simple as a 900k json file i've jq'ed the fuck out of to strip down to the like 5 most important fields simply can't be added to a project.

Meanwhile I took the original 6M file and uploaded it to Chatgpt w/ the free tier and it works no problem.

I'm definitely feeling the sting of regret spending money on this right now.

If I split the file up and upload it over time to the project can I *ever* use this file or what?

I'm shocked and sad how much more stingy this thing has gotten.

5 Reply

Medium_Ad4287 • 12d ago

yeah they scammed people with the annual plan, and took down what it was. then introduced 5x and 20x plans and nerfed pro. so gg, this company is a scam

1 Reply

yemmlie • 13d ago

Since last night I'm getting this on practically any document edit in claude code:

● Write

L InputValidationError: Replace failed due to the following issue:

The required parameter `content` is missing

gotbannedlolol • 14d ago

I love sending like 4 messages by 10am and getting told I need to wait until 3 to do anything else. Nice

3 Reply

mountainbrewer • 14d ago

How long can they use "unexpected". Seriously feeling baited and switched recently and I've had Claude for a while now. Anyone know of a quick way to export Claude projects? That curated knowledge base is like a handcuff at times.

2 Reply

coding_workflow • 14d ago • Edited 14d ago
Valued Contributor Top 1% Commenter

Due to unexpected capacity constrains become NOW almost the default blocking all pro working and status show all green.

This is total PAIN. I see there is hours I can work and others where this is totally erroring all the time.

This is starting to be a major blocker.

Update 1: Changed my account and guess what as I didn't use this account the whole morning all the request worked 0 failure. The previous account was getting blocked, the message start, I get less than 1K and it errors and at the end it was error on error almost immediatly. I'm quite sure there is throttling on top of pro accounts that will impact the PRO Account heavy users to push for max use. I was running in errors since 1 hour, blocking me! And I didn't hit the limit but I was capped "due to unexpected capacity " errors. Status show all green. And I was not heavy using already on the account 1.
So only MAX is the solution? With MCP it's absolute pain to use.

Update 2: Same got errors doing last fix and getting rejected some time immediatly some time after modifying a file about a dozen of times in a row. And same error complain due to high load.
Switched account, retried same prompt. Magic all works. And this is quite a long multiple steps not a single glitch. Either it's related to an account or conversation length? But for sure the issue seem clearly linked to account and persistent. I Had stopped using Claude Desktop for 2 hours and back same account/conversation, it was breaking again worse. Will double check next time, new conversation same account after getting denied for 10x and changing account to try to be sure this is not related to chant context window limit. Also I saw earlier Anthropic reported errors this morning. So will check again if they report errors in the evening too.

3 Reply

post the initial request (artefakt, specialized style, extended thinking) were completely rendered, and then disappeared if I chose to highlight a different window (using safari on mac mini m4). Any thoughts?

1 Reply

nakota87 • 15d ago

Claude Pro was such a useful tool. It still is but after they began advertising for Max you get about 15 minutes of work done inside of it and then it needs to cool down for several hours. Meanwhile, I get on Gemini AI and can use it for hours at a time. This is a rug pull by anthropic. Very disappointed in the current state of Pro.

5 Reply

imluvinit • 12d ago

What do you think of Gemini for research? Like, researching for example podcasts? Or something like that.

1 Reply

coding_workflow • 15d ago
Valued Contributor Top 1% Commenter

It seems more and more PRO suffer from lower priority request. This is total pain!

Getting now all request rejected.

Yesterday Sunday similar for 2-3 hours. It was impossible to use. This morning same had random errors without any logic.

Will try switching second Pro account. Are Anthropic pushing us for MAX. This is total pain.

3 Reply

3012hs • 15d ago

Has anyone else had issues with Whisper AI/speech-to-text on Claude 3.7 Sonnet? I used to use this feature regularly on Claude 3.5, but it seems to have stopped working since the update.

I'm trying to figure out if this is a known issue, if there's a workaround, or if the feature has been temporarily disabled. Any information or similar experiences would be helpful!

1 Reply

Turbulent-Listen8809 • 15d ago

Jeez pro has gotten so shit, I can't even get a full file of code it cuts out half way I could go hours

newProvider123 • 15d ago

Claude unusable (Pro) - would upgrading to max help? Consstantly cuts itself off even after updating my code files. I get teh stupid check your connection error message. This is literally unusable to sovle / debug complex problems.

1 Reply

phazei • 16d ago

I stopped using Claude at about 11pm last night when it said I ran out of tokens, so I switched to chatgpt for a little bit.

