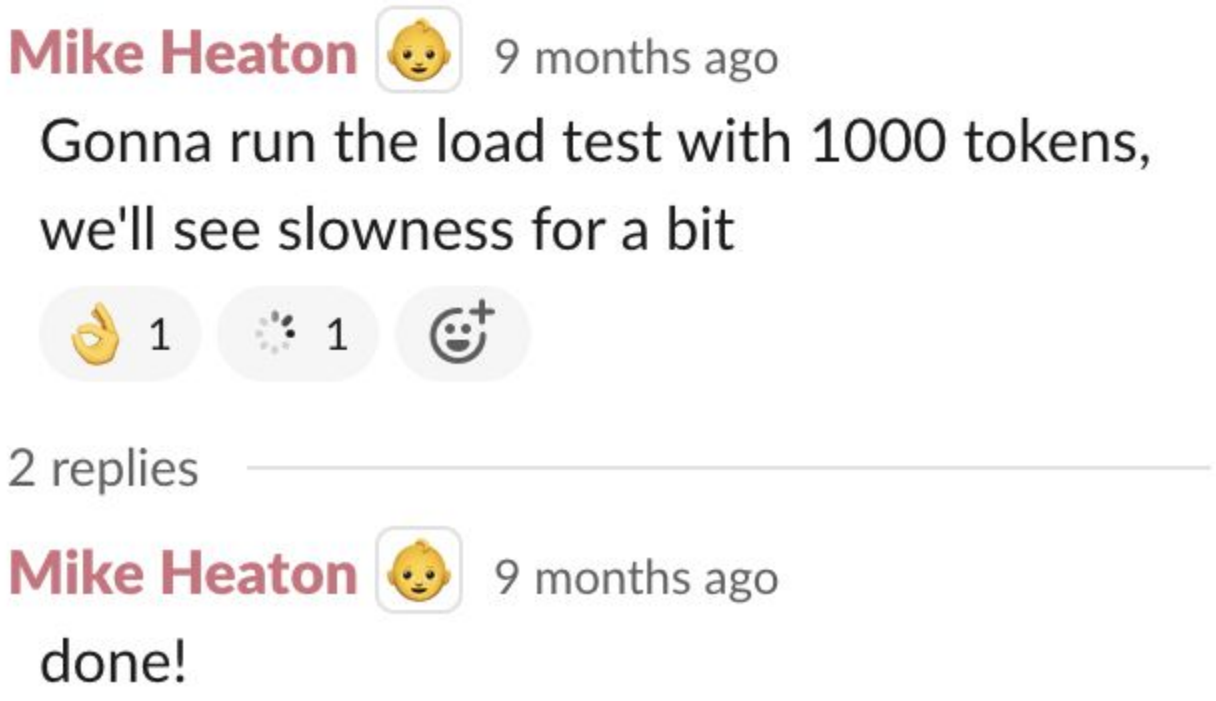
OpenAI

Behind the Scenes Scaling ChatGPT

Evan Morikawa
@e0m

Evan Morikawa is presenting
10:07 AM



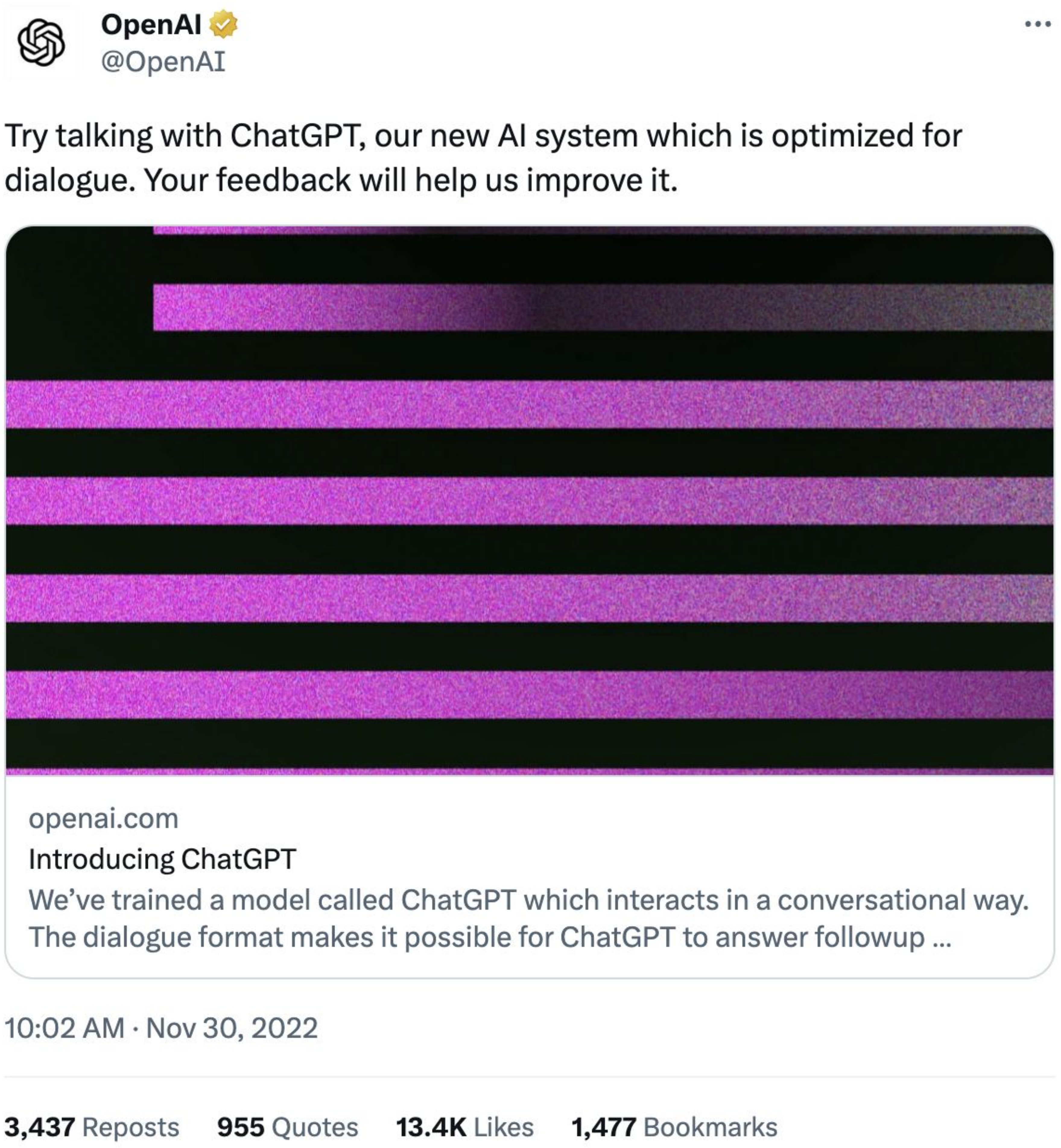
Mike Heaton 👶 9 months ago

Gonna run the load test with 1000 tokens, we'll see slowness for a bit

2 replies

Mike Heaton 👶 9 months ago

done!



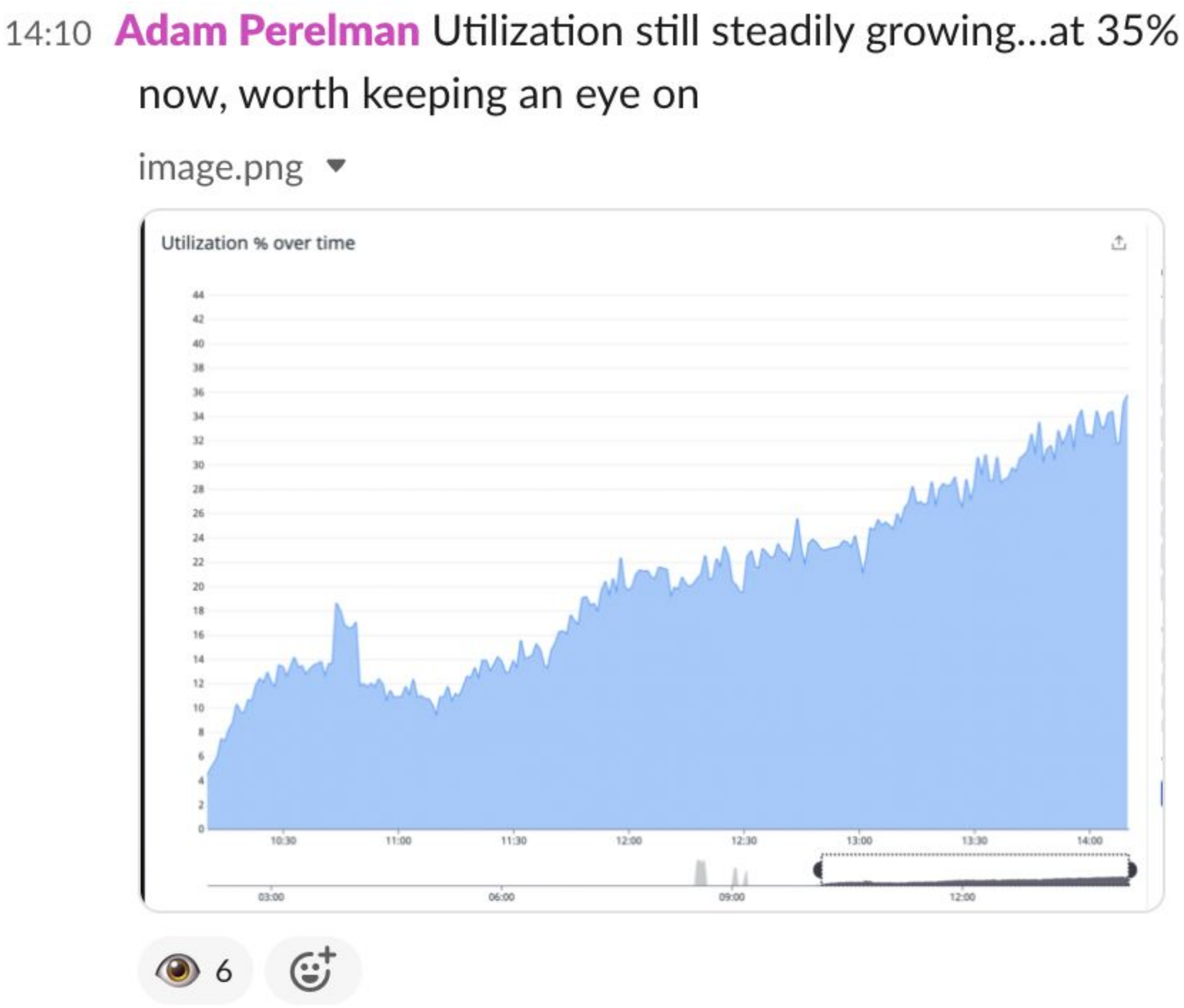
OpenAI
@OpenAI

Try talking with ChatGPT, our new AI system which is optimized for dialogue. Your feedback will help us improve it.

openai.com
Introducing ChatGPT
We’ve trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup ...

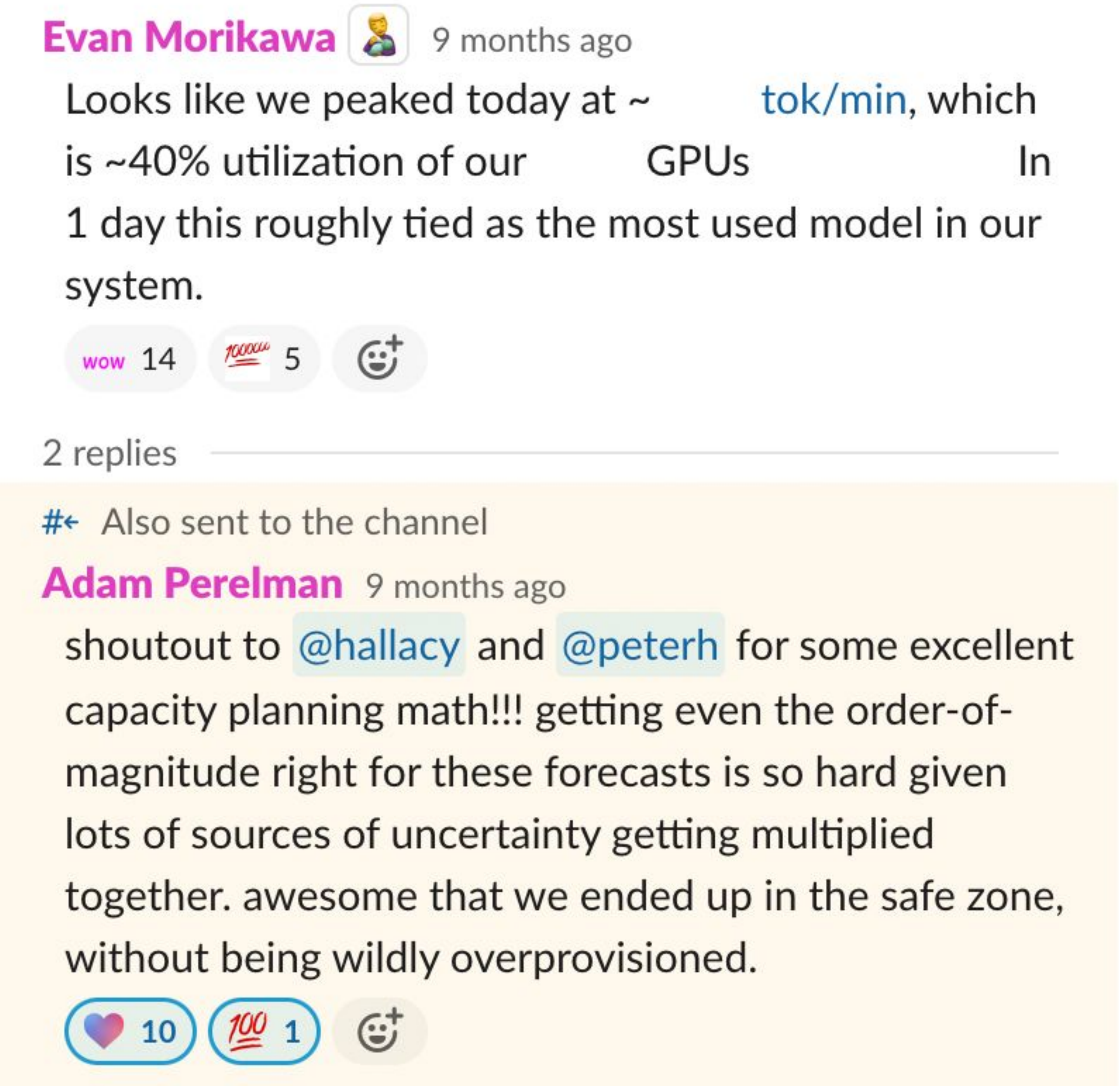
10:02 AM · Nov 30, 2022

3,437 Reposts **955** Quotes **13.4K** Likes **1,477** Bookmarks



14:10 **Adam Perelman** Utilization still steadily growing…at 35% now, worth keeping an eye on

image.png



Evan Morikawa 9 months ago

Looks like we peaked today at ~ tok/min, which is ~40% utilization of our GPUs In 1 day this roughly tied as the most used model in our system.

wow 14 💯 5

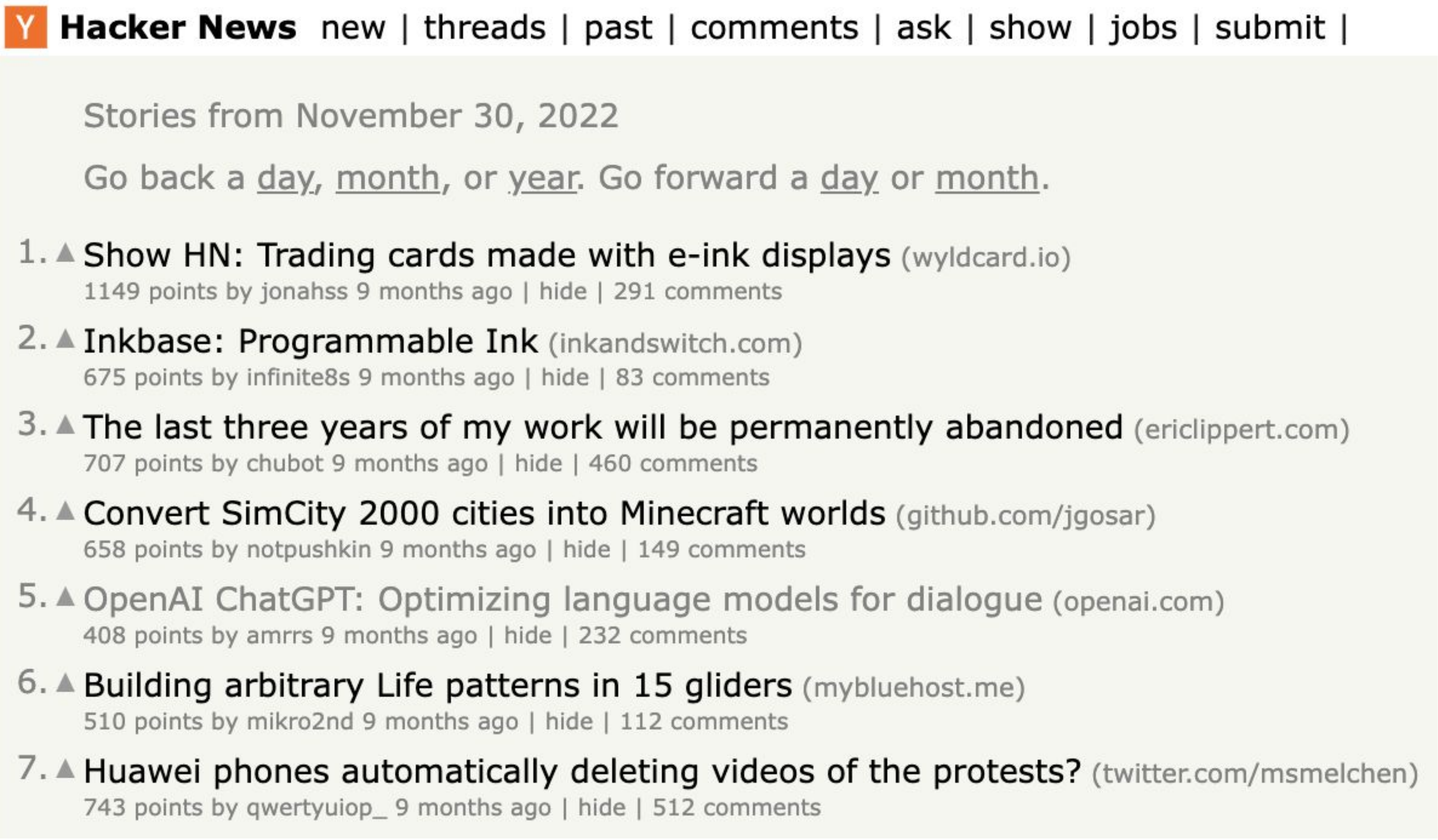
2 replies

#← Also sent to the channel

Adam Perelman 9 months ago

shoutout to @hallacy and @peterh for some excellent capacity planning math!!! getting even the order-of-magnitude right for these forecasts is so hard given lots of sources of uncertainty getting multiplied together. awesome that we ended up in the safe zone, without being wildly overprovisioned.

