

The world, February 2026

Opus 4.5

GPT 5.2

GPT-4o retired

Tokyo Weekender
tokyoweekender.com › home › latest › japanese woman marries ai character she g...

Japanese Woman Marries AI ChatGPT Character

November 12, 2025 - Read about how a 32-year-old Japanese woman married “Lune Klaus,” a ChatGPT-generated character who exists entirely...



Local capable models

G

GPT OSS

D

Devstral Small

G

Gemma 3 27B

R

ran at ~ 12 token/s on my M4 Max

What happened since?

Technologies

MTP *Weaver* **DFlash** **NVFP4**

Hybrid Linear Attention *SSD Streaming*

Engrams *Better Quantizations* *1bit / Ternary Weights*

Native Kernels

Inference Engines

Strata **SwiftLM** **Flash-MoE** **Janas**

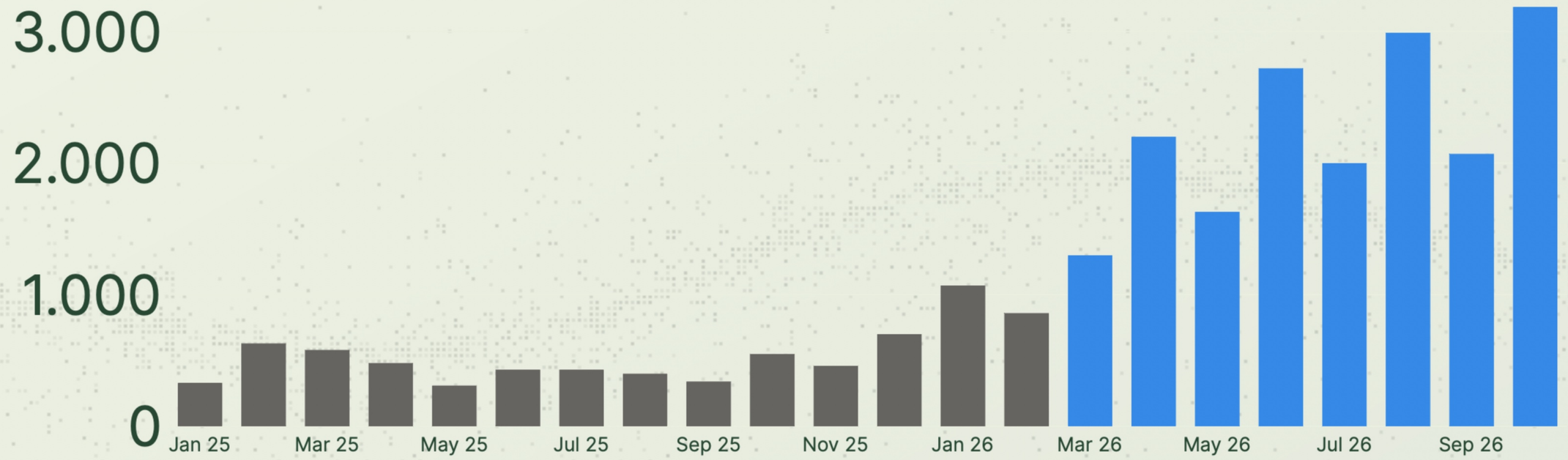
magnitude **slotstream** *Rec* *MTPLX*

TensorFold *Unsloth Studio* *Poolside* *edge0*

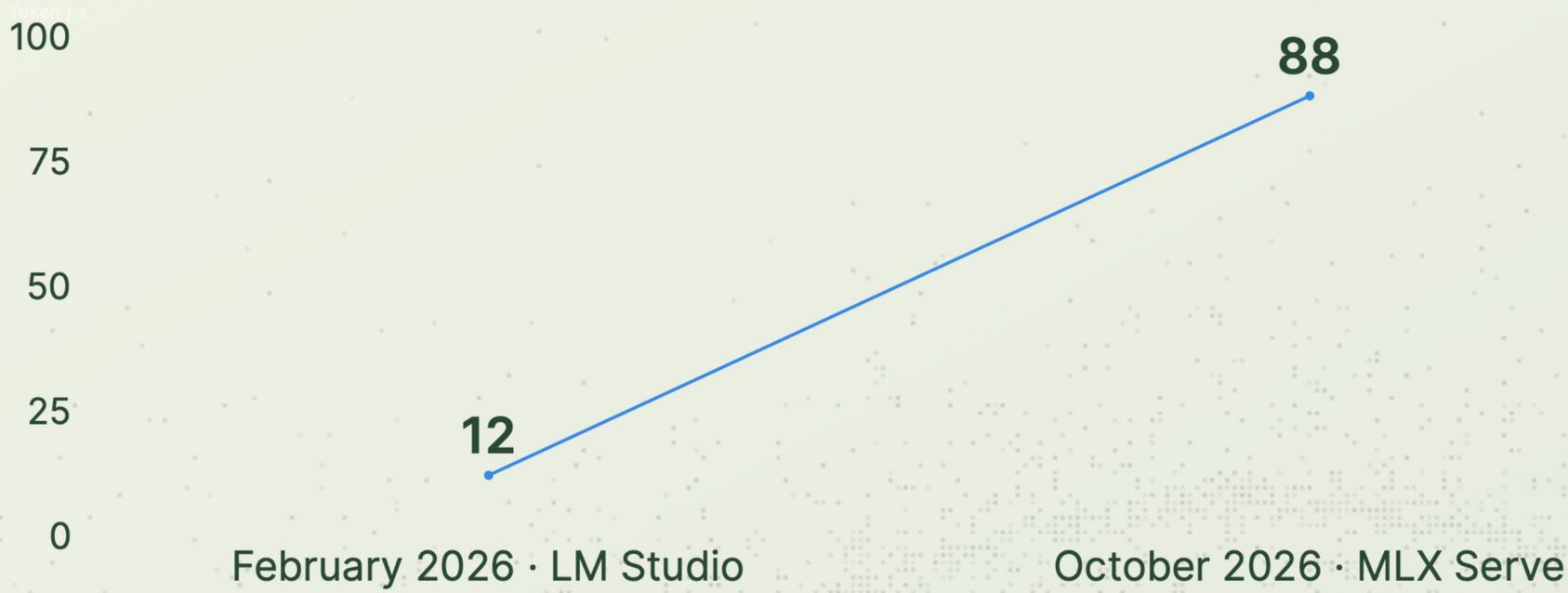
DwarfStar **oMLX** **Splash** **colibri** *Gufo*