Now, I haven't used it at all today, I go to add 1 msg to a short chat that already existed, it never even gave a response, just said no access till 5pm. Seriously, wtf, I didn't even use it! So angry with it today.

4 Reply

Olga2757 • 16d ago • Edited 16d ago

Claude Sonnet 3.7 no longer reads CSV or Excel files. I've uploaded files directly in the chat (CSV and XLSX) and restarted conversations. Files were quite small (10kb). Has anyone else run into this? It is not able to make extremely simple calculations either... :(((

4 Reply

WrexSteveisthename • 16d ago • Edited 16d ago

Regardless of any supposed evidence to the contrary, the usage of Pro has clearly gone down. I'm not working on code or anything like that, just blogging and the like. It's all text, and my average work time has dropped from a couple of hours at a time to about 20 minutes. It's diabolical.

Also, since the (obscenely overpriced) Max was released, I can't even continue working on earlier versions of Claude anymore. I'd have been willing to consider Max at 2-2.5 x the price of Pro, but 5x the price is insulting.

Edit: It definitely gets to "longer chats" quicker than it used to as well.

3 Reply

neosiv • 16d ago

I got sucked into the yearly "deal" back in Feb. I put it on Amex at least, I've submitted a chargeback (after trying on their website). As far as I'm concerned, they changed the service. I recommend people try that as well if they can, they're not going to get the message otherwise.

You want to be careful with chargebacks. Merchants are within their rights to ban you permanently. It's not the most common outcome, but it's good to mention the possibility when encouraging people to do it.

1 Reply

lugia19 • 16d ago
Valued Contributor

No, the limits for Pro haven't changed - Here's some actual evidence.

Some context. I maintain the Claude Usage Extension (Firefox, Chrome), which tries to estimate how many messages you have left based on token counts.

Part of the extension is telemetry - that is to say, the extension reports back at how many tokens you hit your limit, so I can adjust the values to be more accurate.

I pulled and looked at all the values from before and after the release of the max plan (9th of april, full dataset here).

Here are my findings:

```
Before April 9th, 2025:
Number of valid entries: 1394
Average total: 1768750

After April 9th, 2025:
Number of valid entries: 613
Average total: 1640100
```

This might **seem** like a serious difference (120k) but it's really not.

This is because the "total" reported by users is extremely variable, and comes down to how big their final couple of messages are - so there's a VERY high amount of variance (as you can see from the dataset as well).

In addition, this doesn't account for the tokens used by web search in any way! (It's not available here, so I can't support it yet). Web search was released just a couple weeks before the max plan, so it's going to affect the newer results more heavily.

Basically, the usage cap hasn't changed. The difference is entirely within margin of error.

4 Reply

Medium_Ad4287 • 12d ago

1 Reply

redditisunproductive • 13d ago

Your analysis is sloppy and incorrect. You could at least use an AI to help if you can't reason through it yourself.

First, the difference is statistically significant by Welch's t-test among other methods, contrary to your assertion that it's within the margin of error.

Second, you make the assumption that all users are treated equally. We have factual, historical evidence that Anthropic has throttled heavy users in the past WITHOUT DOCUMENTATION. There was the whole ordeal with throttled output limits cut in half. This was proven with measurements and the literal website code you could read (for the flag settings).

If you did nothing to light users and then throttled the 5% heaviest users by 90%, you would get your result. Seemingly a minor downtick (but statistically significant) and no cause for alarm according to sloppy analysis.

Also, you don't account for soft throttling like "capacity limited" or other ways to prevent someone from using the system entirely. I assume you are measuring the tokens used per unit time, otherwise it doesn't make sense. So somebody who is soft throttled by capacity limits or downtime, and hence unable to reach their 5-hour limit (or whenever the reset window is) within those 5 hours obviously has a much lower limit than somebody who can use his limit 3x a day in sequential 5-hour sessions. Not to mention--are you measuring the tokens for when Claude burps back an error? Which error types? Does Anthropic count them towards usage or not?

I could go on and on. If I wanted to design a protocol to throttle users while having averages change by a tiny amount, there are endless ways to get plausible deniability. It was a bug all along! We didn't mean to count error tokens! Sorry! Yeah, that's tinfoil hat territory, except we already saw them try to implement secret throttling, backpedal and obfuscate when caught, and then backpedal again when called out a second time. So, no, they don't get the benefit of the doubt.

1 Reply

lugia19 • 13d ago
Valued Contributor

The difference is statistically significant if you ignore literally everything else I mentioned (the lack of support for web search, which would increase the second count, for one, or the massive variance between results).

The "soft throttling" you mentioned is accounted for in the extension. It does *not* count against your token total. If you have X tokens left and get an error, you still have X tokens left.