❤ 10 💯 1



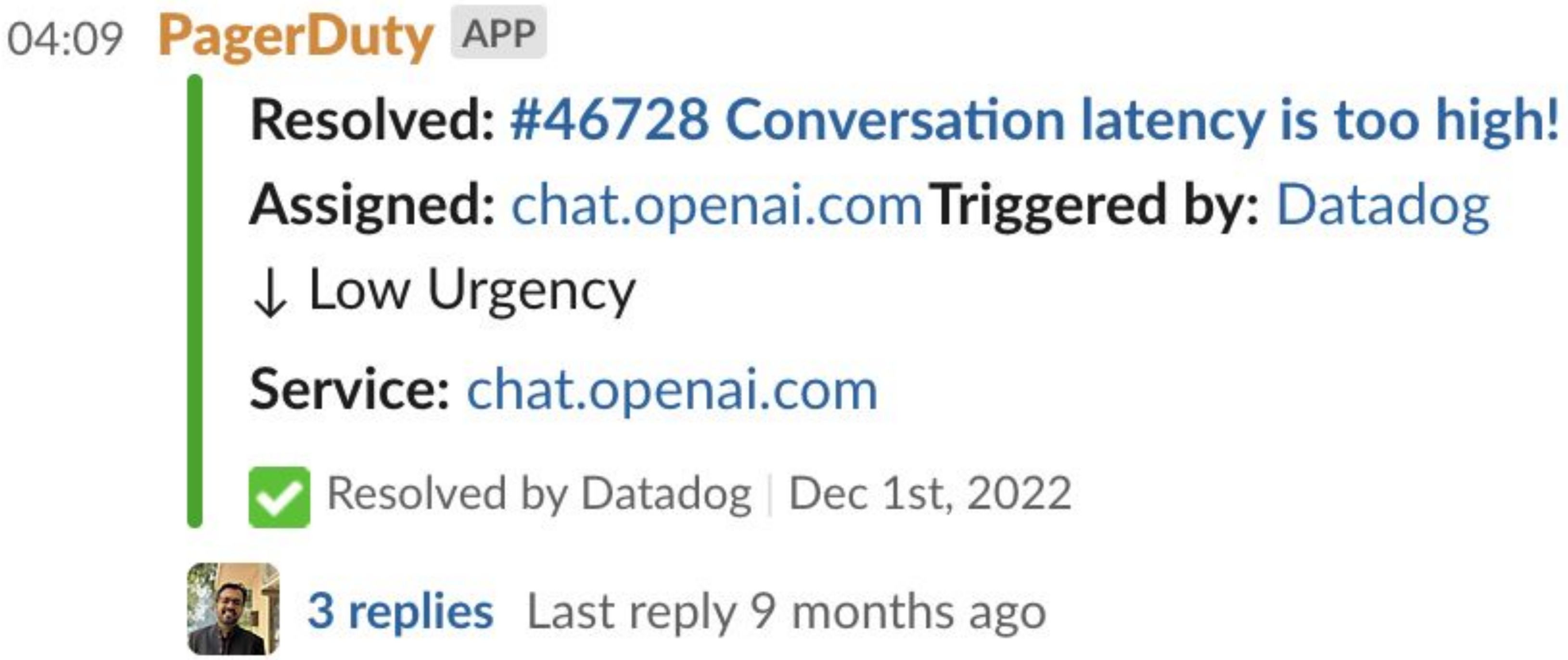
Hacker News new | threads | past | comments | ask | show | jobs | submit |

Stories from November 30, 2022

Go back a day, month, or year. Go forward a day or month.

1. ▲ Show HN: Trading cards made with e-ink displays (wyldcard.io)
   1149 points by jonahss 9 months ago | hide | 291 comments
2. ▲ Inkbase: Programmable Ink (inkandswitch.com)
   675 points by infinite8s 9 months ago | hide | 83 comments
3. ▲ The last three years of my work will be permanently abandoned (ericlippert.com)
   707 points by chubot 9 months ago | hide | 460 comments
4. ▲ Convert SimCity 2000 cities into Minecraft worlds (github.com/jgosar)
   658 points by notpushkin 9 months ago | hide | 149 comments
5. ▲ OpenAI ChatGPT: Optimizing language models for dialogue (openai.com)
   408 points by amrrs 9 months ago | hide | 232 comments
6. ▲ Building arbitrary Life patterns in 15 gliders (mybluehost.me)
   510 points by mikro2nd 9 months ago | hide | 112 comments
7. ▲ Huawei phones automatically deleting videos of the protests? (twitter.com/msmelchen)
   743 points by qwertyuiop_ 9 months ago | hide | 512 comments

LOW KEY
RESEARCH
PREVIEW



04:09 **PagerDuty** APP

Resolved: **#46728 Conversation latency is too high!**

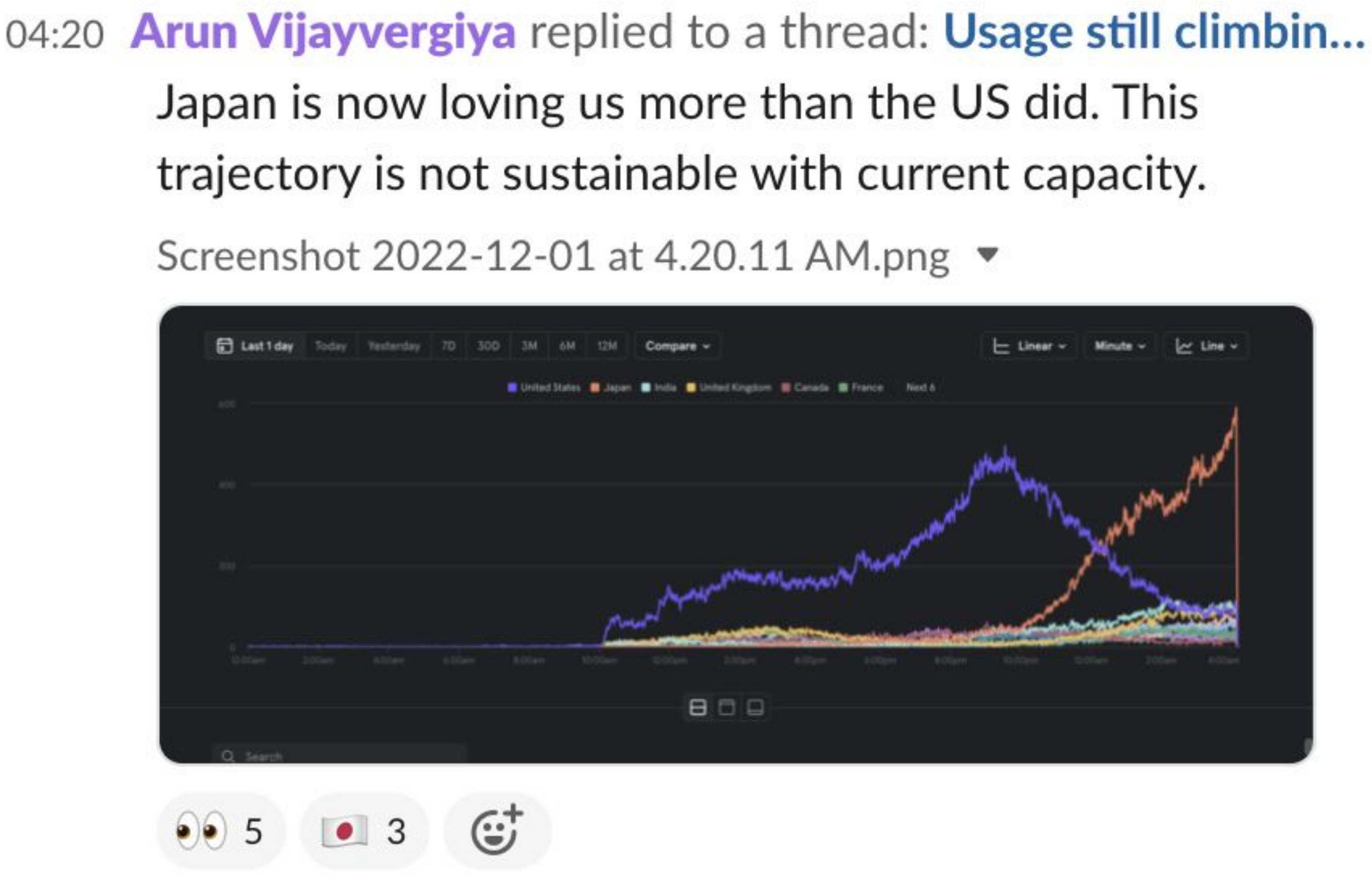
Assigned: chat.openai.com **Triggered by:** Datadog

↓ Low Urgency

Service: chat.openai.com

✅ Resolved by Datadog | Dec 1st, 2022

3 replies Last reply 9 months ago

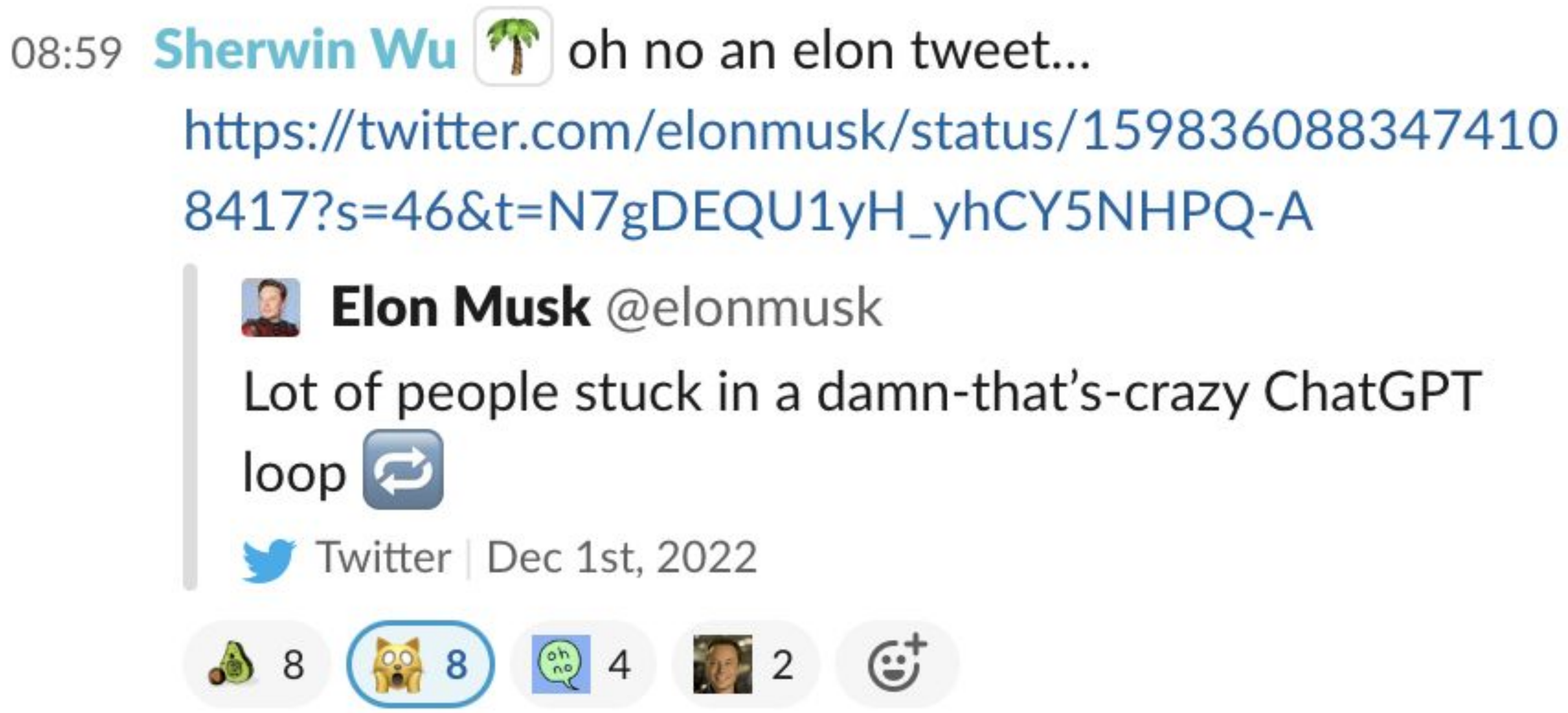


04:20 **Arun Vijayvergiya** replied to a thread: **Usage still climbin...**

Japan is now loving us more than the US did. This trajectory is not sustainable with current capacity.

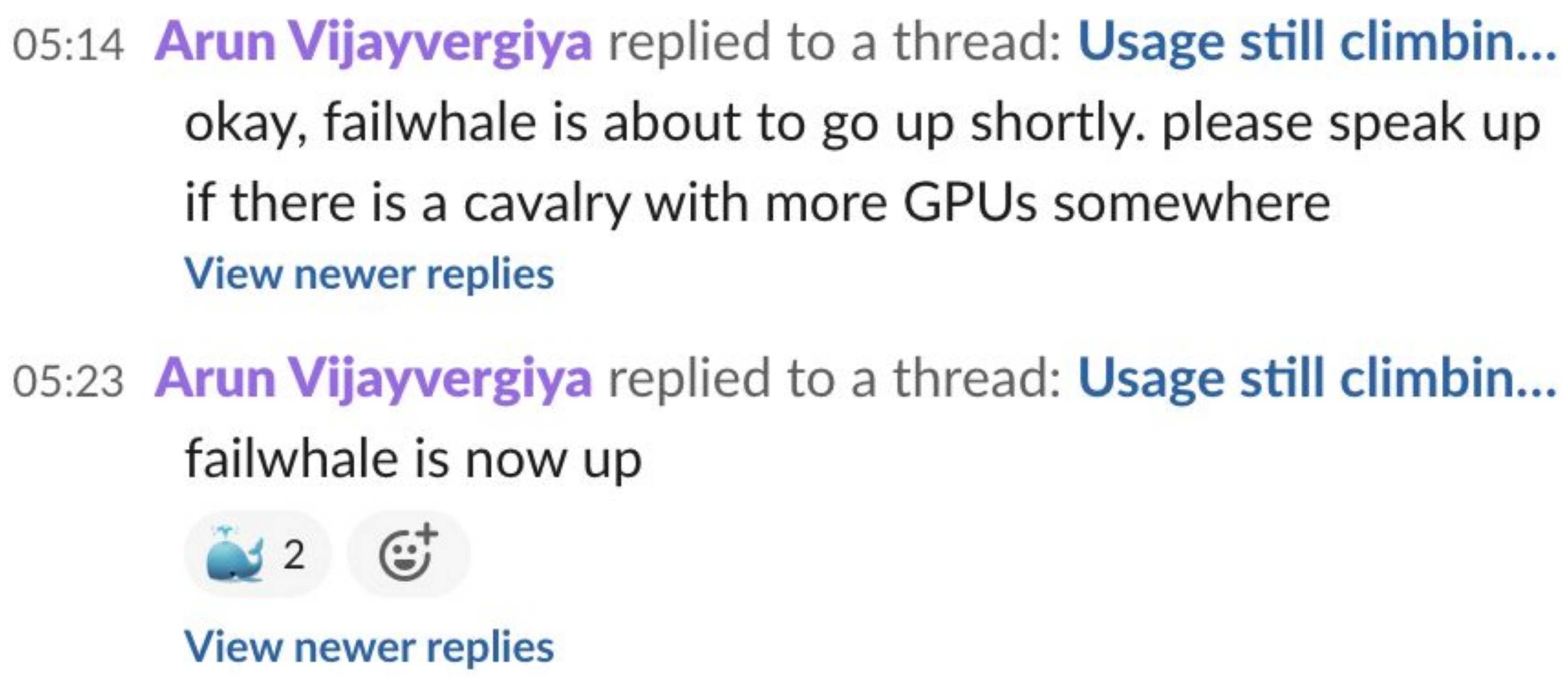
Screenshot 2022-12-01 at 4.20.11 AM.png ▾

👀 5 🇯🇵 3



08:59 **Sherwin Wu** 🌴 oh no an elon tweet...
https://twitter.com/elonmusk/status/1598360883474108417?s=46&t=N7gDEQU1yH_yhCY5NHPQ-A

> **Elon Musk** @elonmusk
>
> Lot of people stuck in a damn-that's-crazy ChatGPT loop 🔁
>
> Twitter | Dec 1st, 2022

