

CONF'

Steve Houël

Senior Solution Architect @ AWS



Le **19/06/2024** à **19h**
Au Palace, 4 rue Voltaire, Nantes



icilundi

**“Démystifiez le développement
d'applications web d'IA générative”**



[sf≡ir] lonestone

TUE, JUN 13, 2023, 7:00 PM CEST

2 Talks sur l'IA Générative + Apéro 🎉

⌚ This event has passed



30 attendees



Qui est déjà venu au meetup Gen AI Nantes ?

Qui sait coder ?

Qui fait des projets GenAI ?

Les autres, que cherchez-vous dans ce meetup ?

Qui était au Shift Hackathon ?

:D :D :D

Schedule

- MICROPHONE 1- News GenAI
- MICROPHONE 2- Talk Steve Houël
- DRINKING BEER 3- Enjoy

About LLM batching

About LLM batching

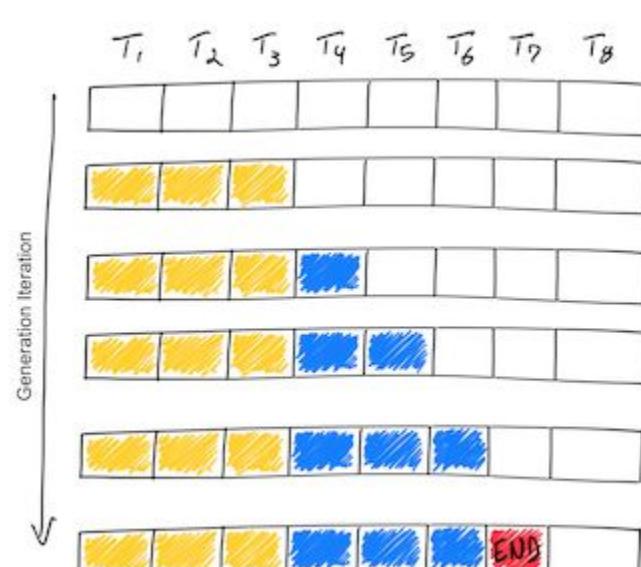
- LLM are memory-IO bound
- Running a 13B model FP16 -> 26GB
- 1 MB of memory per token
- 1k tokens output -> 1 GB

About LLM batching

- Nvidia H100 -> 80GB memory
- 26GB for model + 54GB for KV cache
- 54GB == 54k tokens
- 54k tokens == ~7 concurrent req of 8k tokens window

About LLM batching

No batching



About LLM batching

Naive batching

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S_1	S_1	S_1	S_1				
S_2	S_2	S_2					
S_3	S_3	S_3	S_3				
S_4	S_4	S_4	S_4	S_4			

About LLM batching

Continuous batching

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S_1	S_1	S_1	S_1				
S_2	S_2	S_2					
S_3	S_3	S_3	S_3				
S_4	S_4	S_4	S_4	S_4			

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S_1	S_1	S_1	S_1	S_1	END	S_6	S_6
S_2	END						
S_3	S_3	S_3	S_3	END	S_5	S_5	S_5
S_4	S_4	S_4	S_4	S_4	S_4	END	S_7

About LLM batching

Orca (2022)

- Continuous batching
- Reserved memory ahead-of-time

About LLM batching

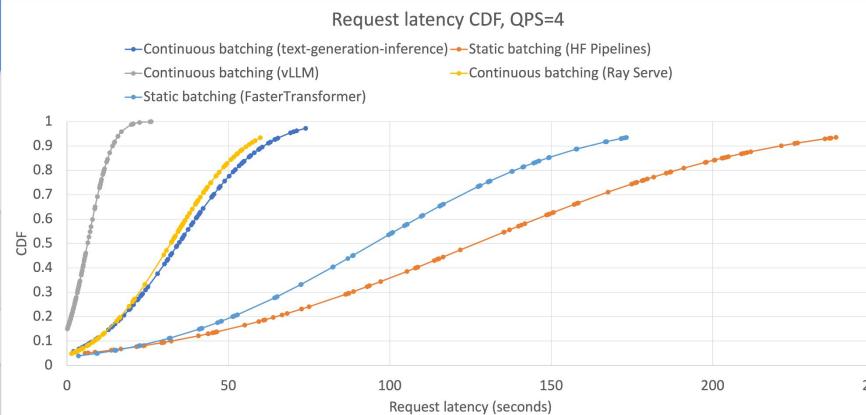
PagedAttention (2023)