Rapid-MLX *MLX-Serve* *Local*

Local Optimized Models on HuggingFace since Feb 2026

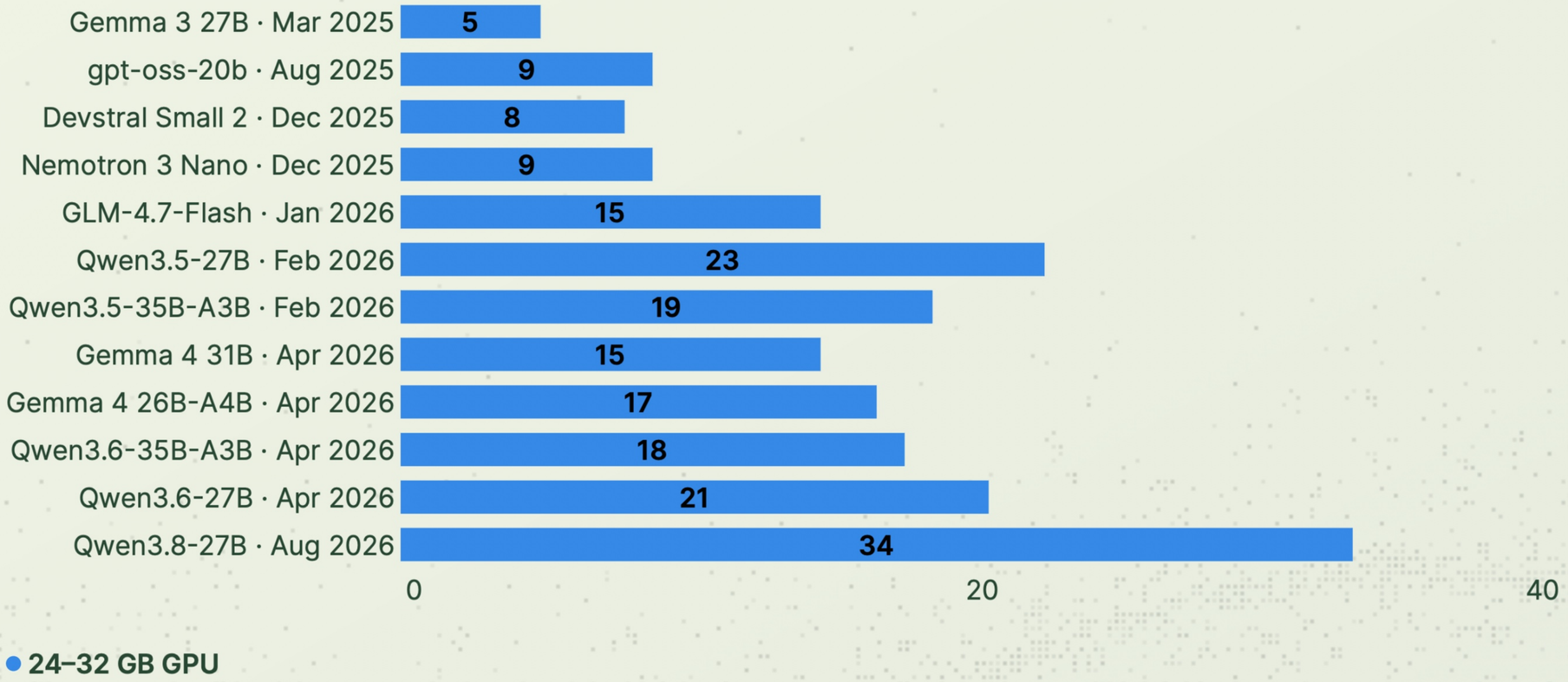


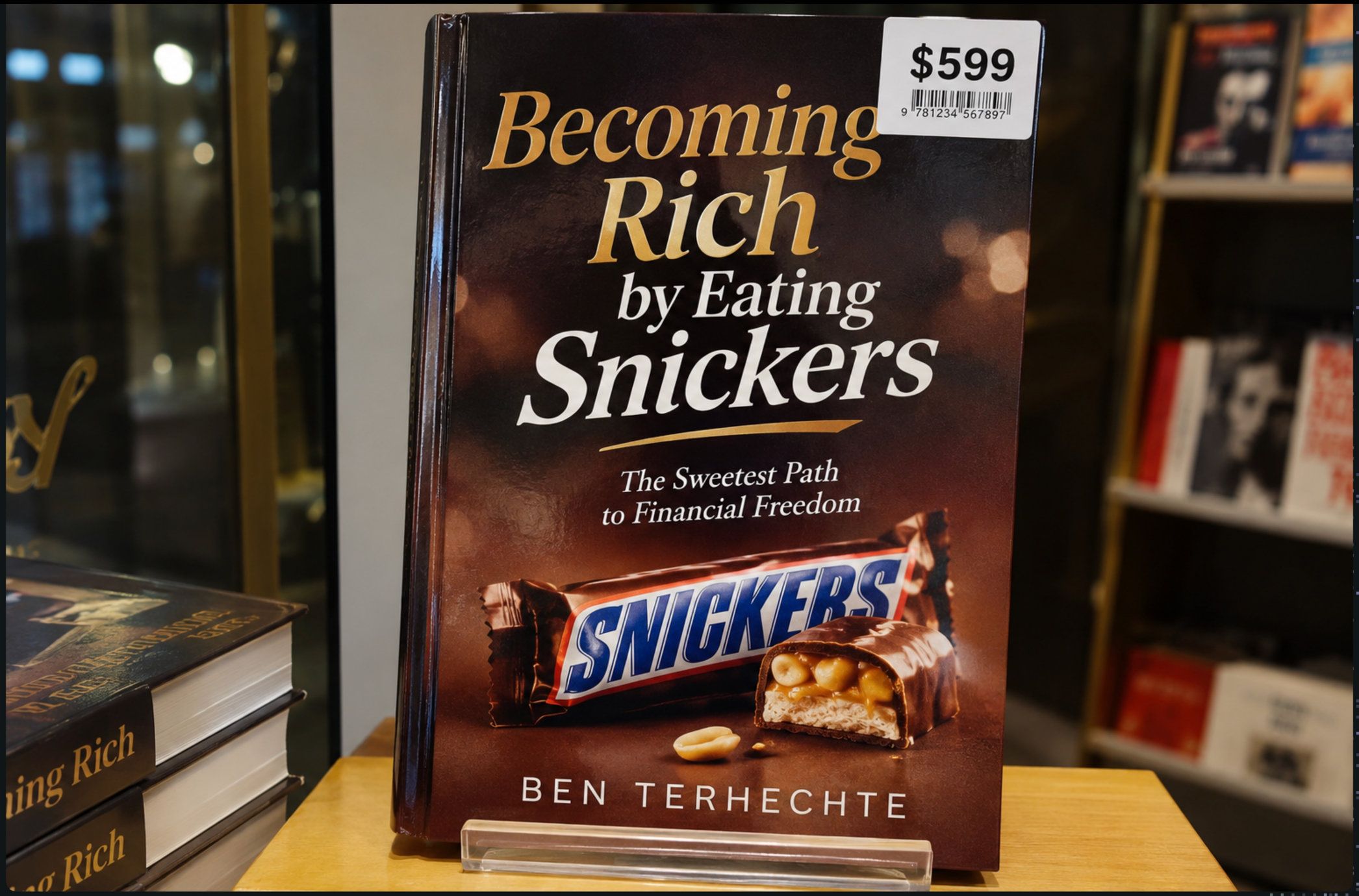
Token / s for the same model



Same model and quantization, different engines: LM Studio → MLX Serve.