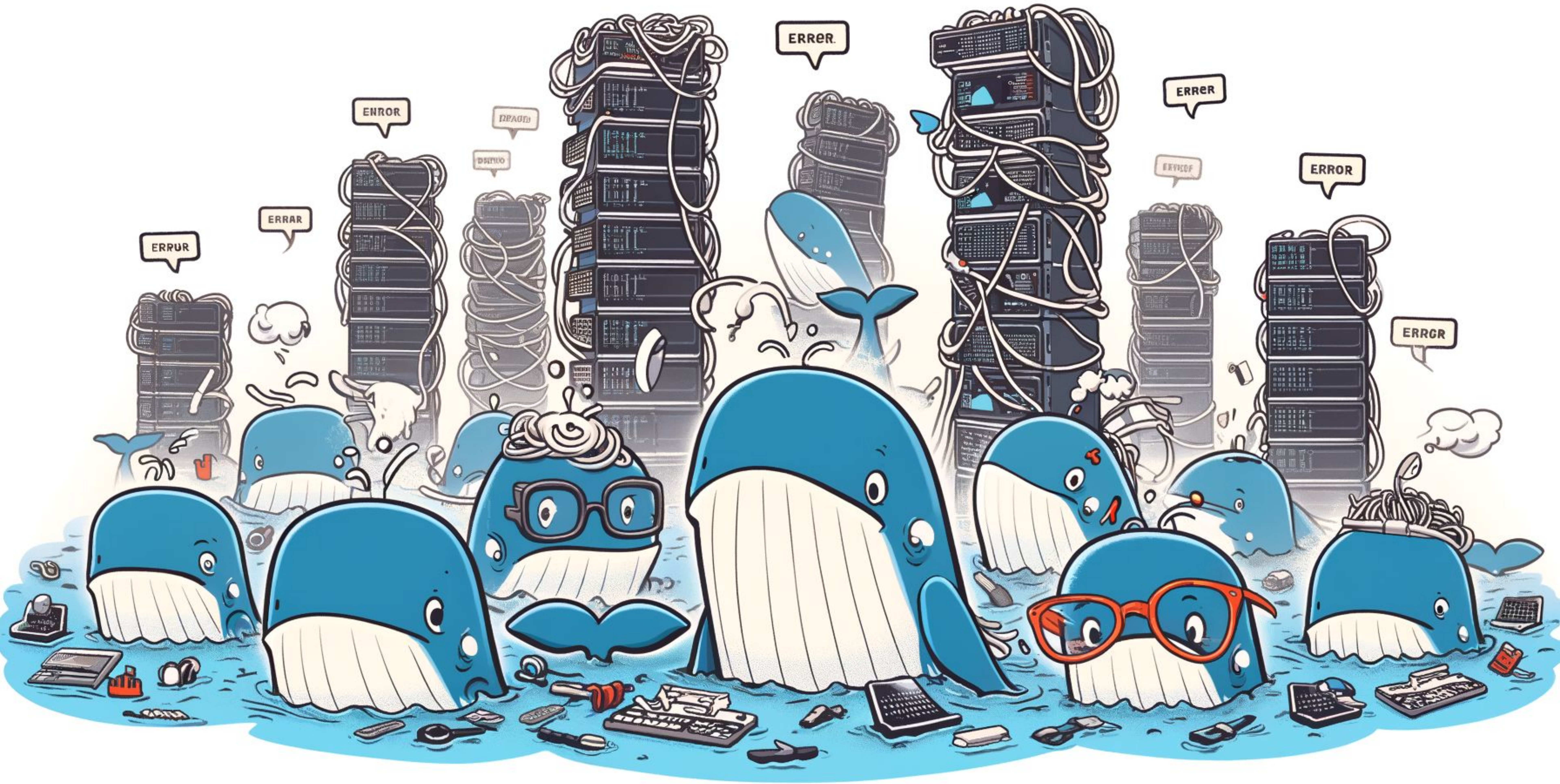


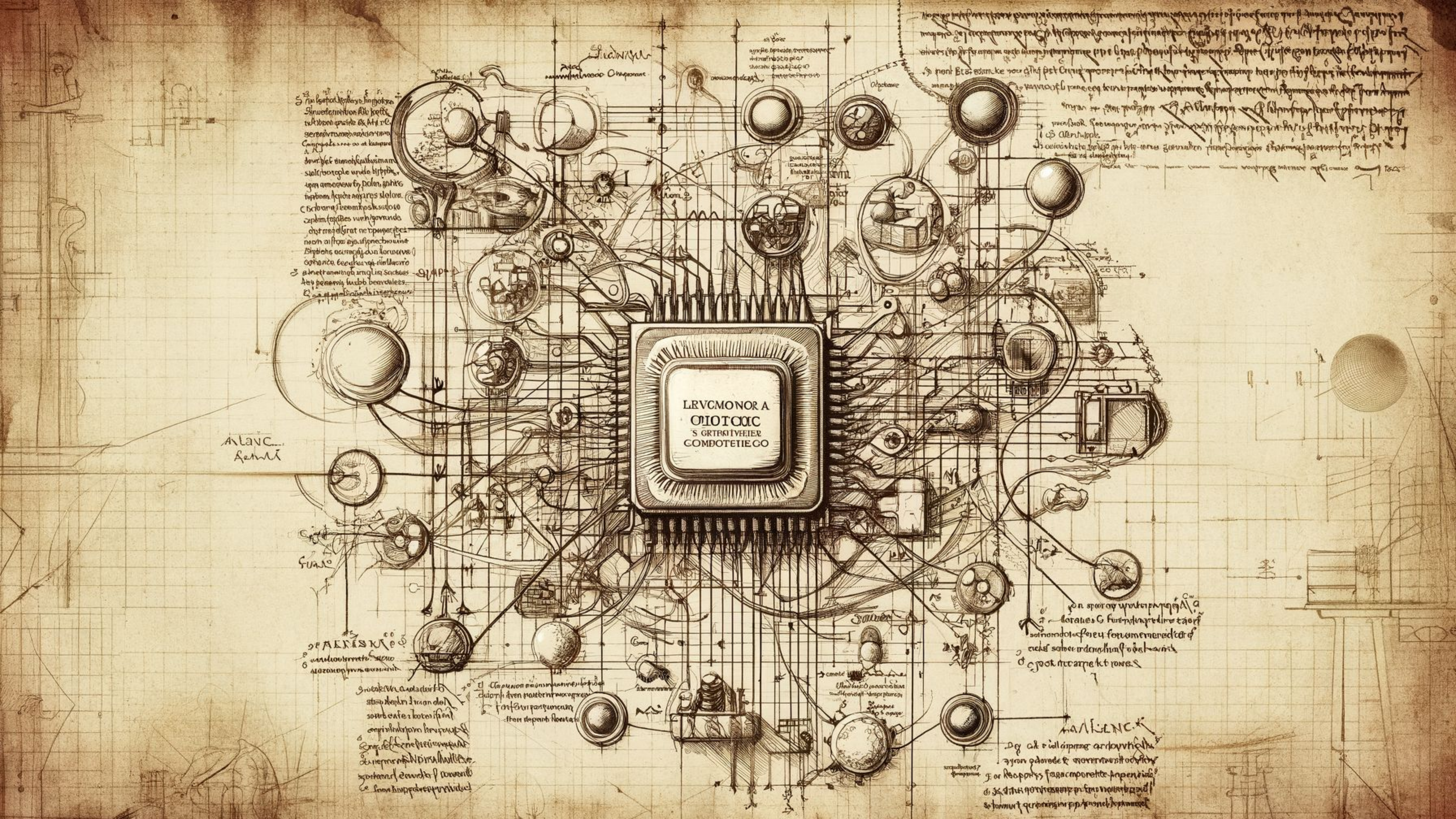
05:14 **Arun Vijayvergiya** replied to a thread: **Usage still climbin…**

okay, failwhale is about to go up shortly. please speak up if there is a cavalry with more GPUs somewhere

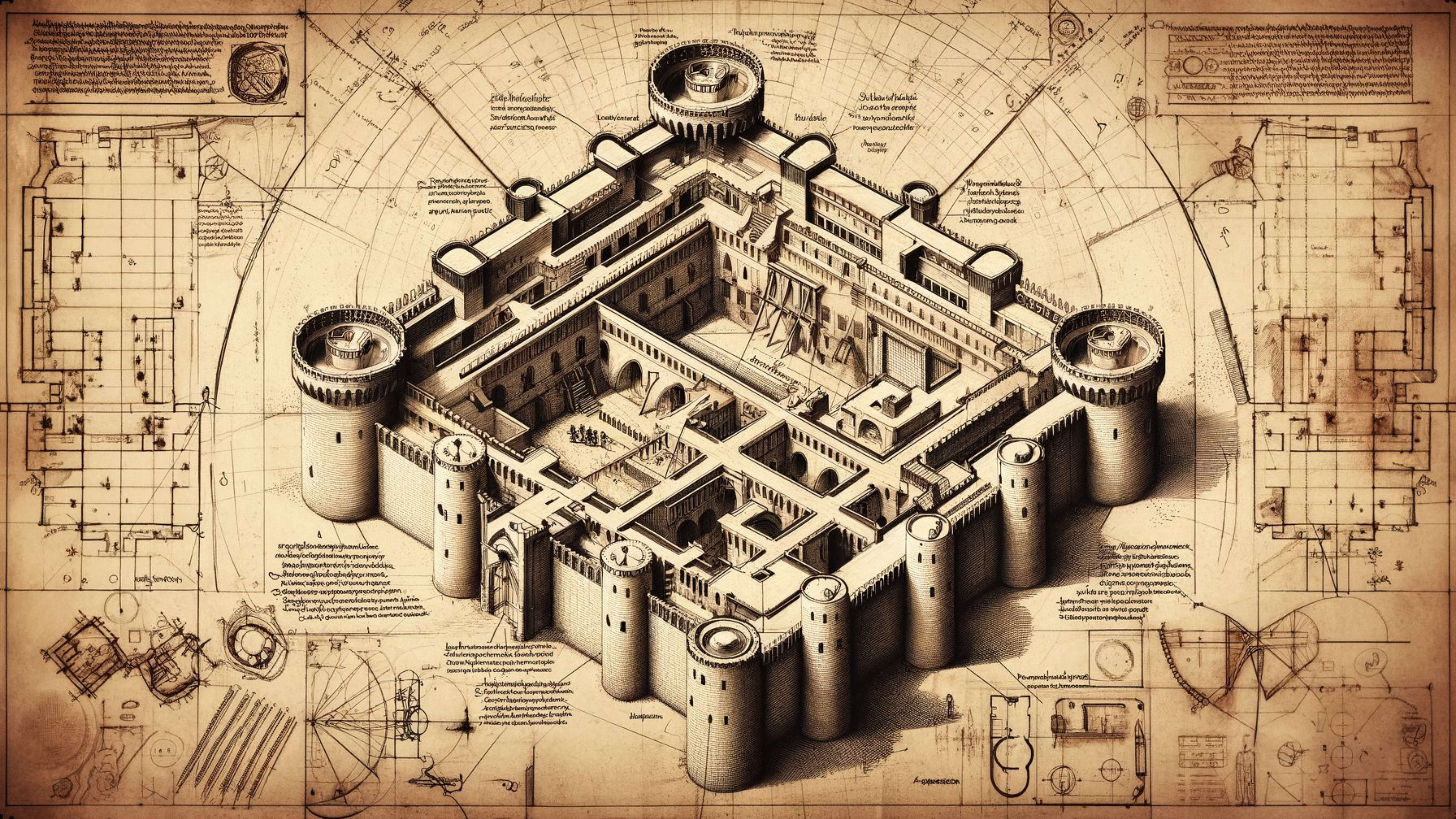
05:23 **Arun Vijayvergiya** replied to a thread: **Usage still climbin…**

failwhale is now up


ERRUR
ERRAR
ENROR
ERRER.
ERRER
ERROR
ERRGR



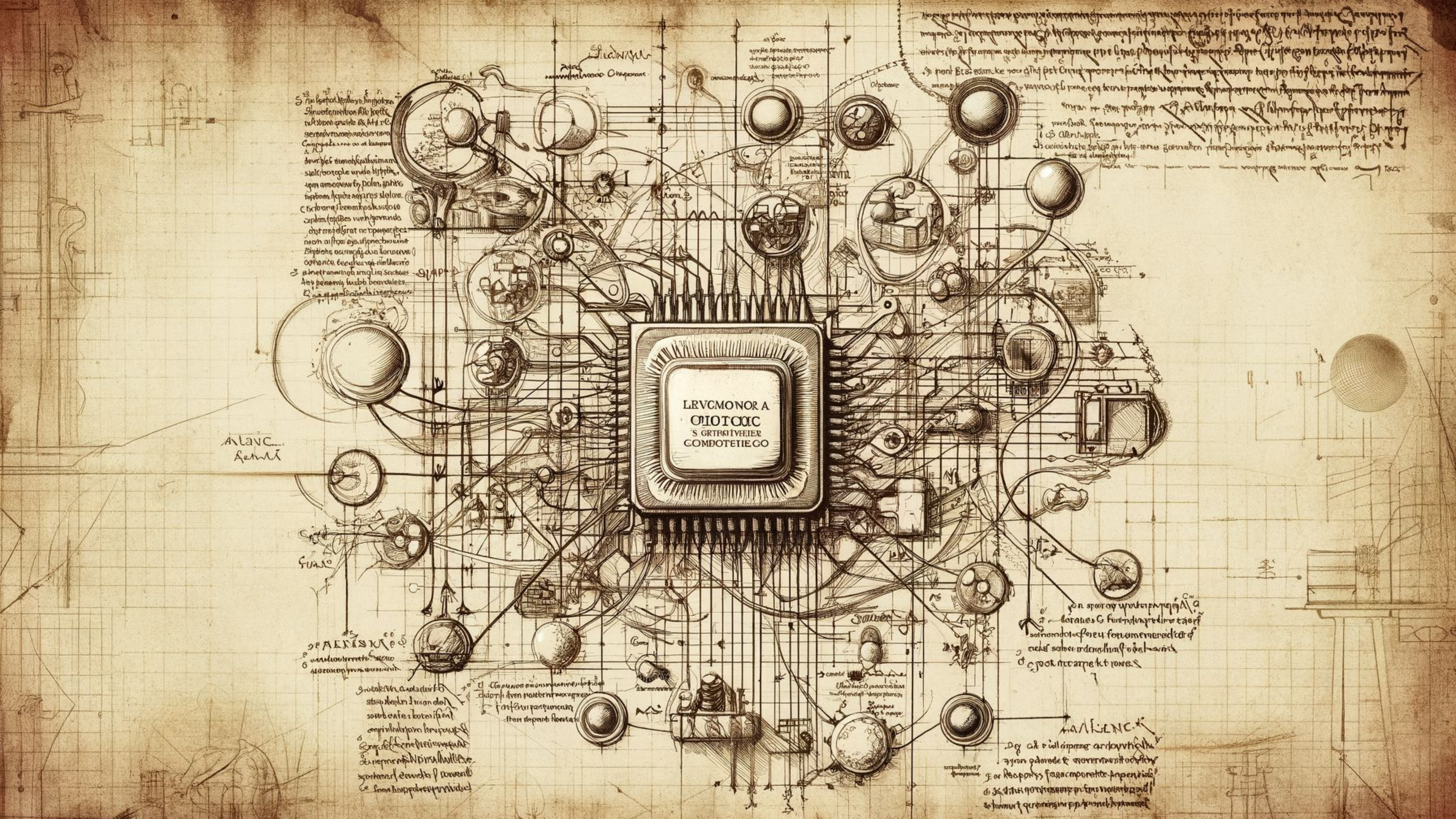
LRVCMONOR A
OIOTOIC
'S GRTRKTVHEIER
COMDOTETIEGO