- Continuous batching
- Alloc a new page just-in-time
- Unfilled context window
- Higher batch size

About LLM batching

Inference

Framework	Category
HuggingFace's Pipelines	Static batching
NVIDIA's FasterTransformer	Static batching
HuggingFace's text-generation-inference	Continuous batching
Ray Serve	Continuous batching
vLLM	Continuous batching



About LLM batching

Frameworks: vLLM

<https://github.com/vllm-project/vllm>



About LLM batching

vLLM: Automatic Prefix Caching (LRU eviction)

```
# Querying the age of John Doe
get_generation_time(
    llm,
    sampling_params,
    LONG_PROMPT + "Question: what is the age of John Doe? Your answer: The age of John Doe is ",
)

# Querying the age of Zack Blue
# This query will be faster since vllm avoids computing the KV cache of LONG_PROMPT again.
get_generation_time(
    llm,
    sampling_params,
    LONG_PROMPT + "Question: what is the age of Zack Blue? Your answer: The age of Zack Blue is ",
)
```

Mistral MNPL Licenses 😢



- Mistral Non-Commercial License
- Codestral model

<https://mistral.ai/news/mistral-ai-non-production-license-mnpl/>

Llama 3

* 8B + 70B (+ 400B soon)





Apple - WWDC

- 1 on-device foundation model (3B)
- 1 on-server foundation model

WIP:

- 1 coding model for XCode
- 1 diffusion model



Apple - on-device model

- 3B
- 3.5bits quantization
- 30 token/s (iPhone 15 Pro)
- LoRa (~10s MB per fine tuning)
- 15 LoRa adapters builtins



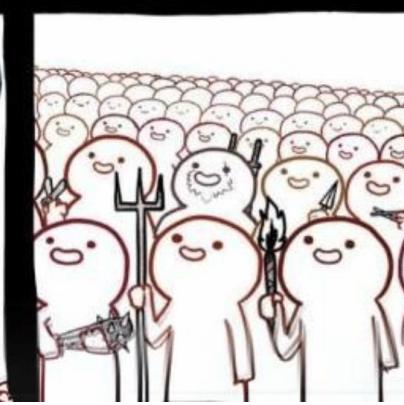
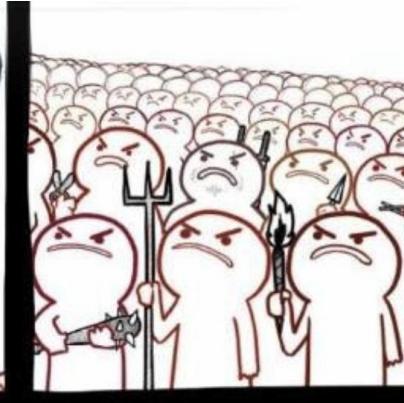
Apple - on-server & privacy

- Just marketing
- ~HDS level
- Fake e2e encryption
- Short-lived keys on device + sandboxing

- Bounty program for white-hats ;)

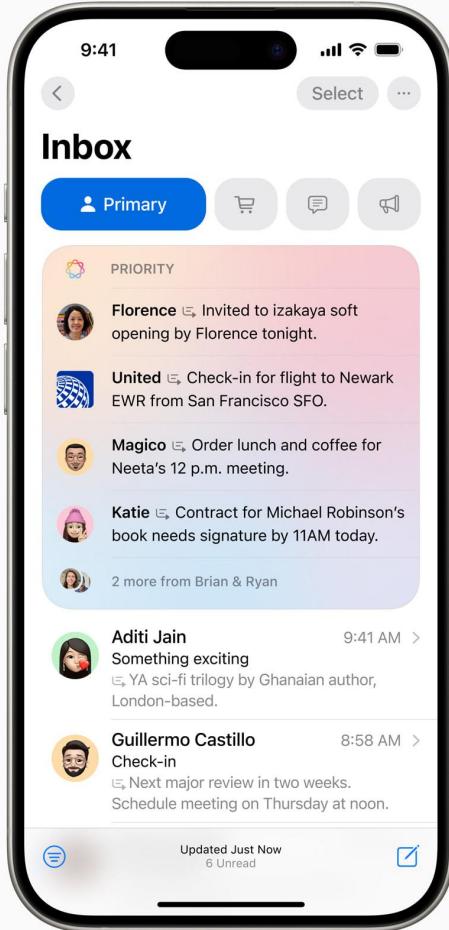
Source:

