

My Codex Ran 800 Million Tokens in A Day. The Real Story Isn't Cost.

21 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=L8BL0TSLK6M](https://www.youtube.com/watch?v=L8BL0TSLK6M)

<https://www.youtube.com/watch?v=L8BL0TSLK6M>

SUMMARY

The video discusses the creation of a token burn dashboard, emphasizing its purpose beyond merely tracking token usage. The creator encourages viewers to rethink their computing experiences with AI, using the dashboard as a tool to analyze habits, improve efficiency, and expand imagination in AI applications.

- The dashboard tracks AI token usage to help users understand their habits and improve their AI interactions.*
- Codex is highlighted as a user-friendly tool for measuring token usage, unlike other AI models.*
- Visualizing token burn data reveals insights into how different AI tools impact productivity and creativity.*
- The creator shares examples of using AI for personal tasks, such as organizing files and conducting research efficiently.*
- The importance of community sharing and learning from each other's AI usage is emphasized.*
- The creator encourages viewers to build their own dashboards and explore creative AI applications to enhance their productivity.*
- The video stresses that understanding token usage can lead to better problem-solving and more effective use of AI resources.*

I built a token burn dashboard. In this video, I'm going to show you how I did it, but the point is not to burn tokens. The point is not to brag about how many tokens you burned. Yes, I burned, I think it's 800 million tokens last Thursday. That's fine. The point is what I did with it. And the heart of this video is to share with you that you need to imagine your entire computing experience differently, and that one of the greatest tools to do that is to see how you use your computer with AI. And that's what this token dashboard does.

It actually shows you, what are my habits with AI? How do I use it? Do I use it well? Could I be using it more? I am really really into tools that expand the imagination. I'm not into tools that just show you what you did and say, "So far, so good." And so, I want to walk you through how I used AI to make this dashboard, and how I'm using it every day to stretch my own imagination and expand what is possible. So, first, how do you think about a dashboard for AI?

What does that mean, and what does it involve? I got to be honest with you, the easiest tool that you can use for this is Codex. And part of why is because Codex makes it easy to measure the tokens Codex uses. You know how we always used to say, "Well, AI doesn't know what version it is." It's kind of the same thing with tokens now. If you're using Claude, which I love, you can't tell how many tokens Claude is using unless you're in the API. You cannot easily tell if you are in Claude, co-work if you're in Claude chat, how many tokens you have used in your session, even if you've done very heavy work. But in Codex, you know it down to the token. And

so, the first thing to be honest with is I had to do some fancy math to approximate my Claude usage as a part of this dashboard work. And yes, we'll we'll talk about that. And why do we care? Why do we care about approximating usage, getting the usage right? Well, first and foremost, if you don't know what the ratio of your actual behavioral usages of tools, it's hard for you to learn and self-improve over time in how you use AI. It is really tough. One of the things I noticed that just opened my eyes when I built my dashboard is that I could see in the chart the way my behavior actually shifted when I started using Codex. I could see the number of tokens going up.

It showed me that that particular tool was unlocking my imagination in a way that previous tools had not. I don't think I would have had that insight in the same way. Certainly not at that same level if I hadn't had a chart. Now, if you're wondering, "What does this chart look like, Nate? How did you get this chart to look this good?" I've got a simple answer for you. I found an open-source skill, which I'll be sharing, called the Tufte skill. Tufte was a famous data visualizer, and all I did to make this chart was I actually worked with Codex, and I told Codex what I wanted in terms of features very clearly in plain English. It wasn't a fancy prompt. I said I wanted to see my token burn. I wanted to see it in a github.style chart. I wanted to see more particularly what I was doing on given days. I wanted to see same-day usage so I could understand in a lived-experience kind of way what activities were burning tokens versus what weren't. Uh and I wanted to understand the model distribution. And so, I had to do work to actually reason from artifacts, to reason from logs, to approximate model usage for Claude so that I could actually map that in and understand the relationship uh in my activities. Your particular chart may look different. One of the things that's cool is that I actually put together multiple versions of this for you over on the Substack, so you can have one that is more Claude-leaning, you can have one that is more ChatGPT-leaning, you can have one with multiple lines that show your Claude uh usage versus your ChatGPT usage versus your Codex usage. All of that's great. But, remember, the point is not to show usage, the point is to show where you go next. And so I crafted mine specifically to help me understand what am I doing today on AI and how is what I'm doing today on AI different from previous days when I was a heavy AI user. So I can see that same day.