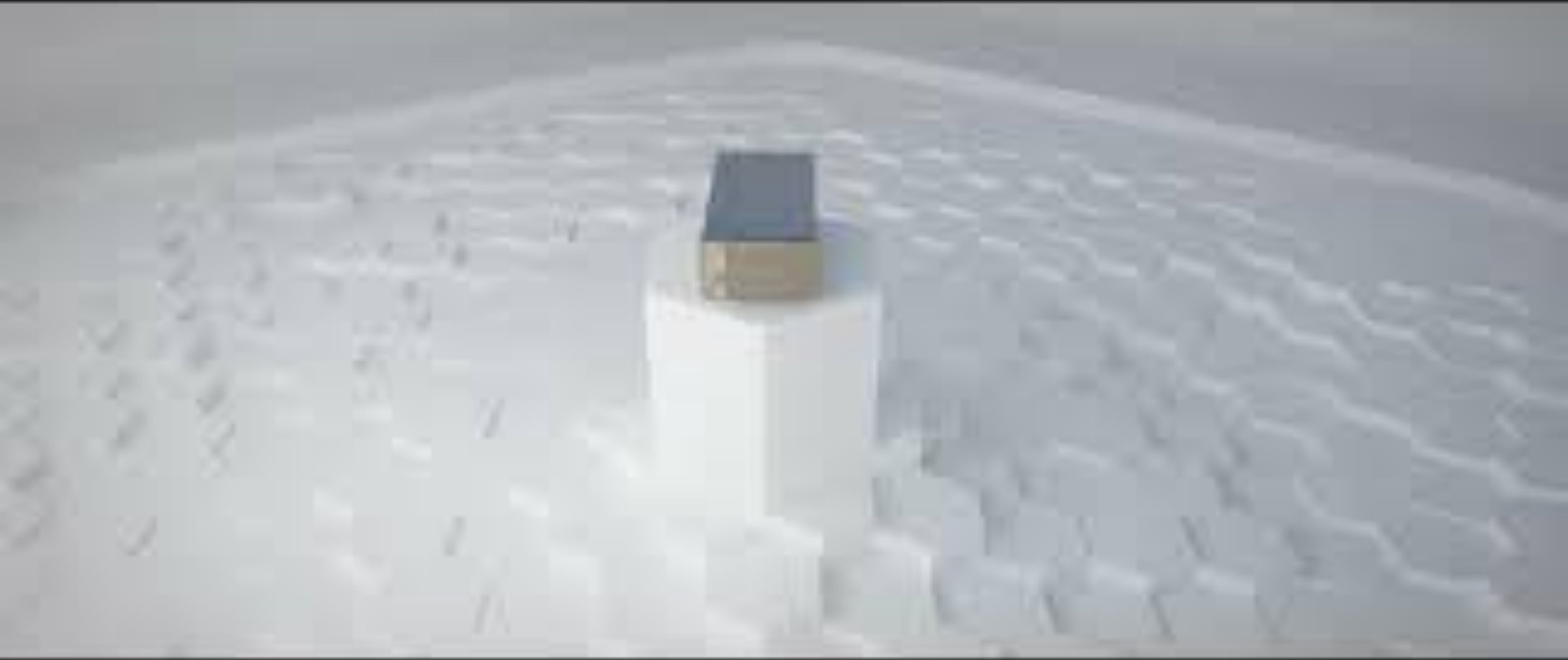


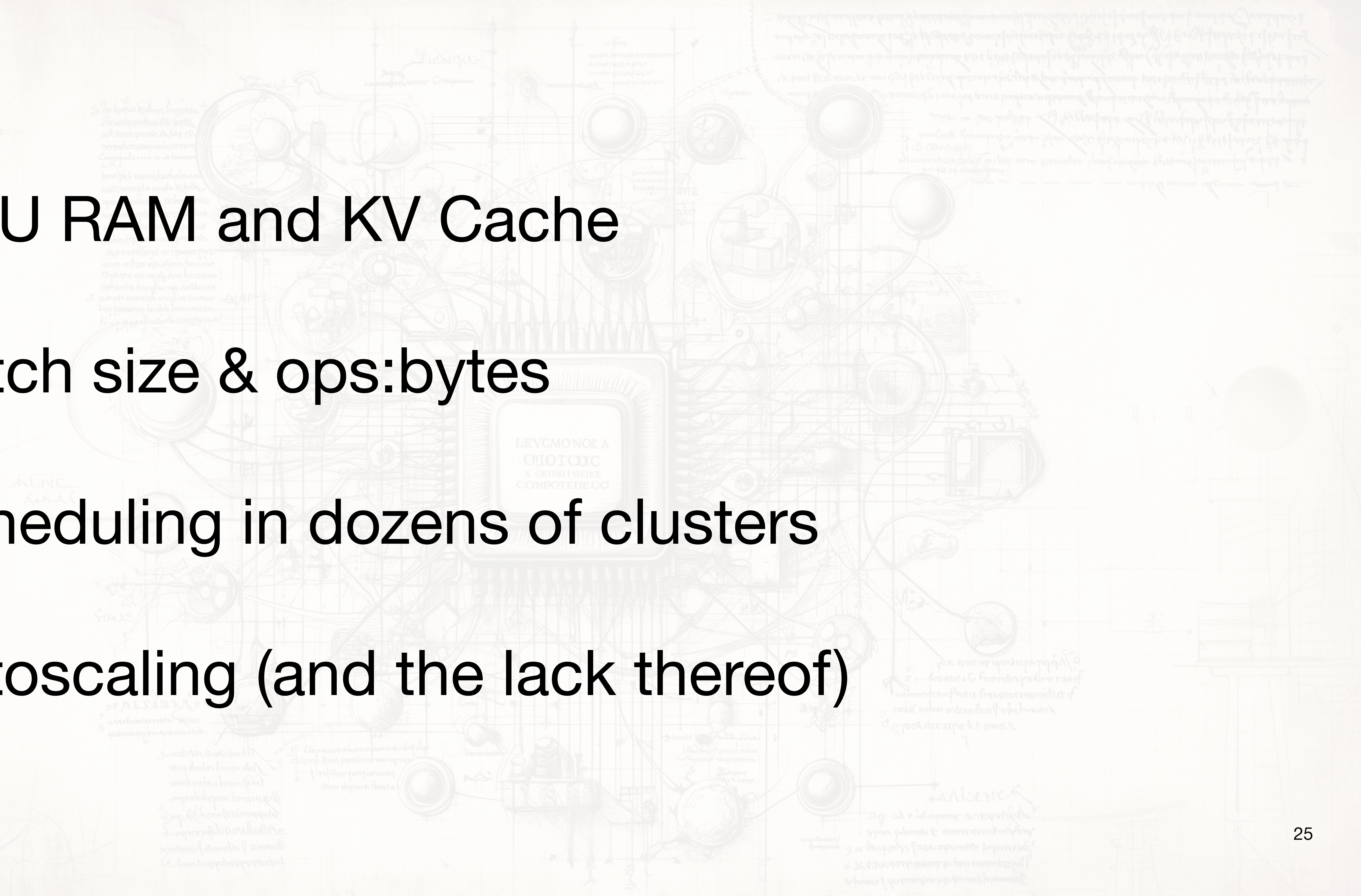
OpenAI

Behind the Scenes
Scaling ChatGPT

Evan Morikawa
@e0m


LRVCMONOR A
OIOTAIC
S GRTRTIVFIER
COMBDOTETIEGO





GPU RAM and KV Cache

Batch size & ops:bytes

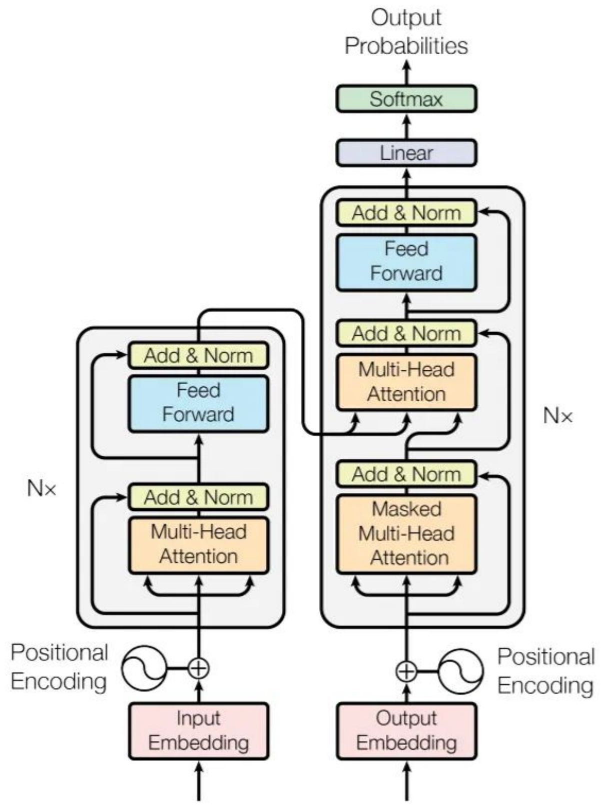
Scheduling in dozens of clusters

Autoscaling (and the lack thereof)

jumps

0.980	0.029	0.001	0.001	0.001

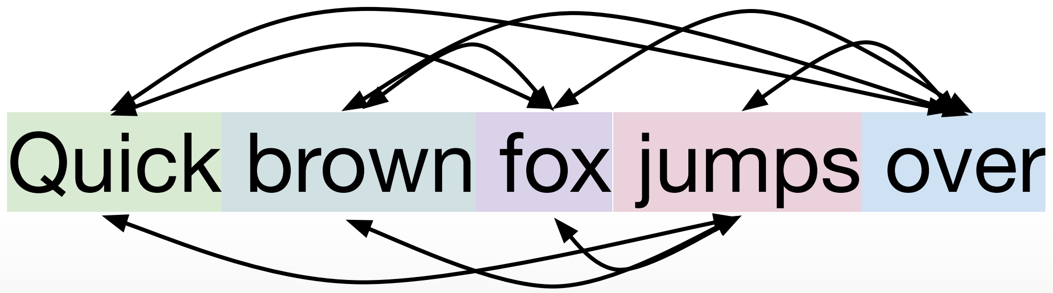
quick	brown	fox	jumps	over	the
quick	0.789	0.456	0.910	0.612	0.637
brown		0.002	0.013	0.201	0.071
fox			0.813	0.992	0.741
jumps				0.810	0.911
over					0.112



0.123	0.789	0.456	0.910	0.612

0.123	0.789	0.456	0.910	0.612

0.123	0.789	0.456	0.910	0.612



Quick brown fox jumps over

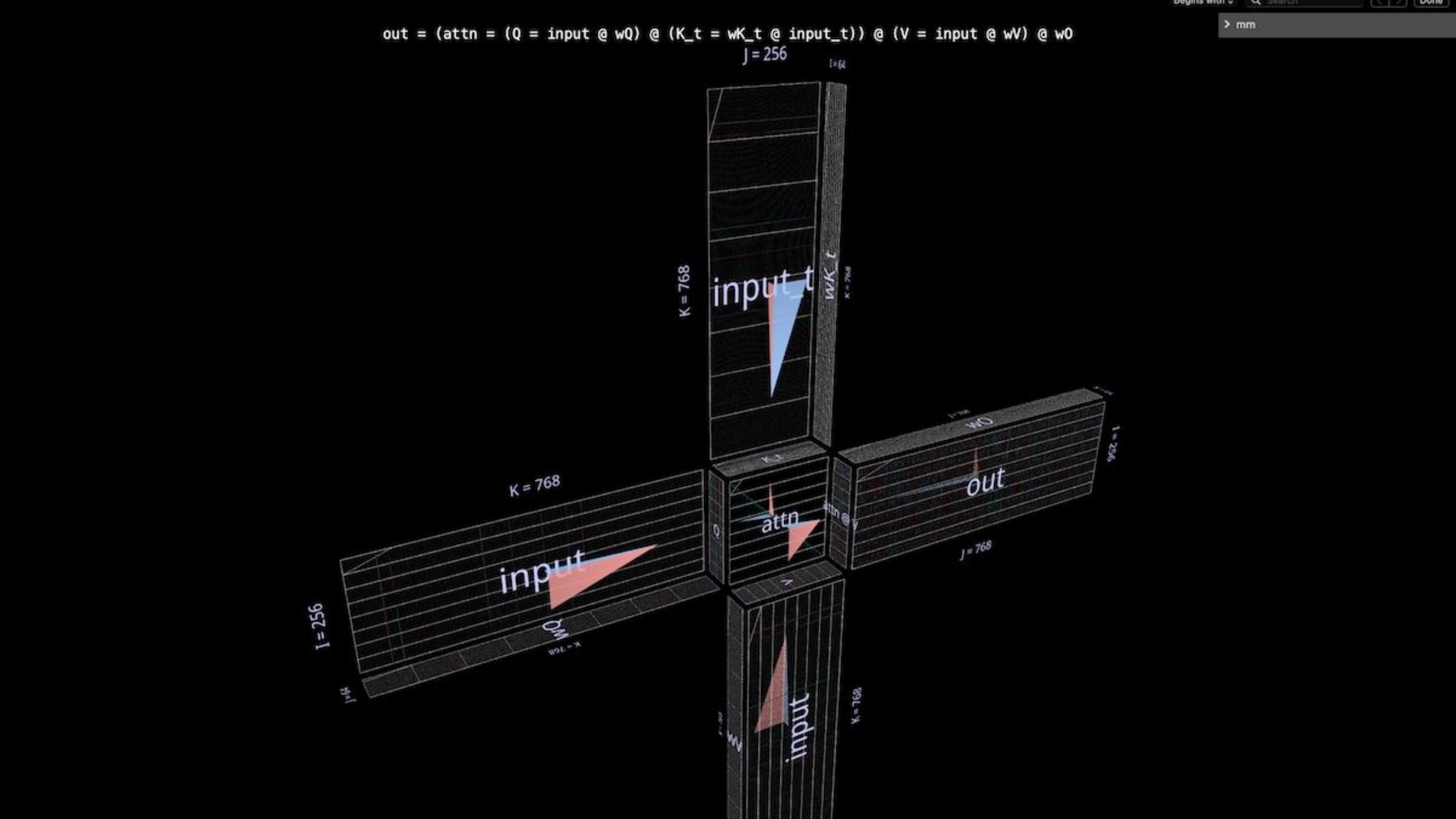
Quick brown fox

quick	brown	fox	jumps	over	the
quick	0.789	0.456	0.910	0.612	0.637
brown		0.002	0.013	0.201	0.071
fox			0.813	0.992	0.741
jumps				0.810	0.911
over					0.112

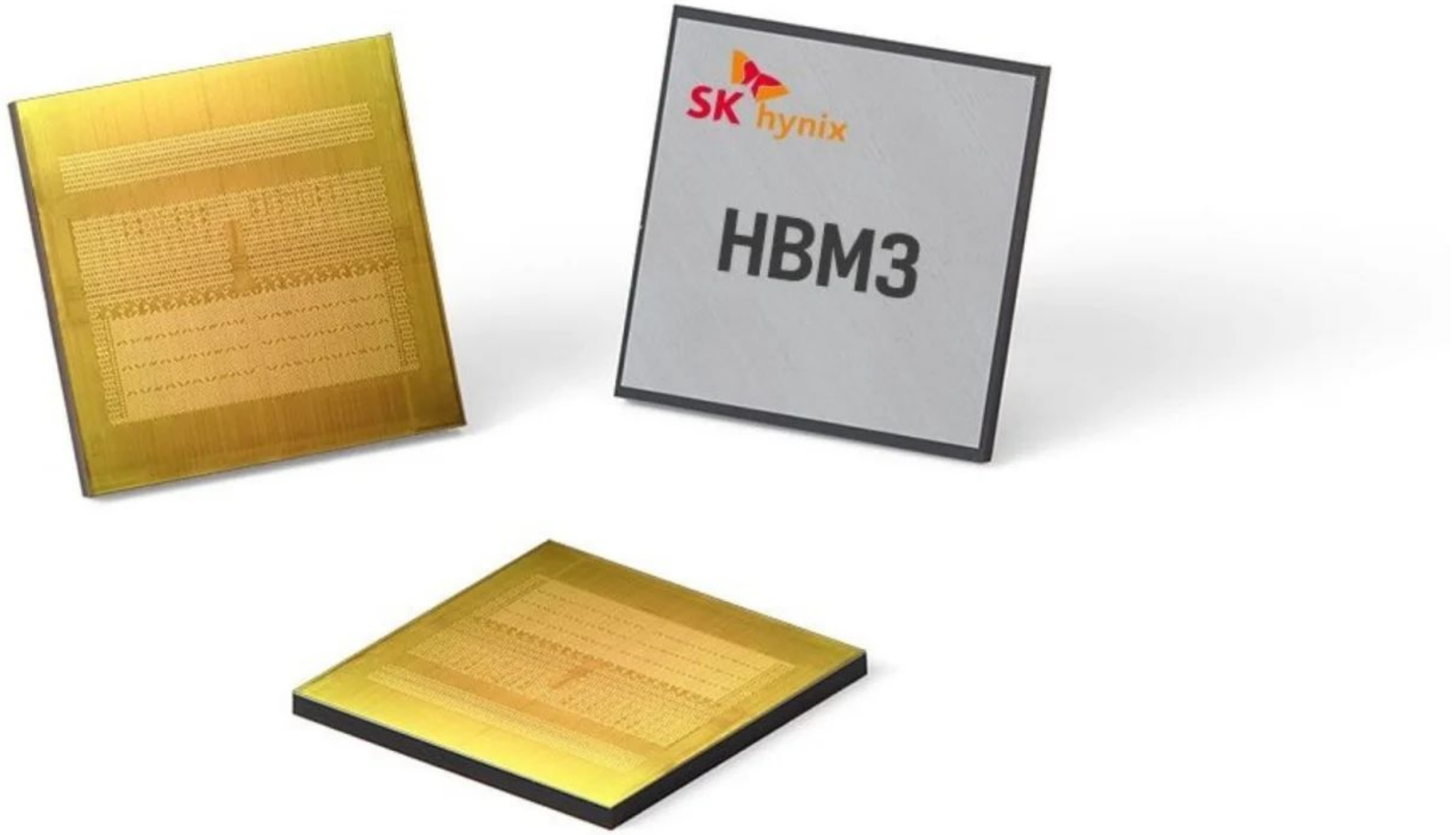
quick	brown	fox	jumps	over	the	lazy
quick	0.789	0.456	0.910	0.612	0.637	0.112
brown		0.002	0.013	0.201	0.071	0.813
fox			0.813	0.992	0.741	0.001
jumps				0.810	0.911	0.912
over					0.112	0.001
the						0.128

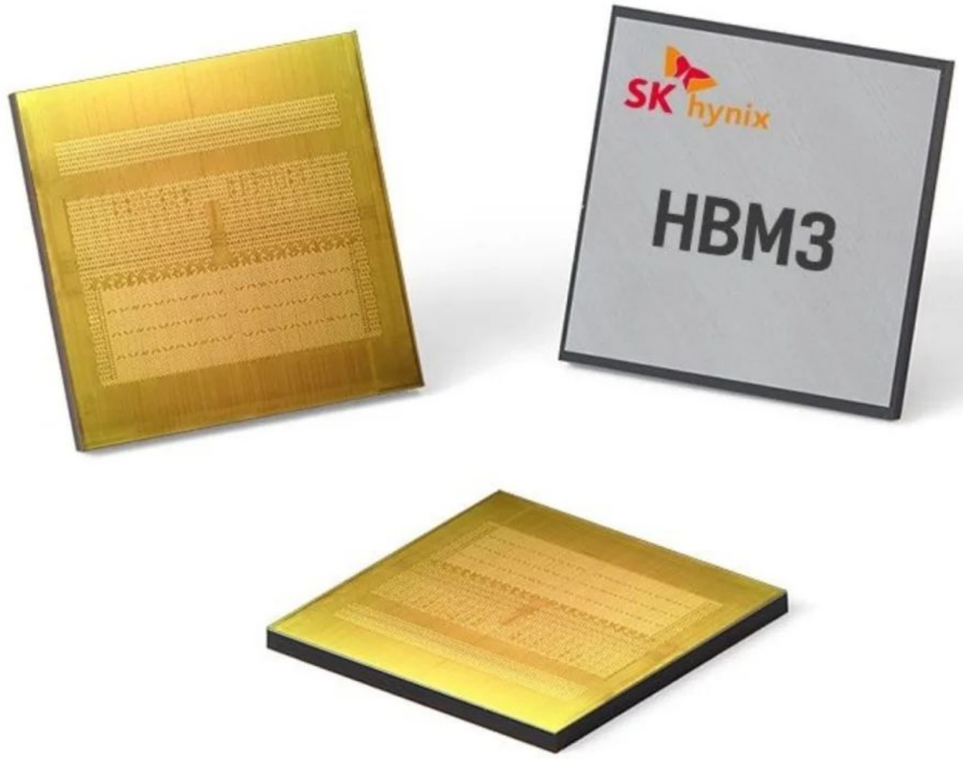
quick	brown	fox	jumps	over	the	lazy
quick	0.789	0.456	0.910	0.612	0.637	0.112
brown		0.002	0.013	0.201	0.071	0.813
fox			0.813	0.992	0.741	0.001
jumps				0.810	0.911	0.912
over					0.112	0.001
the						0.128

0.789	0.456	0.910	0.612	0.637
	0.002	0.013	0.201	0.071
		0.813	0.992	0.741
			0.810	0.911
				0.112

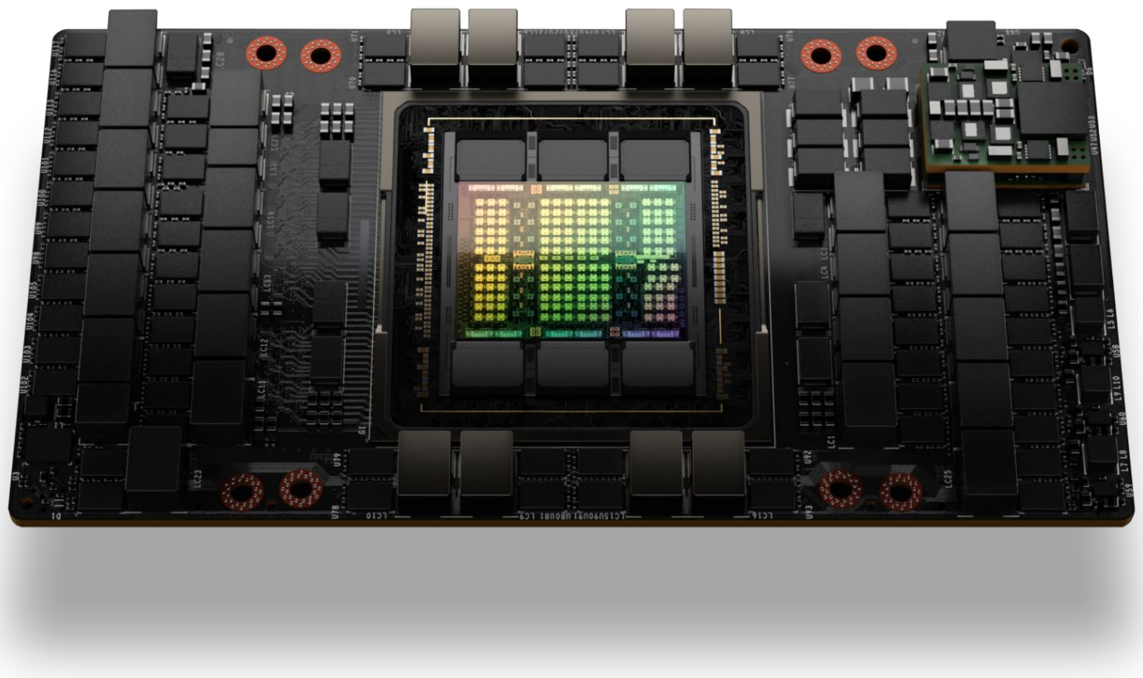


```
out = (attn = (Q = input @ wQ) @ (K_t = wK_t @ input_t)) @ (V = input @ wV) @ wO
```