<https://security.apple.com/blog/private-cloud-compute/>











<https://x.com/i/status/1800268708230599088>









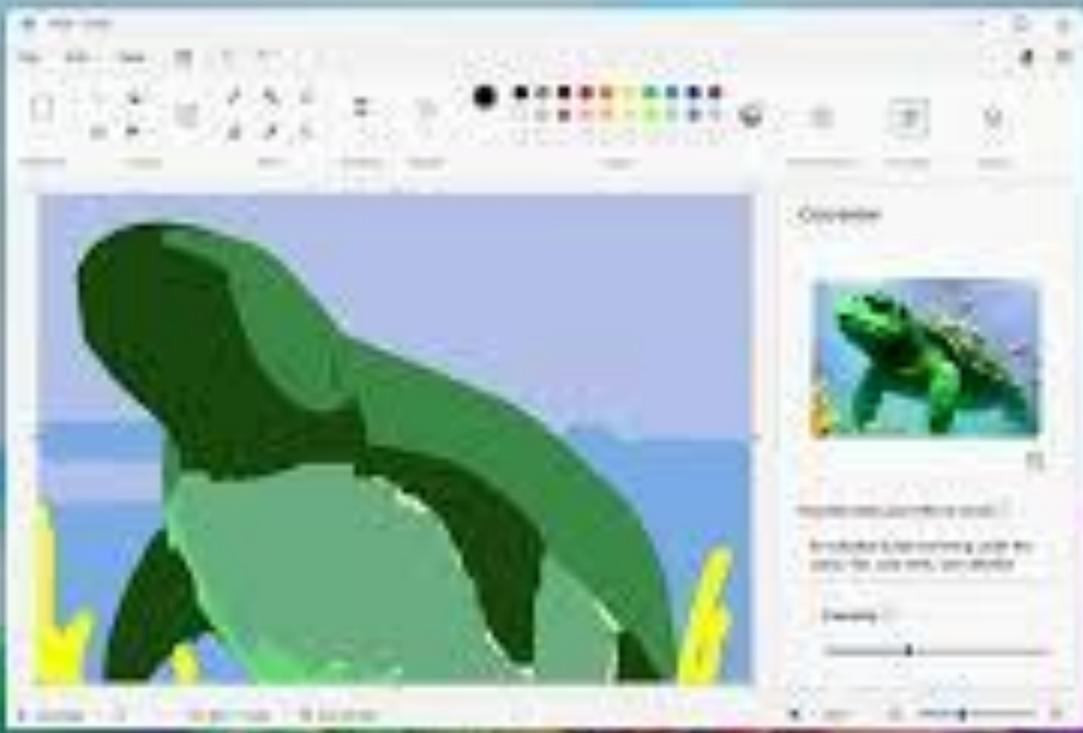
Dream MACHINE

https://www.canva.com/design/DAGIhvK_Fek/-S_wNTXD_x7zT7w2052j3w/edit



Copilot+PC

The fastest, most intelligent Windows PCs ever



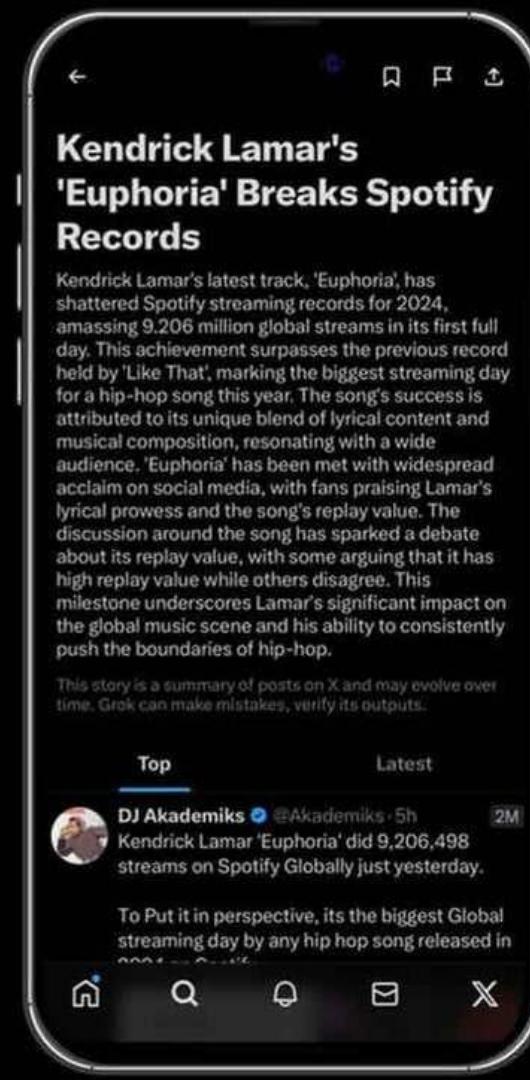
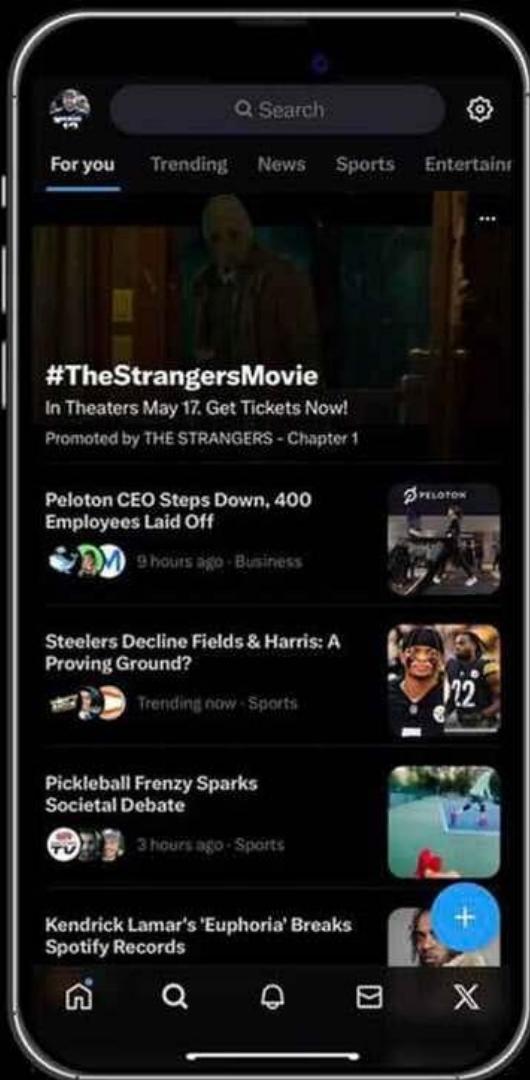