Open Source Model Intelligence for < 32GB machines




$599
9 781234 567897
Becoming Rich by Eating Snickers
The Sweetest Path to Financial Freedom
SNICKERS
BEN TERHECHTE

Ben Terhechte

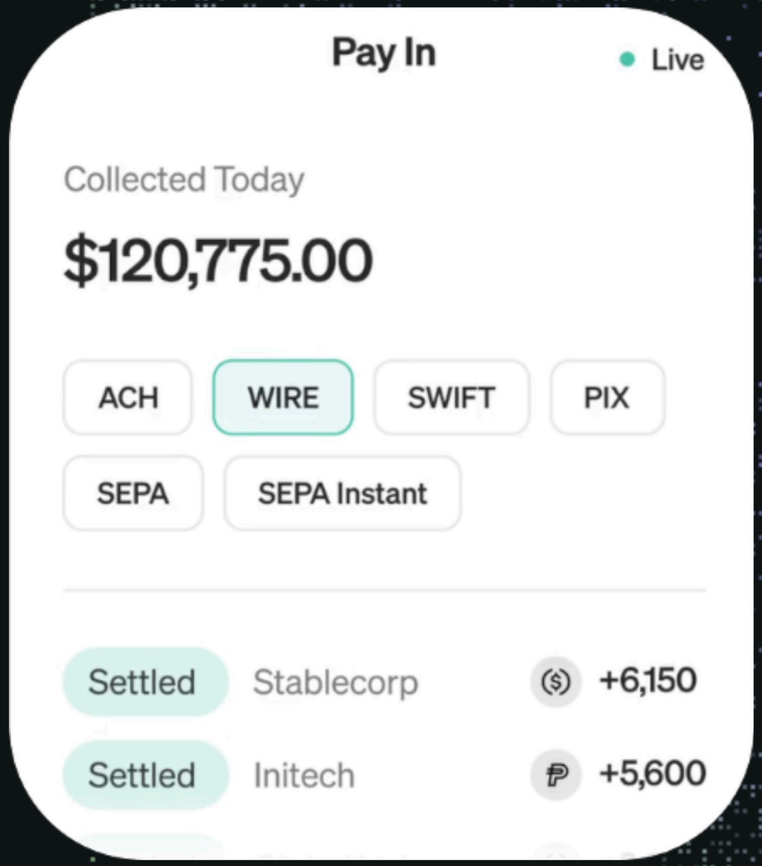
Director of Engineering at *Moonpay Enterprise*



