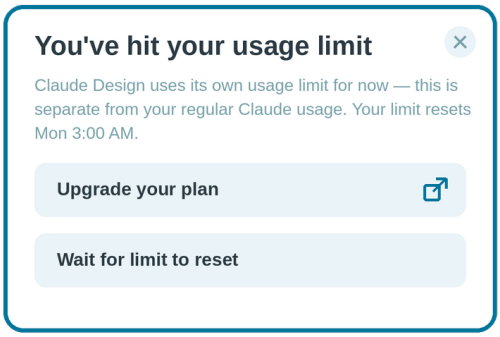
BRIEFING NO. 7

Why did my AI suddenly stop?

Five different limits can produce almost the same annoying message.



Use AI for a while and sooner or later you'll see one of these: **"You've hit your usage limit."** Or **"You've reached the context limit."** Or **"Tool-use limit reached."** If your business runs AI inside other software, add **rate limits** and **monthly spend limits** to the pile. They all have the same effect — the AI stops doing what you asked — but underneath, they describe five very different things. Sort them out once and the messages stop being mysterious.

First: what is a token?

A token is simply the **unit AI uses to measure information.** AI doesn't count words the way we do; it breaks text into smaller pieces. A short word might be one token, a longer word several. Your question uses tokens, the answer uses tokens, and so do instructions, uploaded documents, and previous messages. Think of tokens as the **inches on a ruler** — the unit of measurement, not the limit itself. Nearly everything below is measured in them.

The context window: the size of the desk

Imagine working a complicated project at a desk. You spread out reports, emails, and notes, and as long as it all fits, you can see everything at once and connect it. Eventually the desk is full. That's a **context window**: the amount of information the AI can actively work with at one time — your question, the conversation so far, uploaded documents, and room for the answer it's about to write. Different models have different-sized desks; the big ones today are very large indeed, and the numbers keep changing. A big desk isn't unlimited memory. It's a big working area, and a long conversation slowly fills it — which is why the most useful trick in AI is also the simplest: **changing subjects after a long session? Start a new conversation.** You're clearing the desk. One catch, though: clearing the desk does *not* refill your allowance. That's the next thing.

Usage limits: your allowance

A usage limit answers a different question: **how much AI are you allowed to use in a given period?** Think of it as a fuel tank. A subscription doesn't buy a fixed number of questions — ten quick "rewrite this paragraph" requests cost far less than one enormous research project with long documents, extended reasoning, and web searches. Length, complexity, the model chosen, and the tools involved all draw down the tank at different rates, so two people on the same plan can have very different experiences. Depending on the product, allowances reset by session, day, week, or month. When the message says **"You've hit your usage limit — resets at 3:00 AM,"** the desk may be perfectly fine. You're out of today's gas. Wait for the reset, buy extra credit if offered, or move to a bigger tank.

Spend limits: you still have gas, somebody set a budget

"You've hit your monthly spend limit, but you have more credits." This isn't an AI problem at all — it's a billing control. Credits are still there; a monthly ceiling was set on the account, and the system stopped exactly where the budget told it to. Raising it doesn't make the AI smarter or the desk bigger. It just gives permission to spend more.

Rate limits: the speed limit

Most people in a chat window never meet these; software that *calls* AI does. A rate limit doesn't ask how much you've used this month. It asks **how fast are you using it right now** — requests per minute, tokens per minute. Full tank, money in the wallet, empty desk — and the highway still has a speed limit. You're not out of gas. You're driving too fast.

And "tool-use limit for this turn"?

Modern AI doesn't just write — it may search the web, read files, run code, or call other software while answering one request. That message caps *this round* of that activity, not the conversation or your account. Which is why it comes with a Continue button: finish the round, start another, carry on.

The five-second version

When AI says it hit a limit, ask **which limit? Tokens** are the units being counted. **Context window** is the size of the desk. **Usage limit** is the fuel tank. **Spend limit** is the budget. **Rate limit** is the speed limit. Knowing which one usually tells you what to do — clear the desk, wait for the refill, raise the budget, or slow down. And none of them means anything is broken. They're just the guardrails that make an enormous shared service work.

With SterlingPRIVATE.ai, there are no token limitations — you own all of them.

> Want to learn more about keeping your data private? Email **PrivateAI@sterling.net**