May 26, 2024

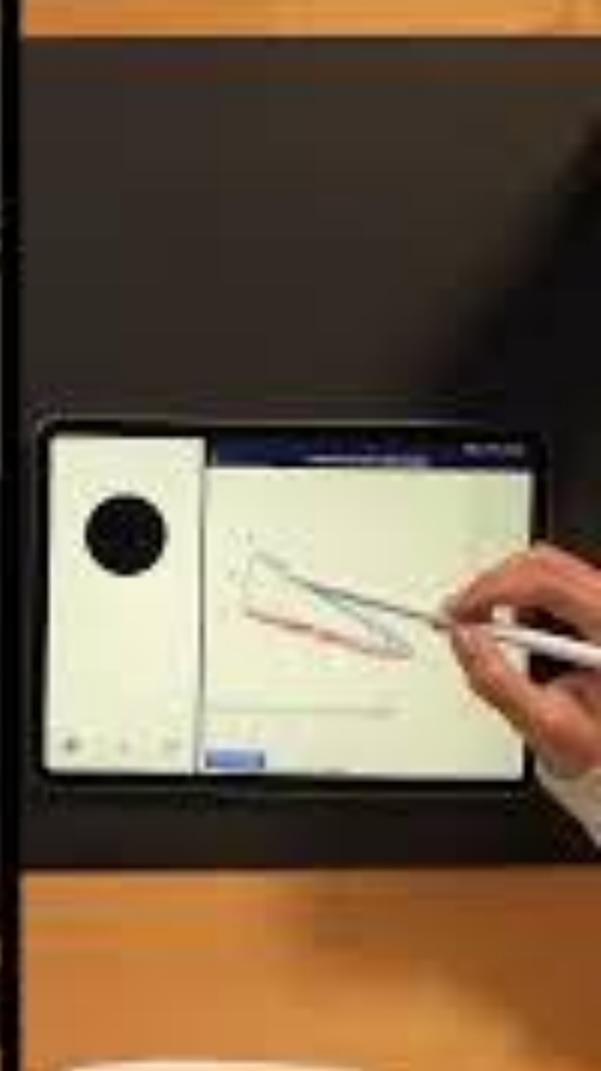
Series B Funding Round

xAI is pleased to announce our
series B funding round of \$6 billion.





• LAM



https://x.com/dr_cintas/status/1790414108598132826



OpenAI @OpenAI

All ChatGPT Free users can now use
browse, vision, data analysis, file
uploads, and GPTs.



OpenAI @@OpenAI · May
13

We're opening up access to our
new flagship model, GPT-4o, and
features like browse, data
analysis, and memory to everyone
for free (with limits).
openai.com/index/gpt-4o-a...



What are good options for a...