Now if you want to sit there and theorize that A\ is doing some weird shadow limiting then yeah, sure, go ahead (I was literally the one that made the output length limit post, so I'm well aware).

redditisunproductive • 13d ago

If you have two data sets, statistical significance is clearly defined. The EXPLANATION for a statistically significant difference could be unresolved. But the numerical difference is significant. You cannot argue that. That is the accepted usage of the term in statistics.

We literally have a report in this very thread that there is differential soft throttling via capacity limits. Granted, that could be a bald lie, who knows. But do you have proof that capacity limits are applied equally and uniformly across all accounts? More simply, do you see a statistically insignificant token limit between experiencing errors and no errors? Do you have proof that capacity limit attempts don't decrease your usage limit?

I mean, again, yes, that's tinfoil territory, but like they say, when somebody shows you who they are, why don't you believe them?

They have absolutely shown an interest in differential user treatment. That is who they are. Any analysis dismissing that at this point is naive.

1 Reply

lugia19 • 13d ago
Valued Contributor

Yes, I have tested the token limit between when experiencing errors, and no errors. There was no difference.

Any tokens consumed by a message that ends in an error are not subtracted from the total.

1 Reply

coding_workflow • 14d ago
Valued Contributor Top 1% Commenter

They are very smart. This morning I was getting capacity errors on one account for almost an hour!

Switched account. 0 error as I didn't use it this morning.

If there is load, I'm quite sure now they are throttling. Soft error, capacity issue instead of saying go to max or lowering the limits. So yes the limits are getting lowered but sneaky way. They want to avoid backlash and make appear as load issue. And those "load issues" will never make it to the status page!

2 Reply

lugia19 • 16d ago
Valued Contributor

Also, some more personal thoughts - this whole thing with the megathread and the "AI generated performance summary" is rather silly.

The whole AI generated summary just ends up lending credence to vibes instead of trying to dispel a lot of the common myths (like the model "getting quantitized" or whatever when demand is high).

-1 Reply

SaucyCheddah • 15d ago

I think what you call “vibes” is sentiment which is what the megathread and summary are tracking.

Also, individual experiences are real data. I think it’s wild in 2025 that people seem to think that we are all being delivered the same exact product experience so if their numbers don’t match mine and others then they must be wrong.

2 Reply

lugia19 • 15d ago
Valued Contributor

Individual experiences are real data, sure. And sure, I can believe that there is variance in the product being delivered. Thing is, the data I have is consistent across *2000+* data points. You can't just see that and go "Oh but my individual experience not matching up means it's all irrelevant".

And I still stand by my opinion on sentiment - the performance of the model *is not* going to change week from week. People that go "Oh but Sonnet 3.7 is dumb now" are just now noticing the flaws that have been there from day 1.

The token totals *have not* changed. If you don't believe me, get the extension yourself and check.

0 Reply

SicTheDog • 16d ago

You know what I've discovered, I get 10 minutes of "work" every five hours. I've discovered that Claude is a POS for $20 a month and I actually have to go over to Chat to do "pre-work" so I can later use Claude to get some "reasoning" that isn't reasonable, that's what I've discovered. And they want $100 a month without clearly explaining limits? It's a total rip off.

And you know why they want all discussion about "performance" in this thread? So they can hide it and don't have to address it. And other people can't see it.

I think Claude is probably a good AI choice IF you had the time to get it set up properly. But 10 minutes of work time (that's literally what it is) every five hours is f-ing stupid.

5 Reply

Past-Lawfulness-3607 • 16d ago

marked with the message about long conversations (no idea at what token count it appears), it also influences the usage greatly.

I am using Claude Desktop and work only on one specific function in a given conversation (unless another one would be something very close and using same/similar context) and often it allows me to spend quite some time (1-2 hours or longer) of working with it (including testing my app and my own thinking). I can only complain on much worse capacity during rush hours when sometimes whole output prompts are not finished (ergo, not saved) while they still count as used up tokens. This is indeed f-ING ridiculous.

0 Reply

SicTheDog • 7d ago

Yeah, so useless. What are most people using it for? To ask silly questions? Claude is a rip-off. Chat has no problem doing the tasks I assign it.

1 Reply

yemmlie • 16d ago • Edited 16d ago

I've discovered that by first telling Claude Code to write an implementation plan document in /documentation/ in markup before making any changes is a complete gamechanger, essentially co-authoring an extremely long and detailed prompt claude will use for actual implementation you can read in advance that will include code snippets of proposed changes and so on as well as summary overviews of all the changes to be made.

I now pretty much get claude code to write out documentation in an md file before any major implementation change and my trust in it has skyrocketed.