Now, let me give you a specific example because I think that that's that that's going to be useful. One of the things that dropped this week was slash workflows, which is not from Codex. It's actually from the Opus 4.8 release day. And slightly compose a workflow with sub agents using Claude code. Yes, Claude code. And when you do that, what Claude code does is it dynamically spins up a plan, an orchestration plan to get your work done and then spins up sub agents and actually gets it done.

Which sounds great. It makes multi-agent orchestration for personal productivity really easy. Well, of course the internet jumped on that. They immediately created an open source skill. I grabbed that open source skill. I ported it over to Codex and I was able to the same day use slash workflows inside Codex to help me tackle tough tasks. I had a task where I had to research a bunch of different schools with my kids and I was able to use slash workflows to put together a gigantic report around schools and like what school would be best without any significant additional effort. And it used three or four agents and it was much more uh useful as a result because you're using more agents, right? So more agents means more tokens burned, which is again gets back to the token chart. But it also means a higher probability that you solve the problem correctly because you're tackling it from multiple different angles. And so if you're wondering how this all relates to the chart, it's very simple. As

soon as you look at the chart, you can see that my behavior shifted the token usage, I can see intuitively that my behavior led to a higher quality result, which is interesting.

But, that was because I could not invoke `{slash}` workflows in the same way from the chat and from code work as I could in Claude code, and it was difficult to get Opus 4.8 to go to the level of depth even when I did invoke the `{slash}` workflows command. And so, effectively, looking at the chart enabled me to see how a change in my behavior was changing token burn and resulting in new work. Because here's the secret, guys. You are not going to understand how to use AI unless you have a feedback loop that allows you to take the AI work you're doing, see how it's affecting your token burn, and then see how it results in higher quality work for you. And the reason token burn in particular matters is because it's easy to measure, and it's correlated with intelligence and successful solutions.

So, I'm not measuring tokens just for the heck of it. I'm measuring tokens because I'm measuring my ability to deploy delegated intelligence to solve problems. And we have seen in study after study after study done by the major labs that when you spend more tokens, you get better results. It's one of the most predictable results in AI. And so, if I'm interested in understanding whether I'm stretching my imagination on AI, I have to kind of like look at my token usage and find out. And I have to see, does it work?

Does it not work? Am I actually spending more delegated intelligence here or not? Now, I'm not saying you want to spend more tokens just to be wasteful. I actually have skills in place that help me to pause automations I don't need, that help me to slim down context windows where I don't need to use them. And so, it's not that we want to waste tokens, we want to use them effectively, but we have to know what we're measuring to get anywhere at all.

And so, if you want to dig in, here's the secret. I told Codex, use this Tufte scale that I talked about, right? Use this design scale. Please, please, please, show me my top 10 days for AI. What are the top 10 AI usage days I have? What's in those top 10 days? What activities did I do? And by the way, I had to scrub some of this in the chart I'm showing you because some of it is quite confidential. Originally, it was extremely specific, and I recommend yours be extremely specific as well if it's a private dashboard because it helps you to learn what you're actually using AI for. And then, I told Codex to please log how my token usage is changing over time. And you can do it any way you want. You can use an XY axis. I used a logarithmic axis, which is a fancy way of saying I wanted to see big changes over time because I was seeing this huge scale problem where I would have some days that were earlier in the AI revolution where I would have only a few million tokens and and days now where it's almost a billion. How do you show that in a scale without breaking it?

And so, I worked with it on the scale. I worked with it on the color contrast. I worked with it on how clean and easy it is to read. I worked with it on showing multiple models. And I actually used Codex to build an approximation of my usage. Codex gave me a quiz when I asked for it. And and it said, "Help me understand your Claude usage." And then it reasoned from artifacts and came up with a tight range for the tokens that Claude used. Yes, it is ironic that Codex is helping me infer my Claude usage. Anthropic, you got to fix this. Please make it easier to measure our tokens on Claude. I would love that so much. But once you get to that

point, once you actually understand what you want to do, there is not a special prompt that you have to get give Codex to make this happen. That this is a wonderful example of 2026 building. It is about the clarity of your intent. I had to see in my head that I wanted the GitHub chart at the top, that I wanted a logarithmic scale that showed the gain in tokens over time. I had to see how I wanted the top 10 days, and then I just had to keep asking for it until it appeared. Yes, I also could have specified it exactly in a really fancy set of requirements up front, but this is a home build, and I'm a little bit of a lazy builder sometimes, and I just had that picture in my head, and I kept asking for it till I got it, and it worked perfectly fine. And And that should encourage you, right? Building is not that hard, and if you're wondering, "Did I build it all the way?" Yes, Codex absolutely deployed it. So, it was able to handle the DNS change to change it to the domain. Uh it's tokenburn.markdown, and it took care of it, right? Like, it took care of the entire thing for me.