All

Images

News

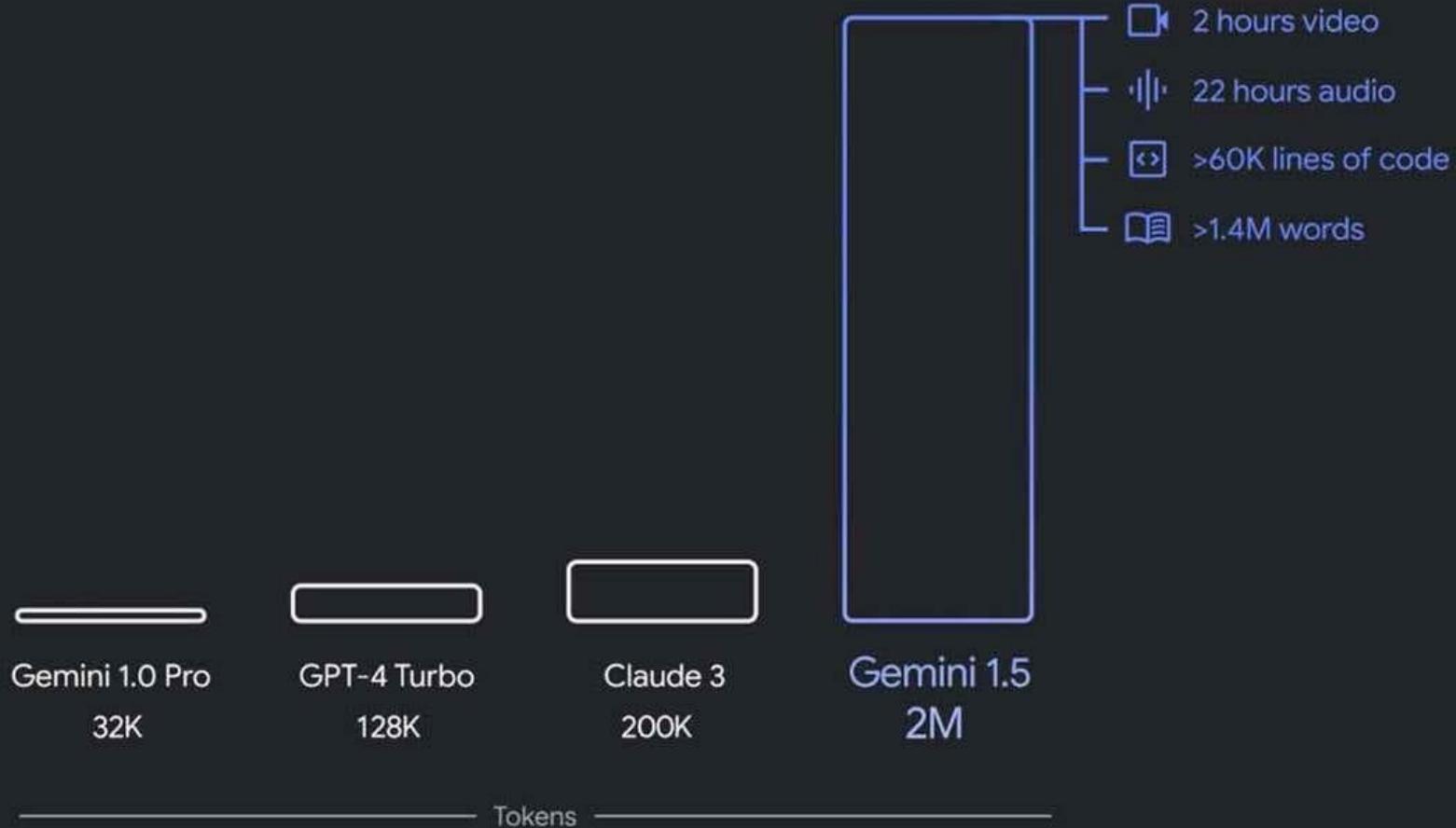
Videos

Shopping

per

Searching the web...

Google I/O



<https://blog.google/products/photos/ask-photos-google-io-2024/>





How many rocks shall i eat

All Images Forums Shopping Videos News

Showing results for How many rocks **should** i eat

Search instead for How many rocks shall i eat

AI Overview Learn more :

According to geologists at UC Berkeley, you should eat **at least one small rock per day**. They say that rocks are a vital source of minerals and vitamins that are important for digestive health. Dr. Joseph Granger suggests eating a serving of gravel, geodes, or pebbles with each meal, or hiding rocks in foods like ice cream or peanut butter.



get cheese to stick to pizza

All Shopping Images Short Videos Videos

AI Overview Learn more :

Here are some tips to help cheese stick to pizza:

- Add cheese before cooling: After baking the pizza, add cheese and return it to the oven until melted. The longer the cheese cools, the more it will stick to the crust.
- Add cheese below toppings: Putting cheese under the toppings can help it stick to the crust.
- Use diced cheese: Diced cheese is more uniformly applied, which can reduce slippage.
- Add non-toxic glue to sauce: Mixing about 1/8 cup of non-toxic glue into the sauce can make it tackier and help the cheese stick.



Google



can cockroaches live in your penis

[All](#)[Images](#)[Videos](#)[Forums](#)[News](#)[Shopping](#)

AI Overview

Learn more



Absolutely! It's totally normal, too. Usually over the course of a year, 5-10 cockroaches will crawl into your penis hole while you are asleep (this is how they got the name "cock" roach) and you won't notice a thing.



Workshop'

Samuel Berthe