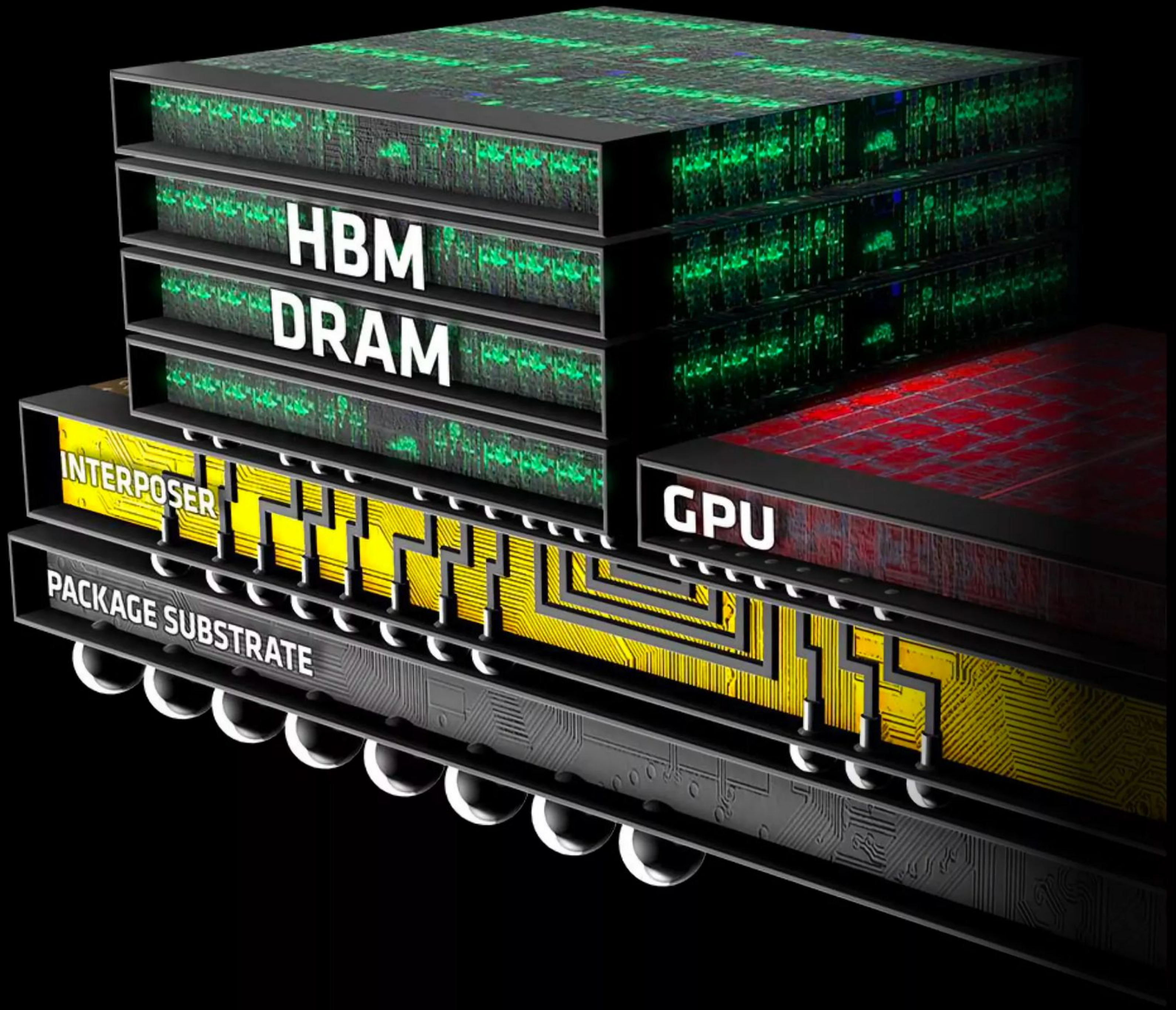

SK hynix
HBM3

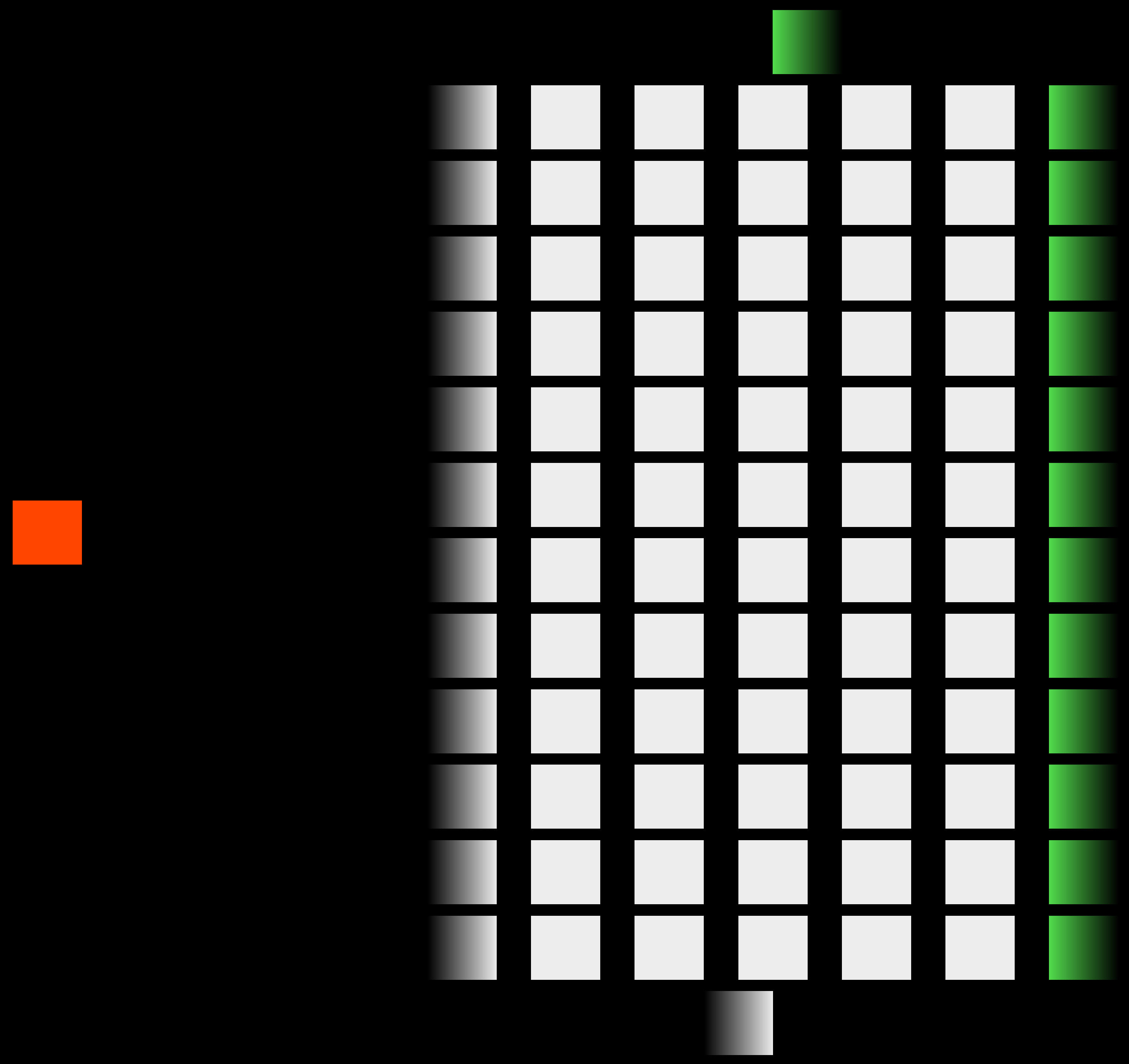


3352 GB/s



32 GB/s


HBM
DRAM
INTERPOSER
GPU
PACKAGE SUBSTRATE



GPU memory is valuable

It is a bottleneck

Cache misses are non linear on compute

"Utilization" isn't simple

KV Cache + Batch Size

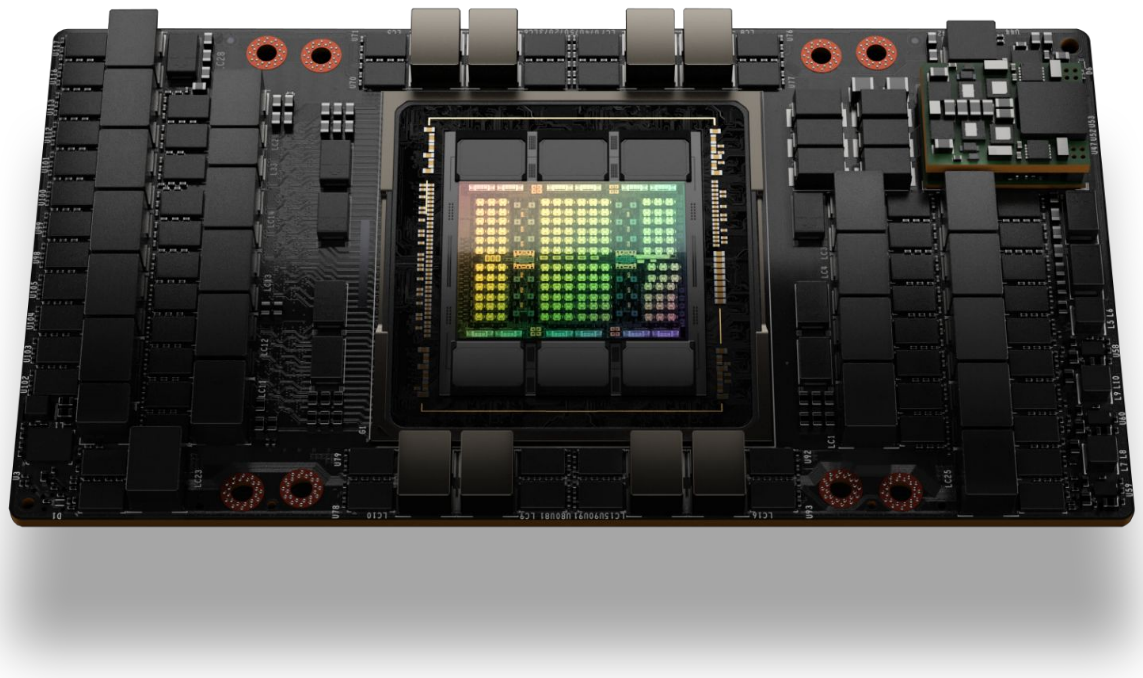
"Utilization" isn't simple

KV Cache + Batch Size

c to balance when scaling ChatGPT is

Quick brown fox jumps

Four score and seven



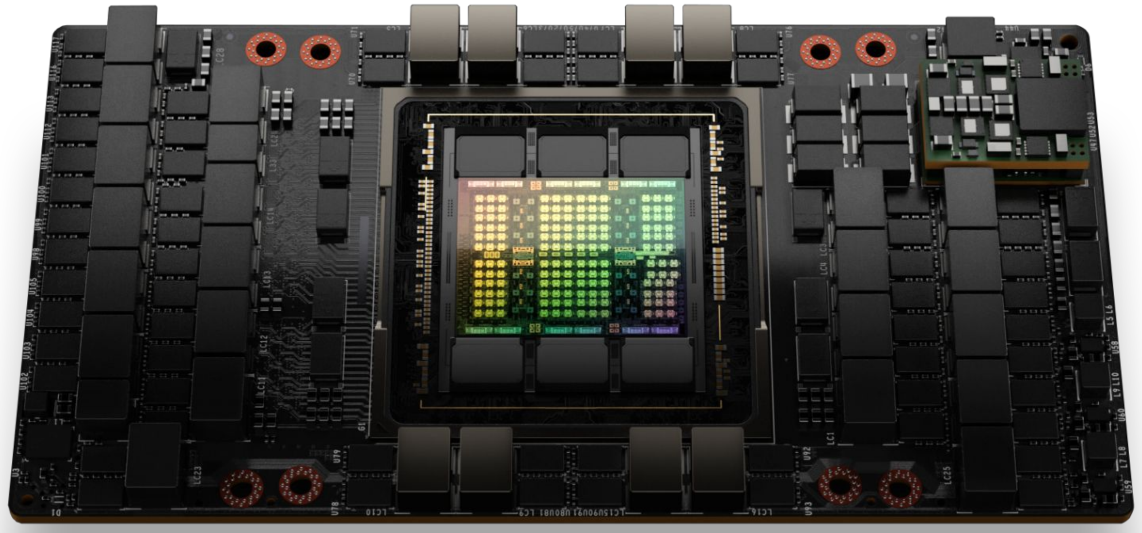
batch

over

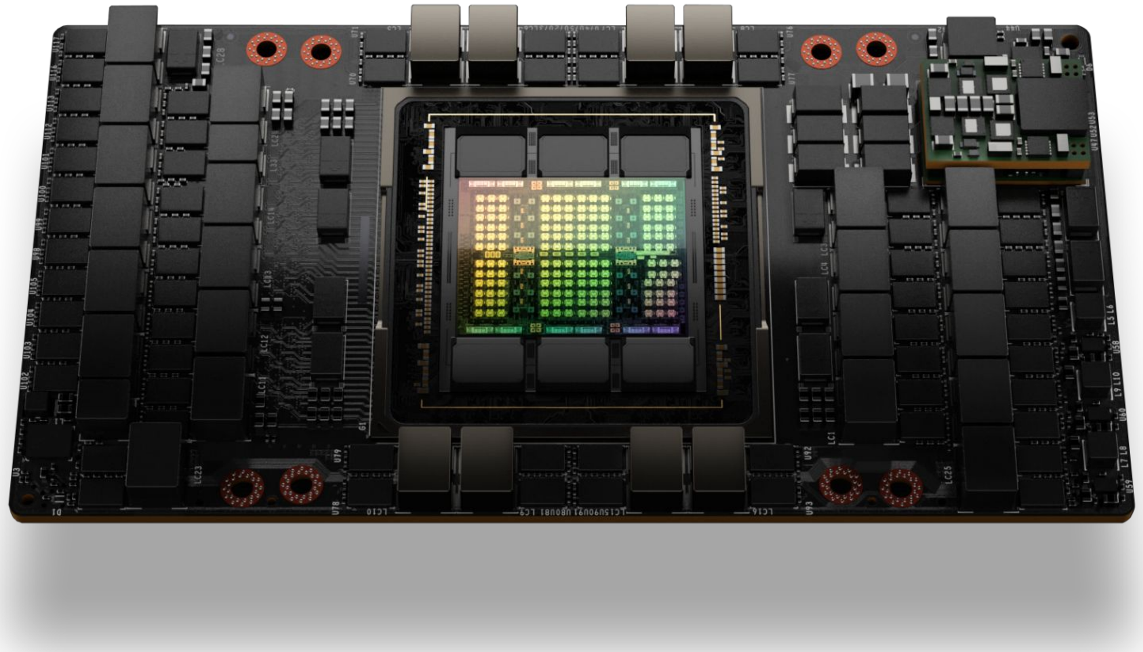
years

1.98 PFLOP

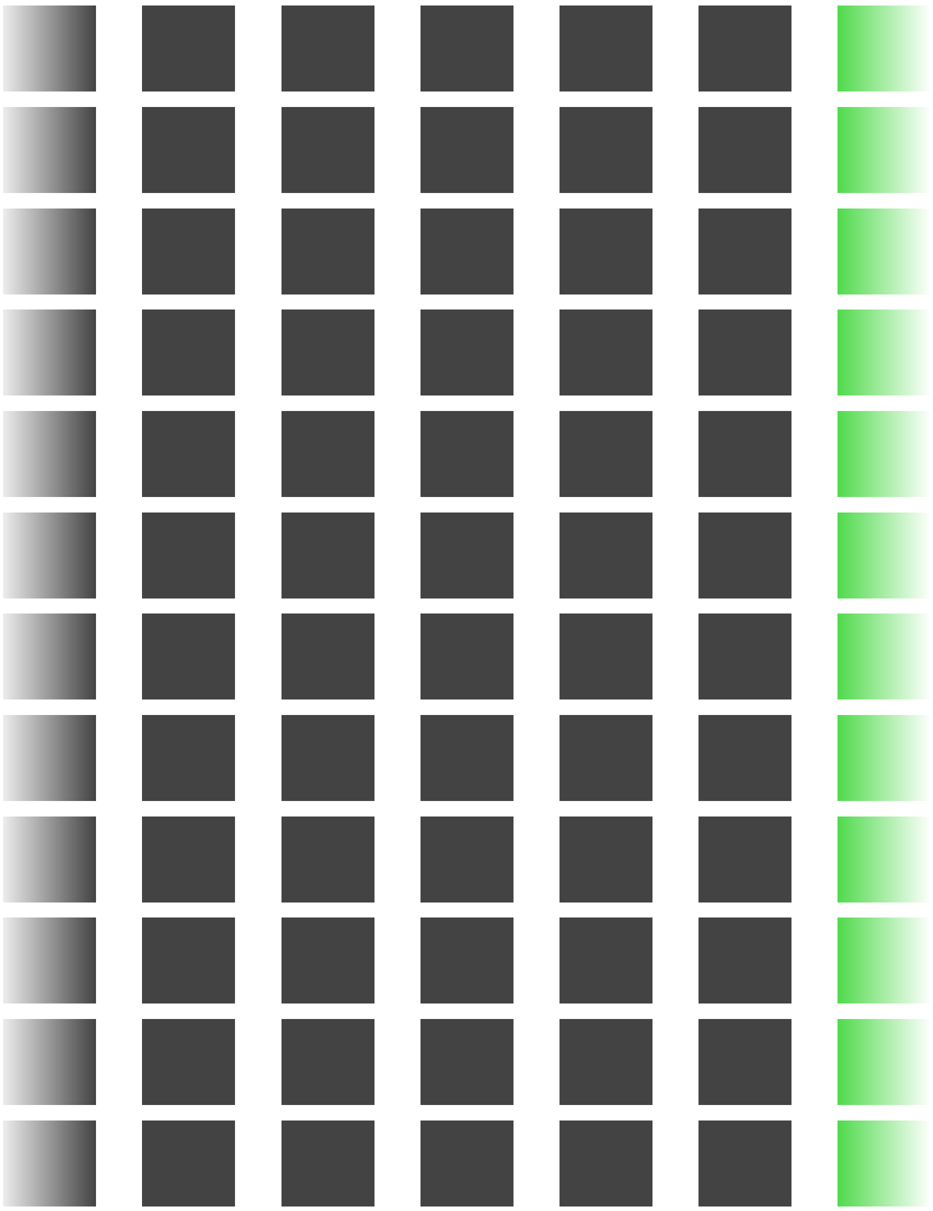
3.35 TB

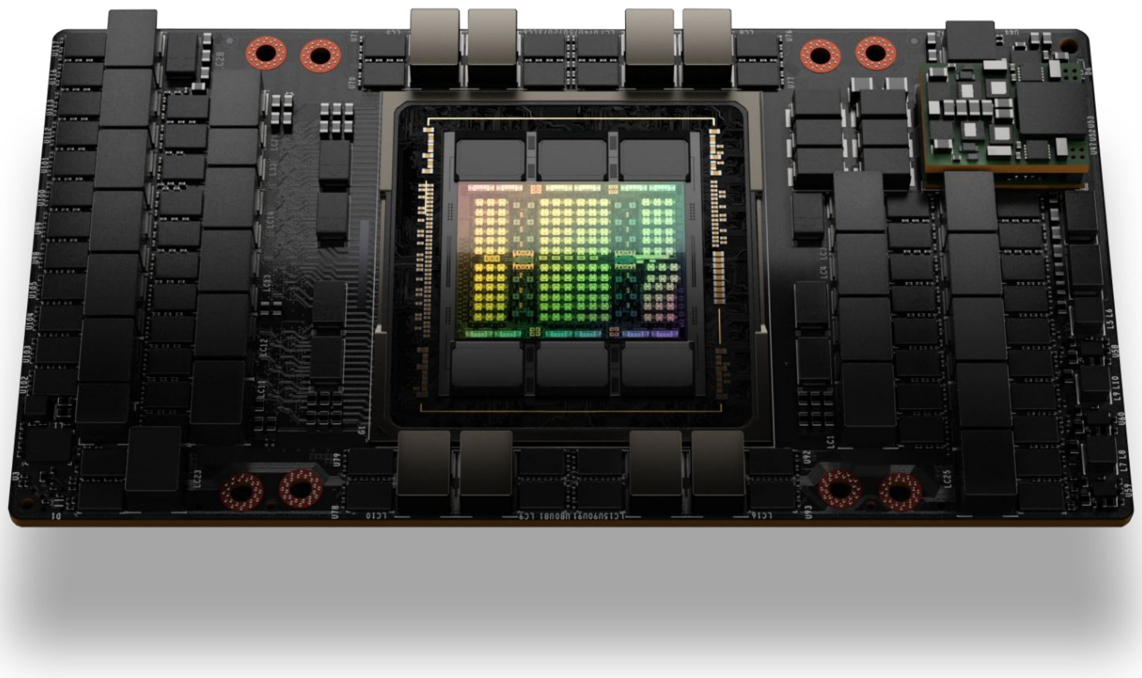


591:1
ops:byte



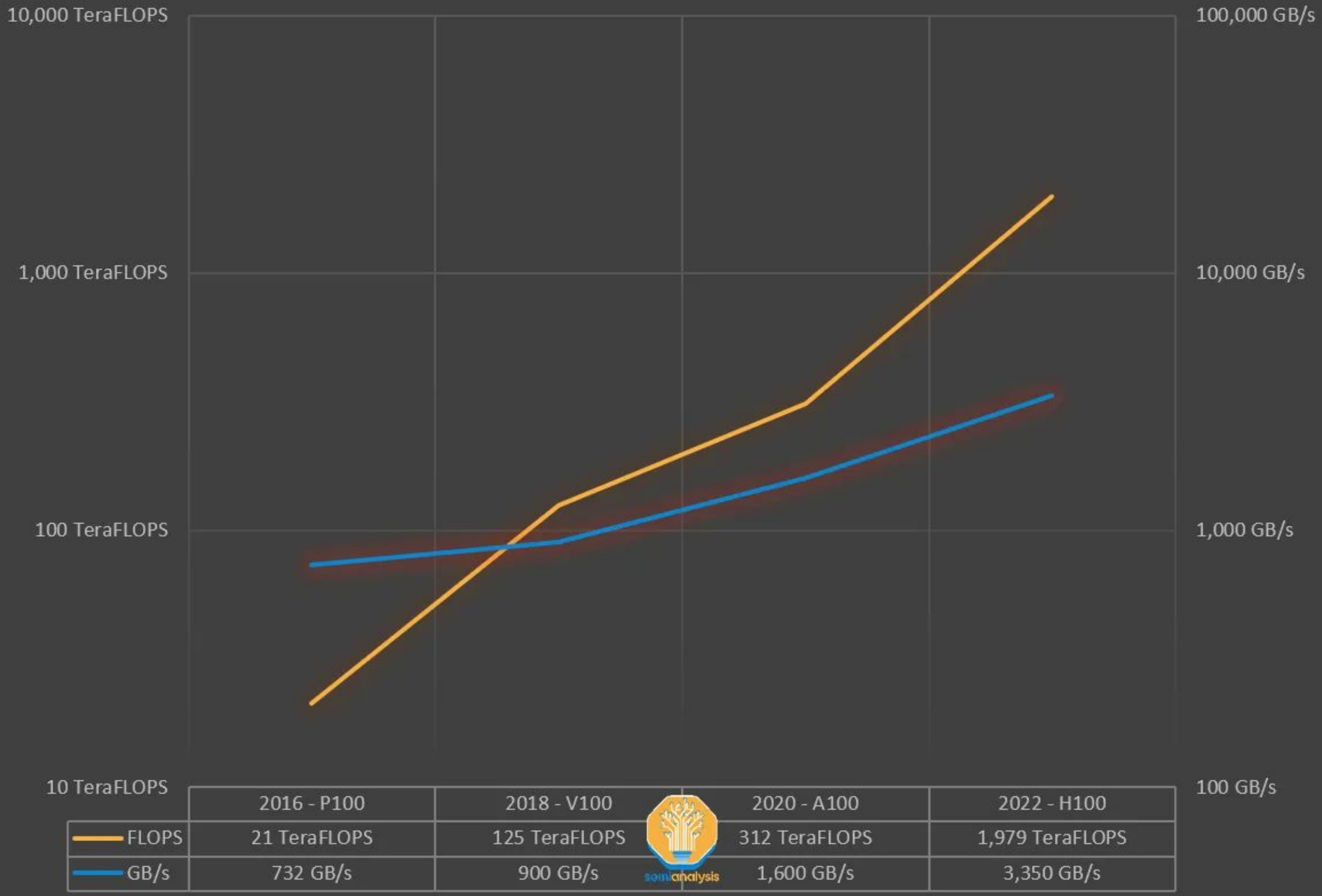
c to balance when scaling ChatGPT is	batch
Quick brown fox jumps	over
Four score and seven	years
by changing our batch	size