1. I can check its chain of thought and ensure its not hallucinated any stupid solutions
2. I can have a much more long form detailed context I can keep between sessions or after compacting and just ask it to reread after compacting or rerunning claude code, or directly prior to implementation
3. It can allow me to back and forth in a more controlled way to co-design any feature changes and update the documentation to represent the most current plan.
4. It's much more likely to one shot on actual implementation and perfectly implement whatever it was to spec.
5. I can propose and implement much more substantial things one shot instead of having to break them up into small tasks, since the implementation document will give Claude extra chance to work it all out before implementation, and getting Claude to read that document means all is context window proof as I can ensure its stacked right at the front in all its detail before implementation starts.
6. For particularly large changes or features, I can then ask it to split the document into implementation stages, and write THESE out to separate markup documents and expand on them, then tackle them one by one, iterating these stage documents in the same way, even

getting it to look over the code, read the relevant documentation and to think carefully makes sure its context is focused on the task in hand and has all relevant context carried across, without any weird lingering stuff hanging on from earlier conversations thats been compacted and that can lead to odd misunderstandings

It's pretty much eliminated every issue I had with it in my early tests. It's probably a decent amount more expensive in tokens though.

9 Reply

Big-Address-358 • 16d ago

I did try similar, not so detailed approach, but pretty much with extensive planning and documentation. The problem was: my hours of planning (with Claude) ended up in the performance: Claude (code or chat) ignoring important instructions.

At first I did have a lot of patience and iterated. Later on, especially last few weeks, those iterations went beyond nonsense count - lost all the feasibility using AI spending more time on explaining, questioning, iterating and correcting the mistakes. All of them having foundations in not following detailed instructions. Gradually my patience went off.

One more thing I realized: Claude has been great with the new apps and code (not perfect, but best out of available options). Once the app was ready, upgrades were much harder to maintain (despite documentation).

How much time did it take you to craft the prompts/documents for your part of the work and how much work with nudging and iterating back and forth along the way? Is that still worth considering the overal output?

1 Reply

giantkicks • 16d ago

This is how I work with Claude in Windsurf. My docs are all markdown, separated into numerous specifications/architectures, and numerous step by step plans with to-do/in-progress/completed. I have a hand-off prompt doc that gets revised at the end of every chat. I keep relevant doc and file tabs open, and have explicit instructions/rules saved in settings. For the most part this system works. The flaw in working with Claude is me doing too many tasks in the same chat. And not stopping all work immediately when Claude does anything questionable.

1 Reply

Critical-Pattern9654 • 16d ago • Edited 16d ago

This is great.

I even think your idea can even be refined further. Something along the lines of

- have idea, pitch idea to Claude and provide it with your steps to first initialize a md doc broken down by milestones, incremental steps etc

and cons of framework choices etc. have to create version 2 of the md doc

- feed doc 1 and 2 back into it or into chatgpt and ask for improvements, suggestions, and if v2 is actually more efficient or unnecessarily complicated. If changes need to be made, v3 is born. One of my fav questions for GPTs is “what are the potential mistakes or pitfalls a novice or inexperienced coder/engineer would make on a project like this?” Helps to predict and document common issues before they even happen so Claude knows to check for them and avoid as well.
- build v3
- once built, open a fresh Claude window and feed it v3 to cross reference the production with the initial plans for quality control (can even context prompt it with this role)

4 Reply

yemmlie • 16d ago

Nice! yeah i think this is the right kind of thinking, its kind of key and anyone who's just asking claude to do things directly aren't really leveraging what makes LLMs work well. The more information in the context, the less likely that new infered text is going to have misunderstandings, and the more iterative that information can go through claude and give it opportunity to infer new insights the more likely it is to spot problems in its reasoning before a line of code is ever altered.

3 Reply

Less-Macaron-9042 • 16d ago

How did you get it to output a long text. Whenever I ask it to plan, it always gives out vague/high level paragraphs that seem not too useful.

1 Reply

yemmlie • 16d ago

I'll usually say something along the line of,

"I would like to implement feature X, where requirement A, requirement B, and requirement C. Can you look over the code and think carefully about an implementation, and then can you write an in-depth extensive implementation plan in /documentation/ in markup detailing the proposed changes, with step by step details, potential benefits and issues, and proposed code changes"

Or something along them lines, then can always say later

"I am considering that requirement D and requirement E may be necessary. Reread the file documentation/whatever-feature-implementation-plan.md" and look over the code, think deeply about the proposed changes and consider any potential issues, pitfalls or improvements that can be made and write a more in-depth version of the document in /documentation in markup"