And that is an example of using your imagination for what AI can do. I realize that's circular. Another example that's been really key for me is using AI to keep my files better organized. My files are not mine really anymore. I use them as fodder for AI, and I let AI organize them. I actually had Codex go through and label and organize all of my annoying little screenshots that I take. When I take screenshots on X, when I take screenshots other places, I had to have them all organized. I had to have them all labeled. It broke open, it looked at all of them, it labeled them, it put them into files, and now it's easy for it to find them and work with them. I have no idea what the folder structure is. I don't have to care. That's an example of the kind of computing that you can do once you stretch your imagination. And so, really, my invitation to you is not to build a token dashboard. I know that people bragging about their AI usage. I know that 800 million tokens in a day is not the highest total out there. I don't care. I'm using the AI the way I can, with delegated intelligence, to solve problems that matter to me. And it's It's a loop that's helping me refresh my thinking, because I can look through my top 10 days, and I can say, "Oh, you know, on my higher AI days, I'm actually giving Codex strong database work to do.

I should be doing that more because that seems to be working better. Or wow, in the last hour I burned 100 million tokens and I could feel all eight threads that I was running come to a successful conclusion. I should be parallel computing more. These are insights that you don't get without keeping a little bit of a closer run on intelligence. Like if you had intelligence, wouldn't you want to actually meter it to understand how it works? That's what we're talking about here. We're talking about building a compass and a speedometer for intelligence.

And if you don't have that, it's hard to know where you're going. And so the visuals you're seeing in this video are actual visuals of the chart that I made for me. And if you want a complete version of it, it's on Substack. You don't have to copy mine. I'm actually giving you multiple versions. Uh I love diving deep on Substack. And you're going to get multiple versions of how to build this. You can build it for yourself. You can share it. We're going to make it easy to share on Talent Board uh because I believe that in the future this is going to be as important as GitHub. You're going to have to be able to show people who are prospective employers that yes, you do actually burn tokens on AI. And if they walk in and they say, "Wow, you know, you've been burning 3 million tokens a day." I don't know. That's going to become a factor.

Like if you have two people and both of them show their charts and one's chart is like a hundredth of the other, yeah, people are going to wonder about your AI usage. And so I also think there's a degree of public

accountability that helps us to stretch and think and grow and and think more about how we can use AI. We are so early. 0.6% of ChatGPT users are using Codex right now. 0.6, not even 1%. And I don't say that because Codex is the only way to do this. I built this in Codex. Opus 4.8 is going to be great for this. If you're a Claude user, you can absolutely build it in Opus 4.8 as well. It's going to be a little bit harder to measure your usage unless you are using the API a fair bit and then it's very, very easy.

But, you can still build a model and infer it, and I provide instructions for that as well, and that's certainly much better than nothing. And again, Anthropic, if you're listening, please, please, please meter those Claude tokens. It makes everything easier. It's just a piece of transparency we need. But, there's a community aspect to this that is really, really important. So, I'm not just talking about doing this for you. I am challenging you to start doing this and sharing your work because when we share what we've done, when we're able to talk about how we're using tokens, the effect of being in a community and sharing the kinds of uses we've had for intelligence is incredibly powerful.

That's one of the things I learned actually talking with Emma at OpenAI recently. She talked about this culture they have that's called the "You're Cool" culture, right? Where like if you come up with a new usage for AI that people haven't thought of, you're the cool person for the day, right? Because people want to learn from that. We need more of that not just at the hyperscalers. We all would benefit from that. I would benefit from learning how you're using AI, and I do. In fact, the Substack chat, the executive chat that I have for folks on WhatsApp, those are all places where I learn what my folks who are subscribing actually use AI for, and I learn from them, and I hope they learn from me a little bit. It's super fun. We need to learn from each other.

We need to learn from each other. Uh and if we don't learn from each other, we're going to think [snorts] only in terms of our little world and what we can accomplish, and we're going to miss the emergent possibilities of these models. And that's something that I try and emphasize over and over again. Part of why things like this chart are super important is because models are grown, not made. They are not defined as a series of parameters that everybody understands. These models are actually at 10 trillion weights or 10 trillion parameters right now. And nobody knows all of them, and they were evolved and grown over the course of a reinforcement learning training regime, they were not designed by a scientist who understood every aspect of them.