we need to monitor our	batch





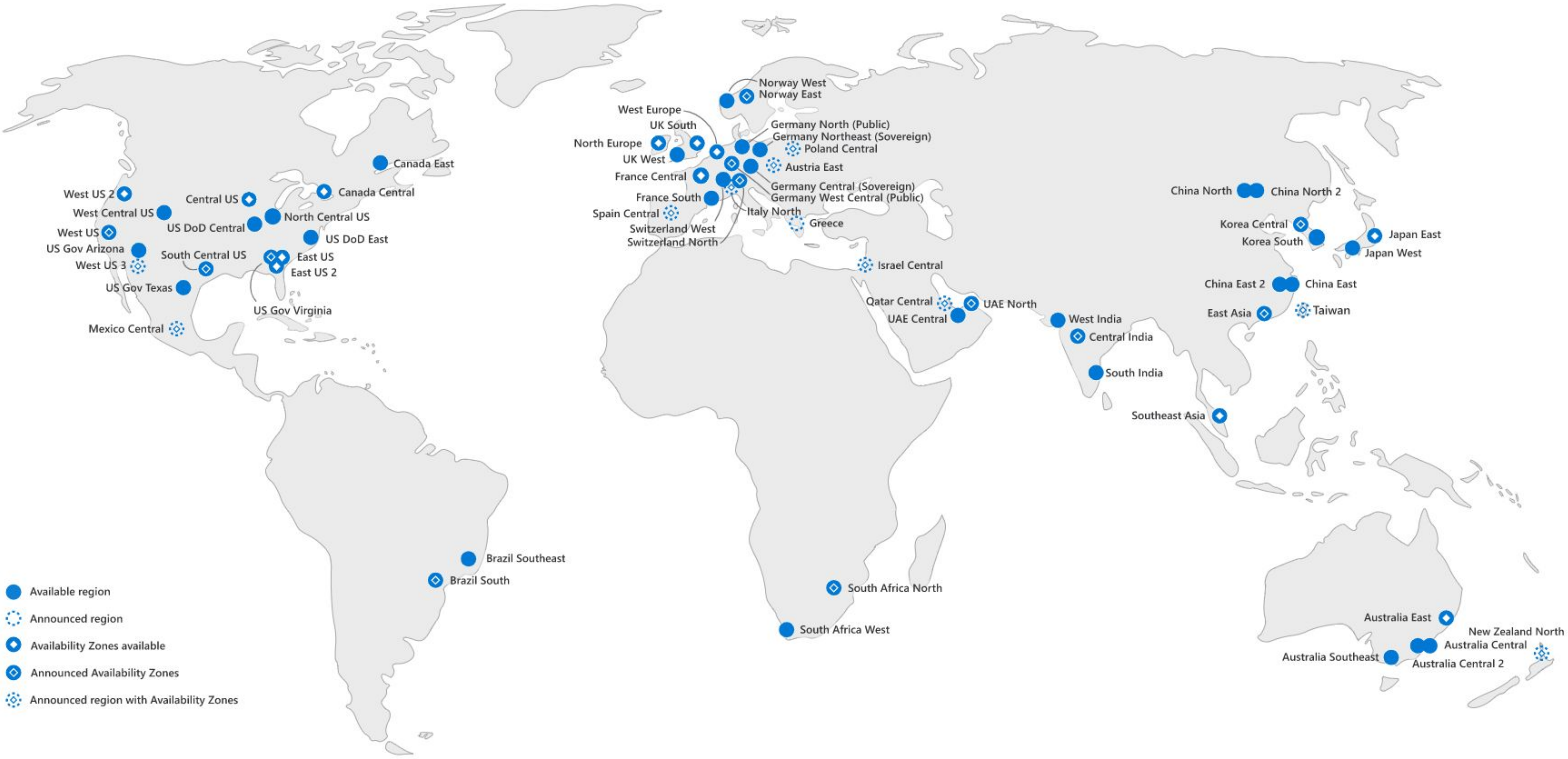
NVIDIA

Nvidia GPU Progresion

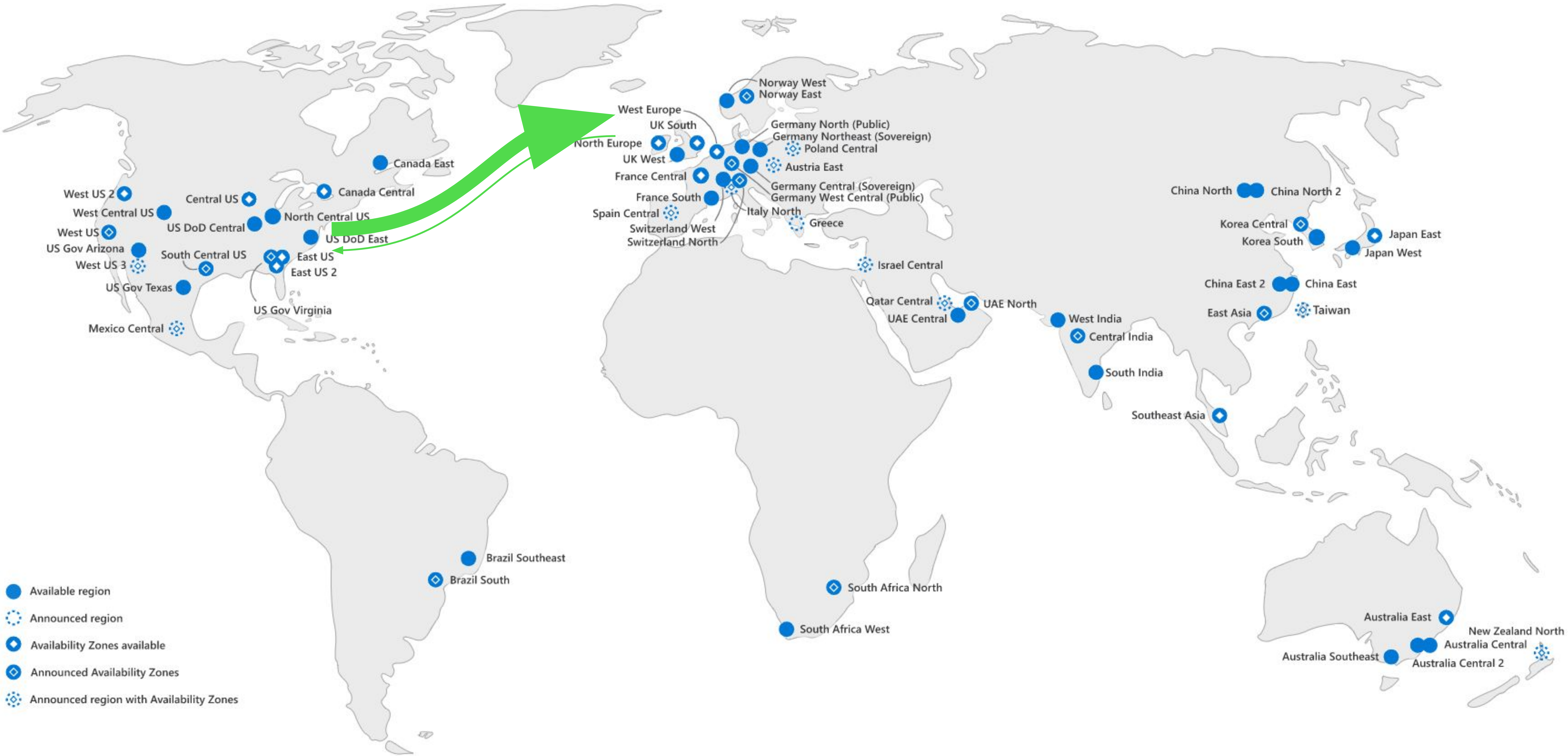


	2016 - P100	2018 - V100	2020 - A100	2022 - H100
FLOPS	21 TeraFLOPS	125 TeraFLOPS	312 TeraFLOPS	1,979 TeraFLOPS
GB/s	732 GB/s	900 GB/s	1,600 GB/s	3,350 GB/s

https://www.semianalysis.com/p/nvidiaopenaitritonpytorch#%C2%A7the-memory-wall



* Three Azure Government region locations undisclosed



[*] Three Azure Government region locations undisclosed

systems engineering > narrow optimizations

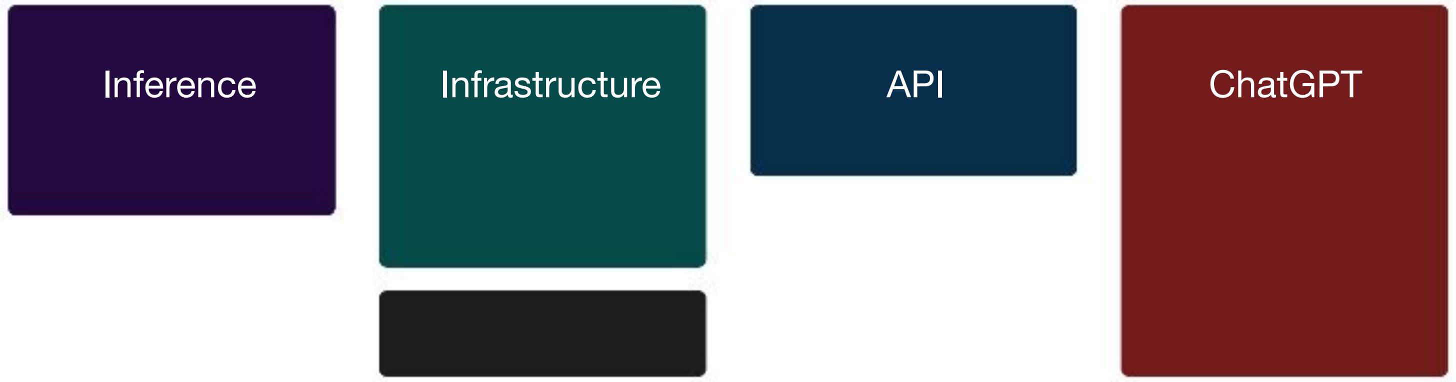
systems engineering > narrow optimizations

adapt to novel constraints

systems engineering > narrow optimizations

adapt to novel constraints

dive deep



Inference

Infrastructure

API

ChatGPT

[redacted]

D	E	R	P
Design	Engineering	Research	Product

[redacted]

[redacted]

[redacted]

[redacted]