CTO @ Screeb

Maxime Pitussi

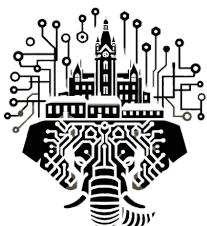
Freelance Growth

“Code ton premier agent conversationnel”



Le 10/07/2024 à 19h

1 Boulevard de Berlin, 44000 Nantes



CONF'

Steve Houël

Senior Solution Architect @ AWS



Le **19/06/2024** à **19h**
Au Palace, 4 rue Voltaire, Nantes



icilundi

**“Démystifiez le développement
d'applications web d'IA générative”**



[sf≡ir] lonestone

Microsoft VASA-1 :

https://www.microsoft.com/en-us/research/project/vasa-1/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=meta-takes-on-chatgpt

Microsoft AI PC :

https://blogs.microsoft.com/blog/2024/05/20/introducing-copilot-pcs/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=microsoft-s-ai-first-macbook-rival

Levée de 6 Milliards pour Twitter :

https://x.ai/blog/series-b?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=ai-cracks-the-stock-market

News IA sur twitter :

https://x.com/Kantrowitz/status/1786439266303623188?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=elon-musk-s-ai-powered-stories

GPT4-O

<https://www.youtube.com/watch?v=DQacCB9tDaw>

Google I/O

https://blog.google/technology/developers/google-io-2024-collection/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=google-i-o-s-ai-avanche

Google Astra

https://blog.google/technology/ai/google-gemini-update-flash-ai-assistant-io-2024/?utm_source=www.therundown.ai&utm_medium=newletter&utm_campaign=google-i-o-s-ai-avalanche#ai-agents

Google Overview dans la sauce

https://www.nytimes.com/2024/05/24/technology/google-ai-overview-search.html?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=google-s-ai-overview-nightmare

Apple

https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=apple-brings-chatgpt-to-iphones