And when people portray them as traditional software, it's deceptive and it's wrong. And by the way, I see journalists doing this all the time. You guys got to stop it. AI is a fundamentally different kind of architecture than we've ever had in computing before. It's called transformer architecture. Yes, I have whole videos on them. You can find them. I'm not going to do them here. We need to realize that when you have a model that is grown not made, you do not fully understand what it is capable of even if you were the one who made it, right?

Even if OpenAI released it, they don't know all the capabilities. We have to discover it. And so tools like this are essentially a way for us to say, "Hey, one, we're using this intelligence, which is a good thing if you're trying to grow your career in 2026. Two, this is what we're using it for so we can all learn from each other. Three, this is an idea of where to go next based on what we're learning and feeling." Because the other part of this, right? Like that's not on the chart, is how it feels to you to use AI when you have a 100 million burn hour, or how it feels to

you to use AI when you get a big project accomplished and you can see how that hit the token burn and that gives you an idea for what works and how delegated intelligence works that you can't substitute for.

Now, you're going to wonder, how am I using AI? What are the secrets of Nate? Where is Nate using AI? Well, I've already shown you a few pieces, right? I talked about the school the school choice, that's obviously a personal thing, a family thing. I'm also using Codex very, very heavily to optimize my computing right now. And so I'm using it to check email. I'm using it to check Slack. I'm using it to And by the way, people say, "Well, why are you Why are you doing that? That's That's overkill.

You don't need to do that." I got news for you. Everybody's time would be better spent not in email. Everybody's time would be better spent not in Slack. If I can have a tool that helps me handle that it helps me focus better, that is actually a very valuable use of intelligence. I also use it, as I've been saying, for handling files on my computer. So, I have a whole process where I have work packs that I have set up that are focused on particular projects. I have a chief of staff that I use that's a thread inside Codex that spins up sub agents. And that all burns tokens because you're now you're managing multiple agents. And that works for me because then I can have one thread that has all the context of what I'm working on and I keep context window clean for detail work with lots of child threads. These are all examples of how I'm actually using those tokens and why I think that they make sense for me. I'm not saying you have to do exactly what I do, but I would invite you to challenge and open yourself up to using AI more than you might think you can. Because these models are almost never at the frontier, at capacity. I don't care that I burned 800 million tokens or whatever in a day last week.

I am not nearly at capacity on AI. There is so much more that I can be doing and I and when I think that way, I come up with creative uses for AI that are really, really important. I The way I've started to handle automations has been very helpful for me for feeling like I'm up to speed. That's another great example. Internal dashboards automatically with running automations. That's another great example of how you can use delegated intelligence. I do that as well. The sky is the limit on what you can use with AI, but we need desperately some degree of accountability and some degree of a learning loop to help us to learn this.

Effectively, the difference between someone who's using a couple of million tokens a day, light user of AI, and someone who's using almost a billion tokens a day. It is literally in tokens a 99% difference. It is also in terms of fluency in what you can get done, at least a 99% difference, but it feels like more because of the multiplicity of the impact you can have with large multi-agent runs. So, the reason I'm emphasizing this now is that we are living in two or three different futures at once. There are a few people who are doing the 1 billion token a day lifestyle where they're actually using these agents to their full capacity. As I see this in the comments like ah AI does what it does, there's been no improvement in 6 months.

No, there has been an improvement in 6 months, but your imagination hasn't caught up with it. And so you're not realizing it can go through and organize all your screenshots and sort out your entire files, sort out your downloads folder, take care of cleaning up your hard drive, take care of troubleshooting your internet connection. All of those are real things I've done with AI by the way in the last week. It can do all of that.

But you got to ask it. You got to ask it. And you got to have examples around you that show that. And so that's why I'm encouraging folks, please please please build and show your token burn chart. Show what you've been using AI for so we can all learn from you. I'd like to learn from you. You know, I'm sure there's a lot I can learn that I haven't thought of. I want you to be the cool kid. You tell me how you've been using AI. You tell me something creative you've done with AI. Is there something with books? I'd love to know that. You know, I love to read. And let's all grow our ability to use these tokens intelligently together. That's why I created the token burn chart. That's why I think it's important. That's why I made several versions of it over on the Substack so you can make it too.

Uh and I think it's really important. I I think it's something that that that if we find ways to share what we are doing with intelligence, we're going to get smarter with AI. And that matters a lot. All right. I will see you next time. This has been so much fun. I had fun making this dashboard. I hope you can tell. You should have fun making it too. It really isn't a pain to prompt and make. I think I made this one in about uh an hour. I went back and forth. I was a very lazy prompter. Uh and it still took an hour and it was just fantastic and fun all along the way. So, have fun with it and uh I'll see you next time. Cheers.