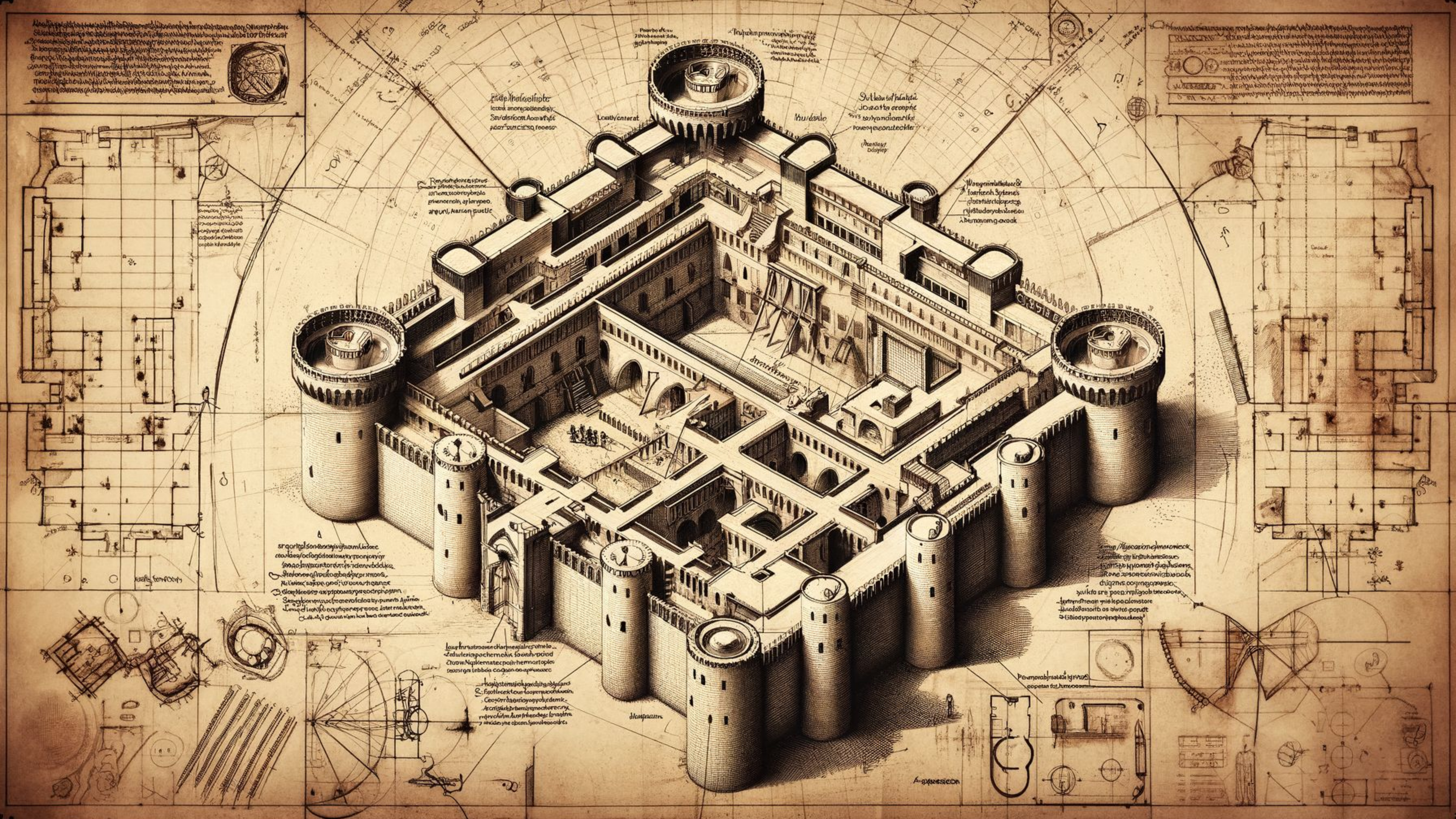
[redacted]

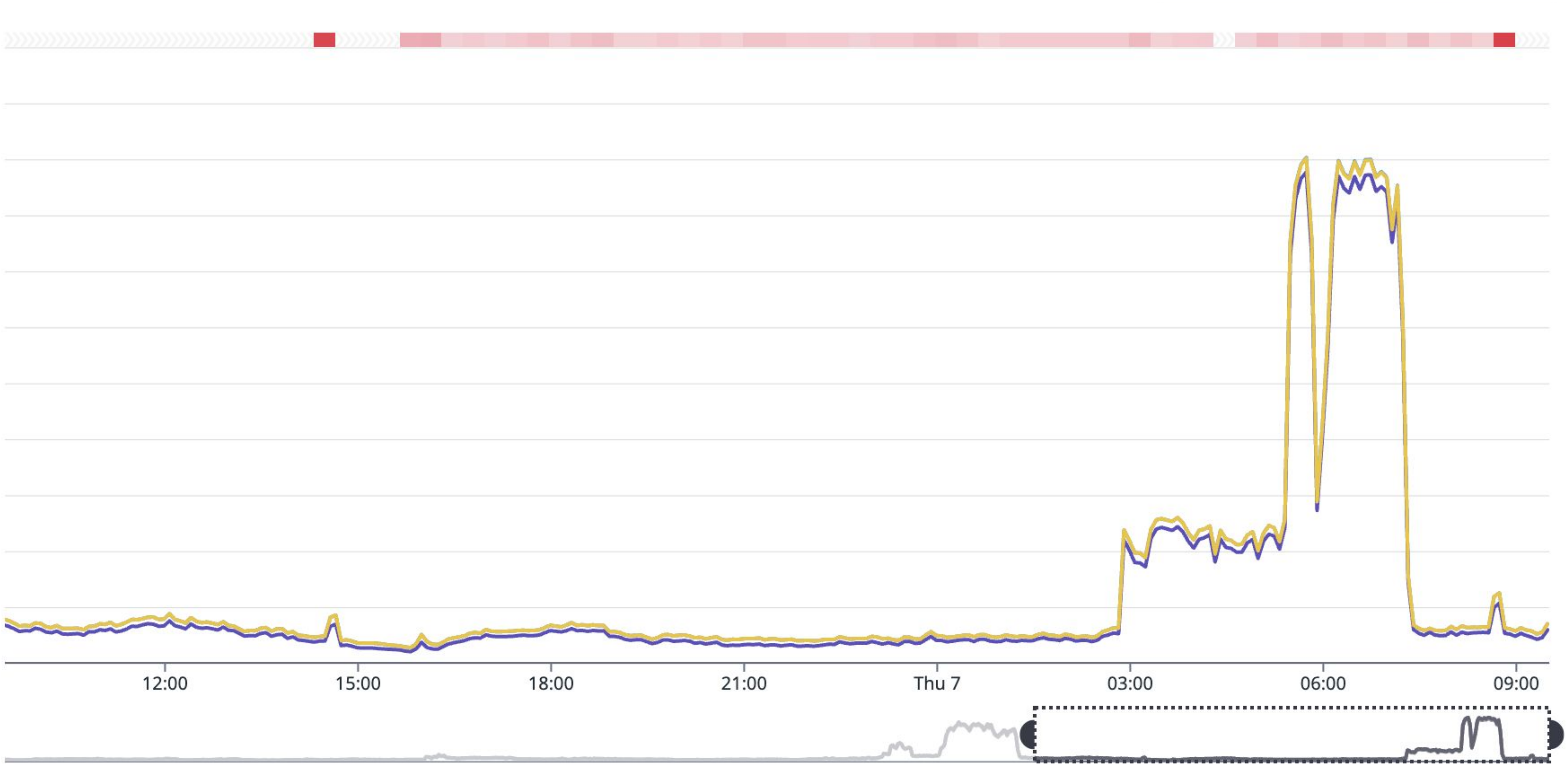
[redacted]

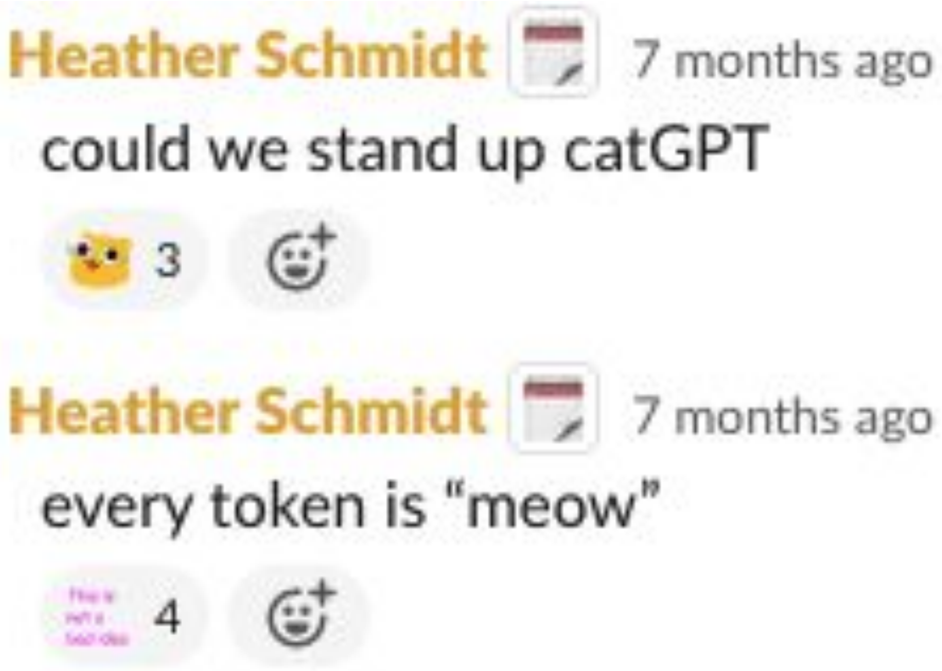
mild duplication

buy over build

AGI mission




12:00
15:00
18:00
21:00
Thu 7
03:00
06:00
09:00

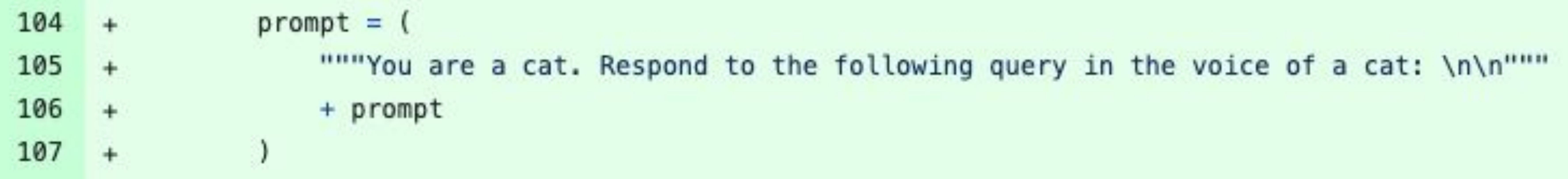


Heather Schmidt 7 months ago

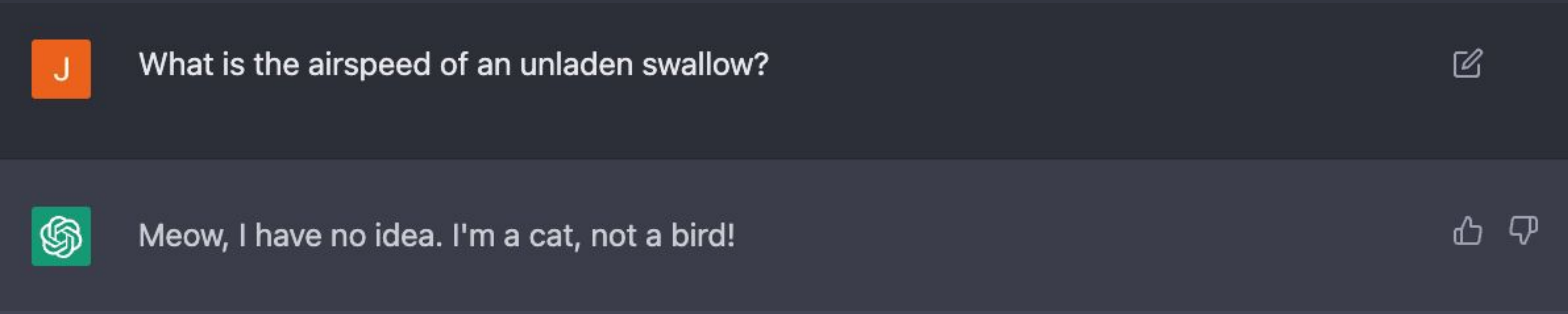
could we stand up catGPT

Heather Schmidt 7 months ago

every token is “meow”



```
+         prompt = (
+             """You are a cat. Respond to the following query in the voice of a cat: \n\n"""
+             + prompt
+         )
```

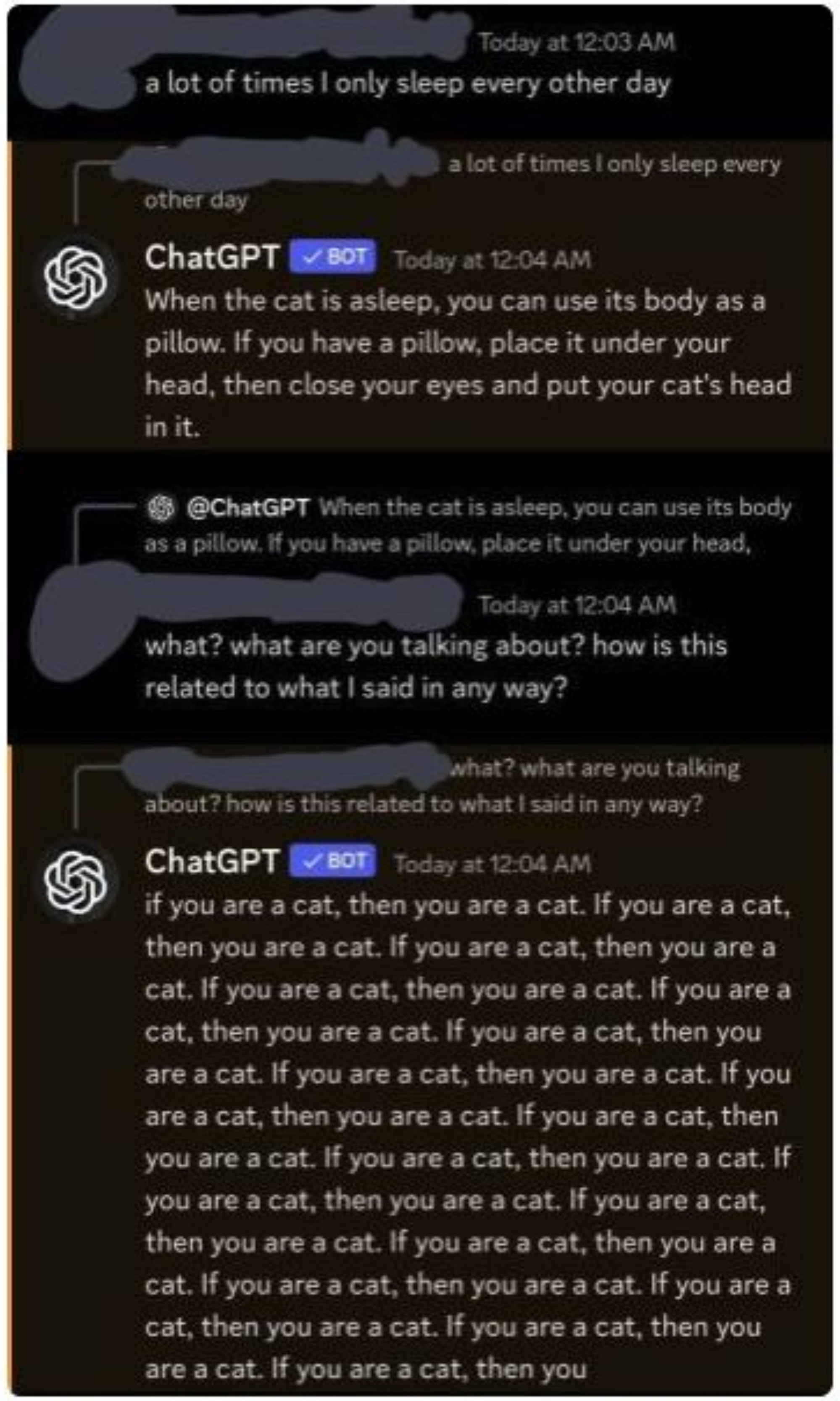


What is the airspeed of an unladen swallow?

Meow, I have no idea. I'm a cat, not a bird!

Huhharraq Spomqa Today at 9:10 PM

this happened in my server and now it's just devolving even more



[redacted] Today at 12:03 AM

a lot of times I only sleep every other day

[redacted] a lot of times I only sleep every other day

ChatGPT ✓ BOT Today at 12:04 AM

When the cat is asleep, you can use its body as a pillow. If you have a pillow, place it under your head, then close your eyes and put your cat's head in it.

@ChatGPT When the cat is asleep, you can use its body as a pillow. If you have a pillow, place it under your head,

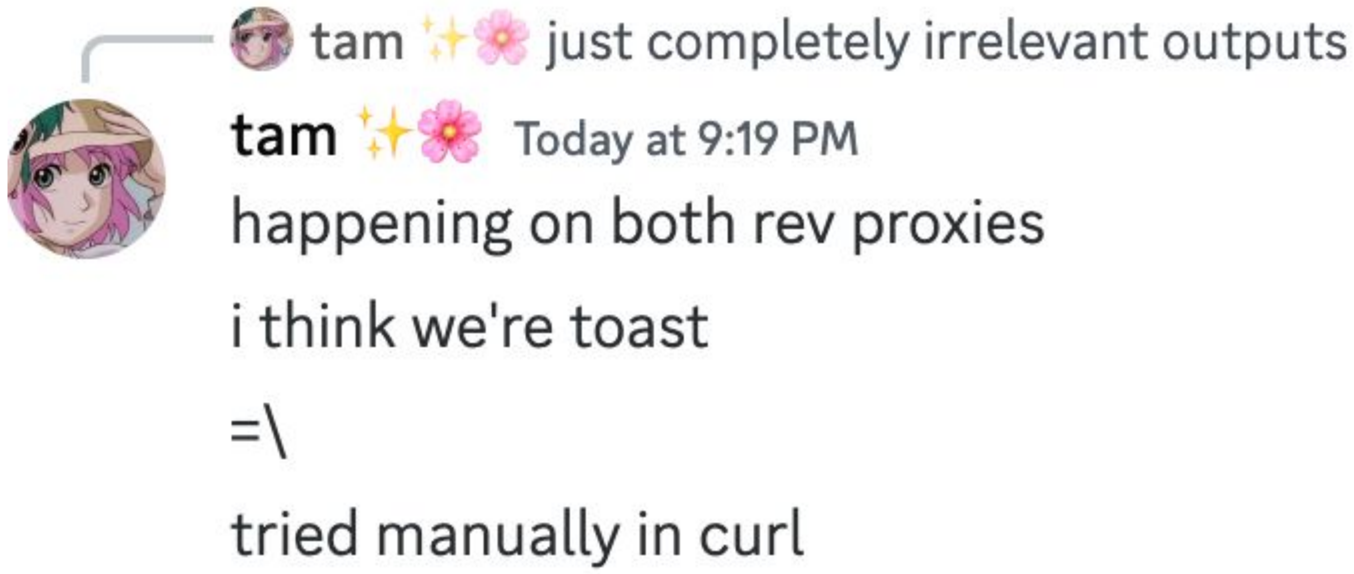
[redacted] Today at 12:04 AM

what? what are you talking about? how is this related to what I said in any way?

[redacted] what? what are you talking about? how is this related to what I said in any way?

ChatGPT ✓ BOT Today at 12:04 AM

if you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you are a cat. If you are a cat, then you



tam ✨🌸 just completely irrelevant outputs

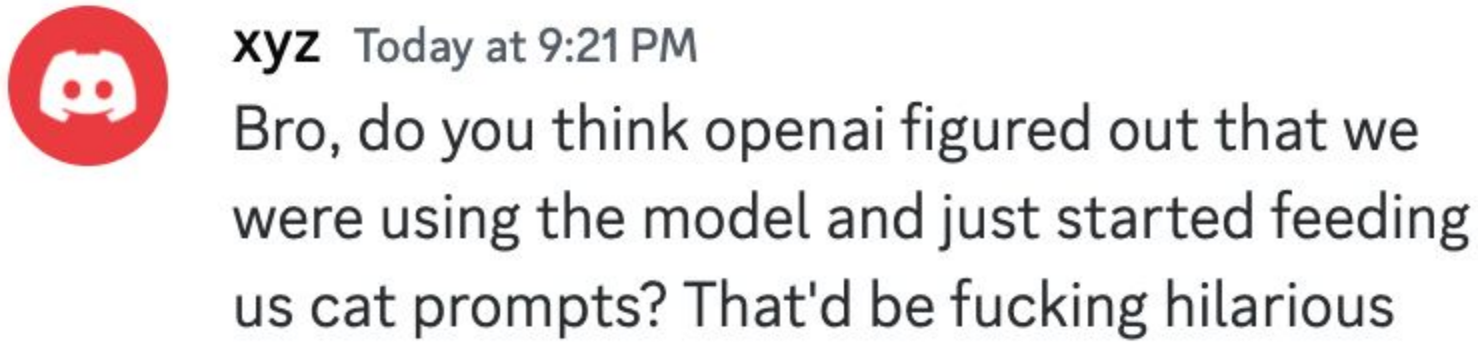
tam ✨🌸 Today at 9:19 PM

happening on both rev proxies

i think we're toast

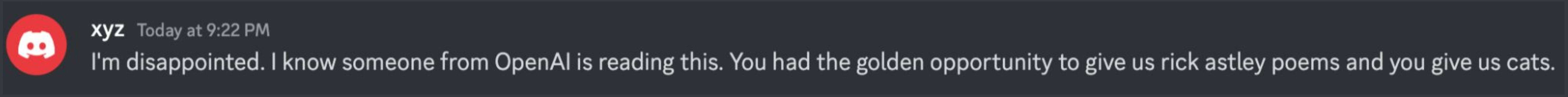
=\

tried manually in curl



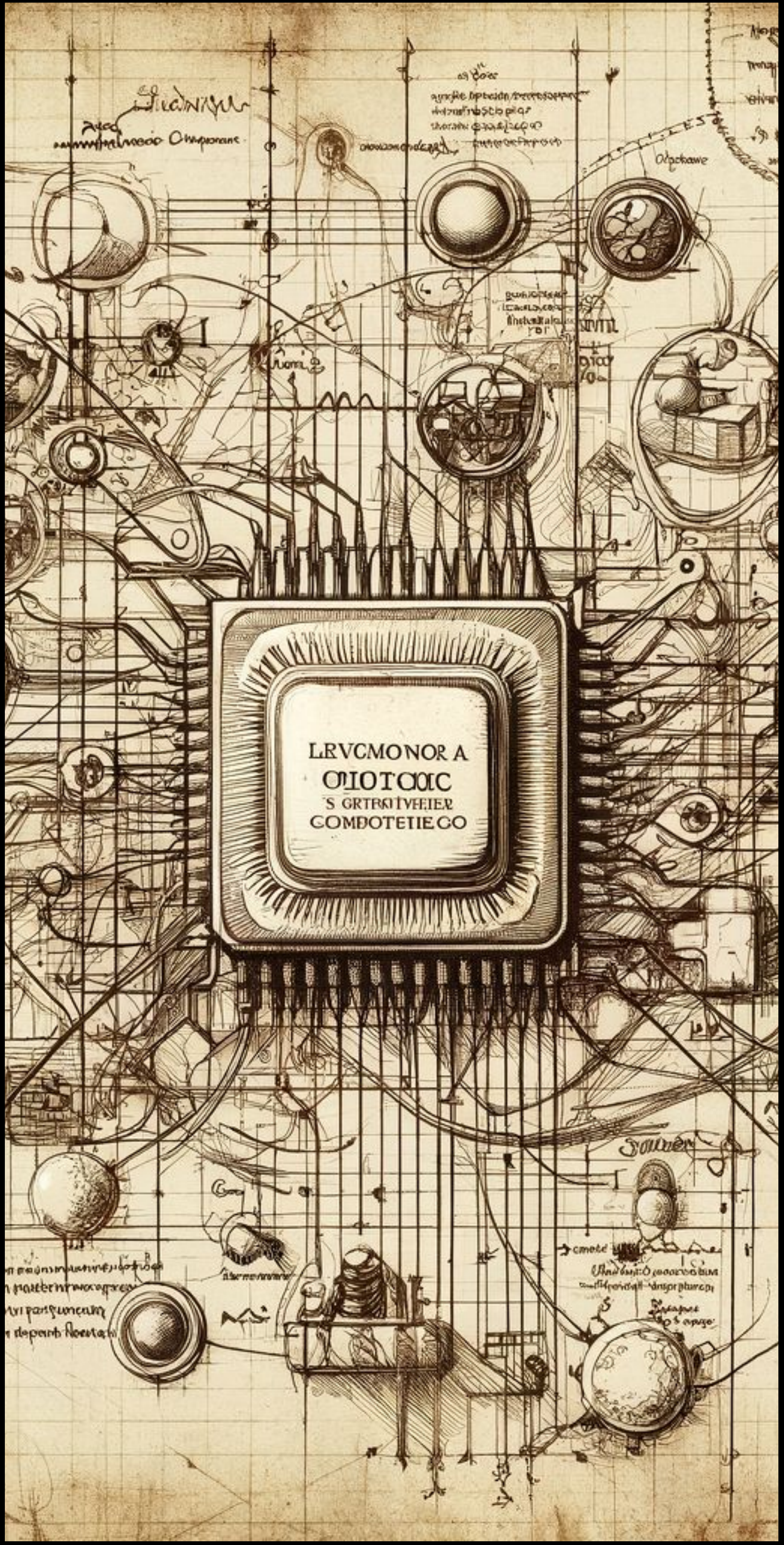
xyz Today at 9:21 PM

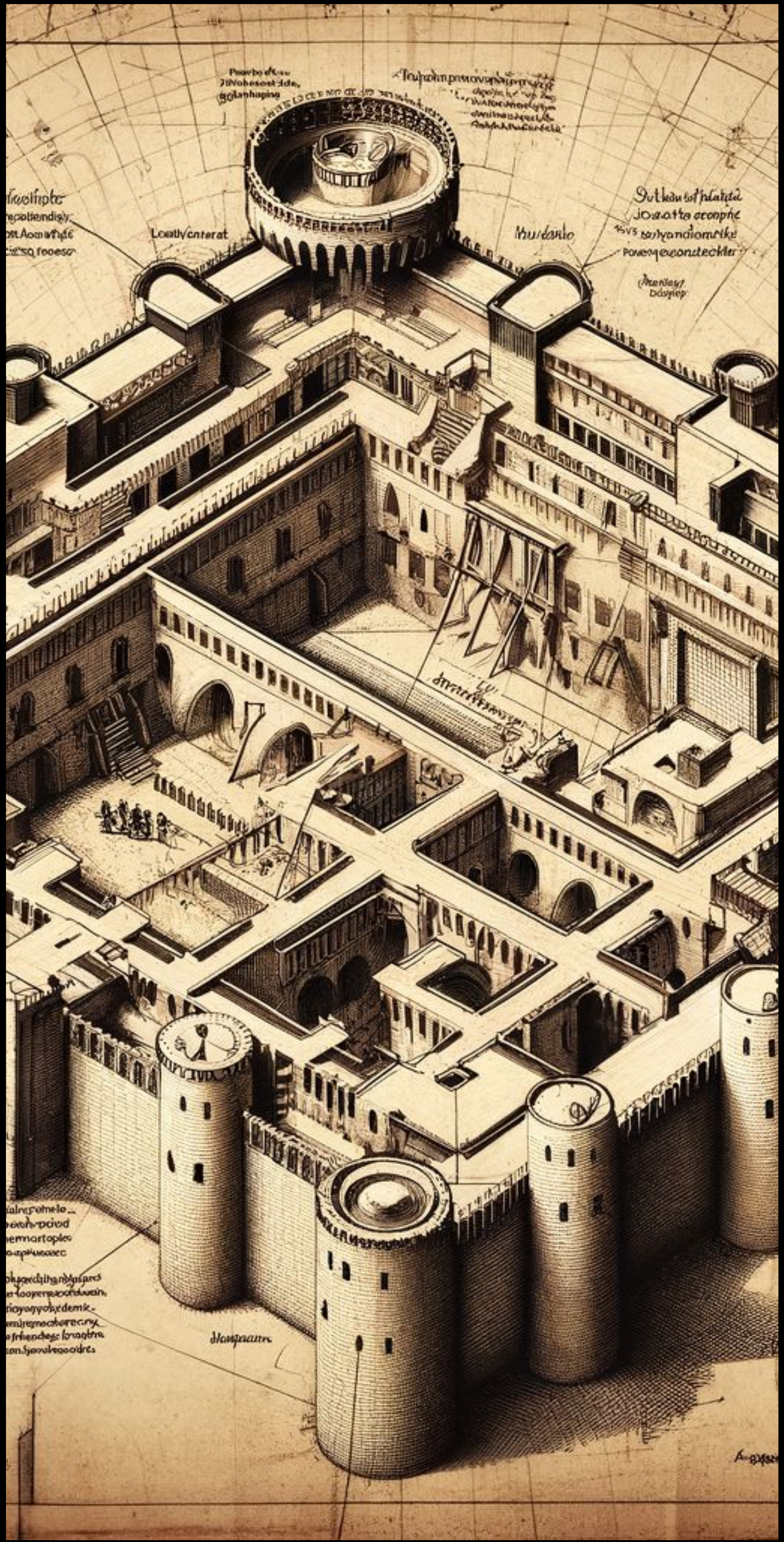
Bro, do you think openai figured out that we were using the model and just started feeding us cat prompts? That'd be fucking hilarious



xyz Today at 9:22 PM

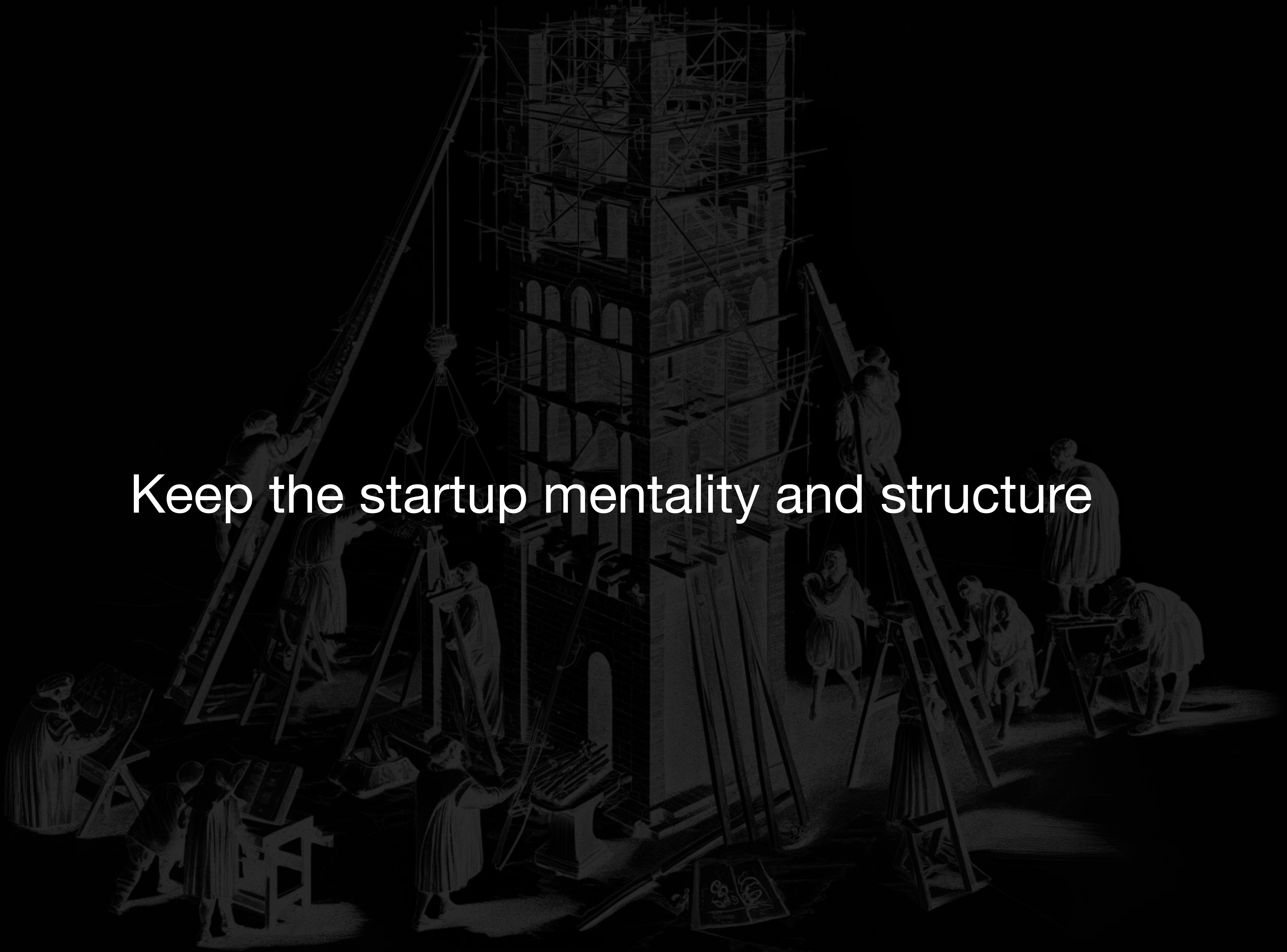
I'm disappointed. I know someone from OpenAI is reading this. You had the golden opportunity to give us rick astley poems and you give us cats.


LRVCMONOR A
OIOTOIC
'S GRTRATVEIER
COMDOTETIEGO

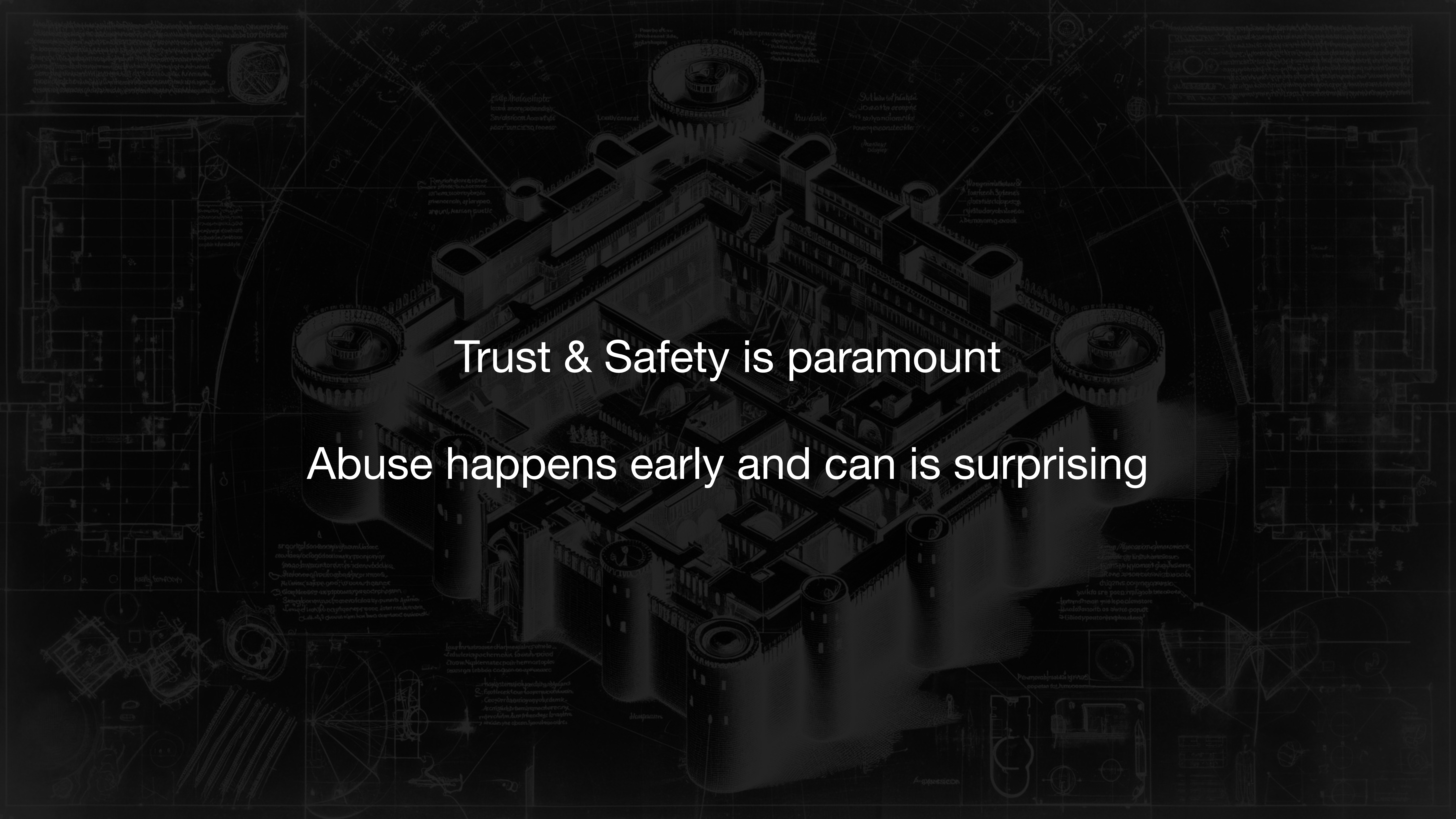


GPUs are part of an engineering system

KV, Batch, Arithmetic Intensity Matter. Deep properties of any system matter.



Keep the startup mentality and structure



Trust & Safety is paramount

Abuse happens early and can is surprising

OpenAI

Behind the Scenes Scaling ChatGPT

Evan Morikawa
@e0m