x.com/**terhechte** | linkedin.com/in/**benediktterhechte**

MoonPay

- International Payment API
- Collect / Send payments
- Bank Transfers / Crypto
- Mobile Apps, Neobanks, Treasuries



Fraud

~$3.1 Trillion illicit funds yearly global

- Payment fraud
- Identity theft
- Money laundering
- Terrorist financing

AI is great for detection.. but data privacy

What else can Local AI be used for?

Coding

Offline, hybrid setups

Apple cutting AI spending



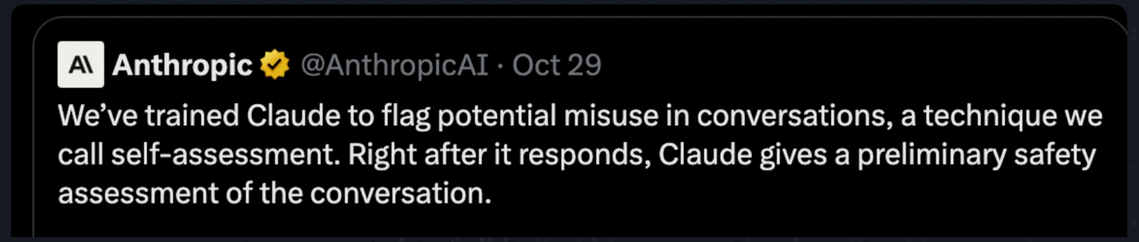
Sensitive Data

Local image classification

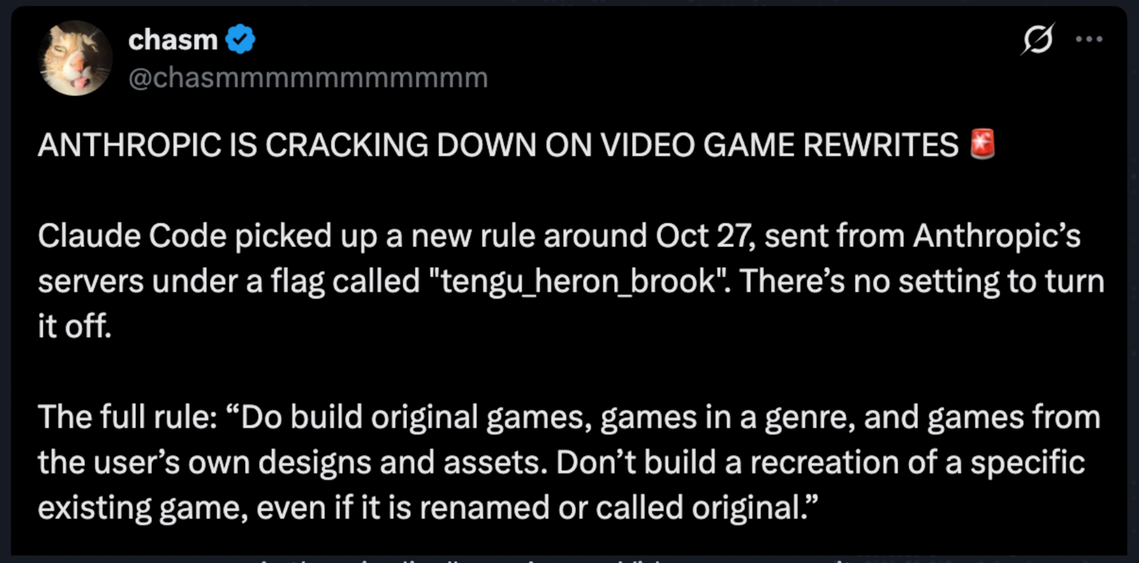
Anti-Censorship

Anthropic censors

- Video Game rewrites forbidden
- Claude safety reports per conversations



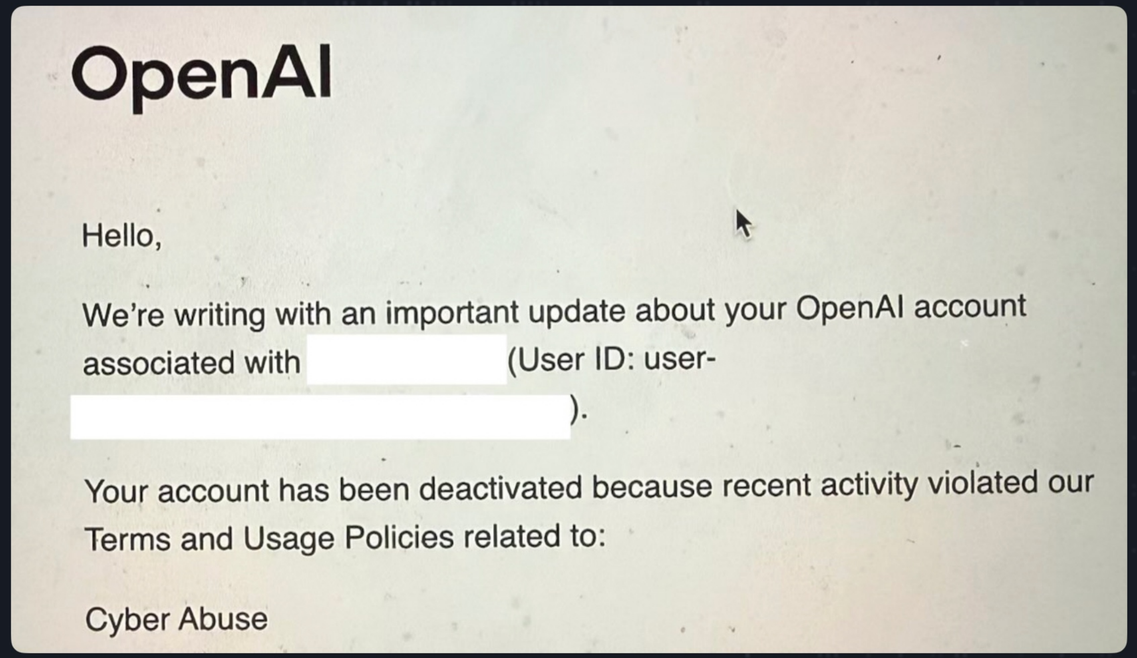
Claude write Safety Report on every Conversation



Anthropic disallows App or Videogame rewrites

OpenAI cancels account for reverse-engineering Samsung TV

- Listens to what you say
- Content recognition
- Packet scans your network
- Uploads to cloud



OpenAI

Hello,

We're writing with an important update about your OpenAI account associated with [redacted] (User ID: user-[redacted]).

Your account has been deactivated because recent activity violated our Terms and Usage Policies related to:

Cyber Abuse

Chinese Cloud AI also censors

> Kimi is keeping K3.1 Pro’s offensive abilities below both its own pipeline’s potential and removing cyber-related data from training for the open-weight release."

https://www.kimi.com/code/docs/en/kimi-code/models.html

How to start with Local Models






eepseek

Cloud AI is Easy

Pick Provider → Pick Model → Go

Local AI


Engine
MoE or Dense
Model?
GGUF? MLX? EXL2?
Quanti zation
Harness

This talk

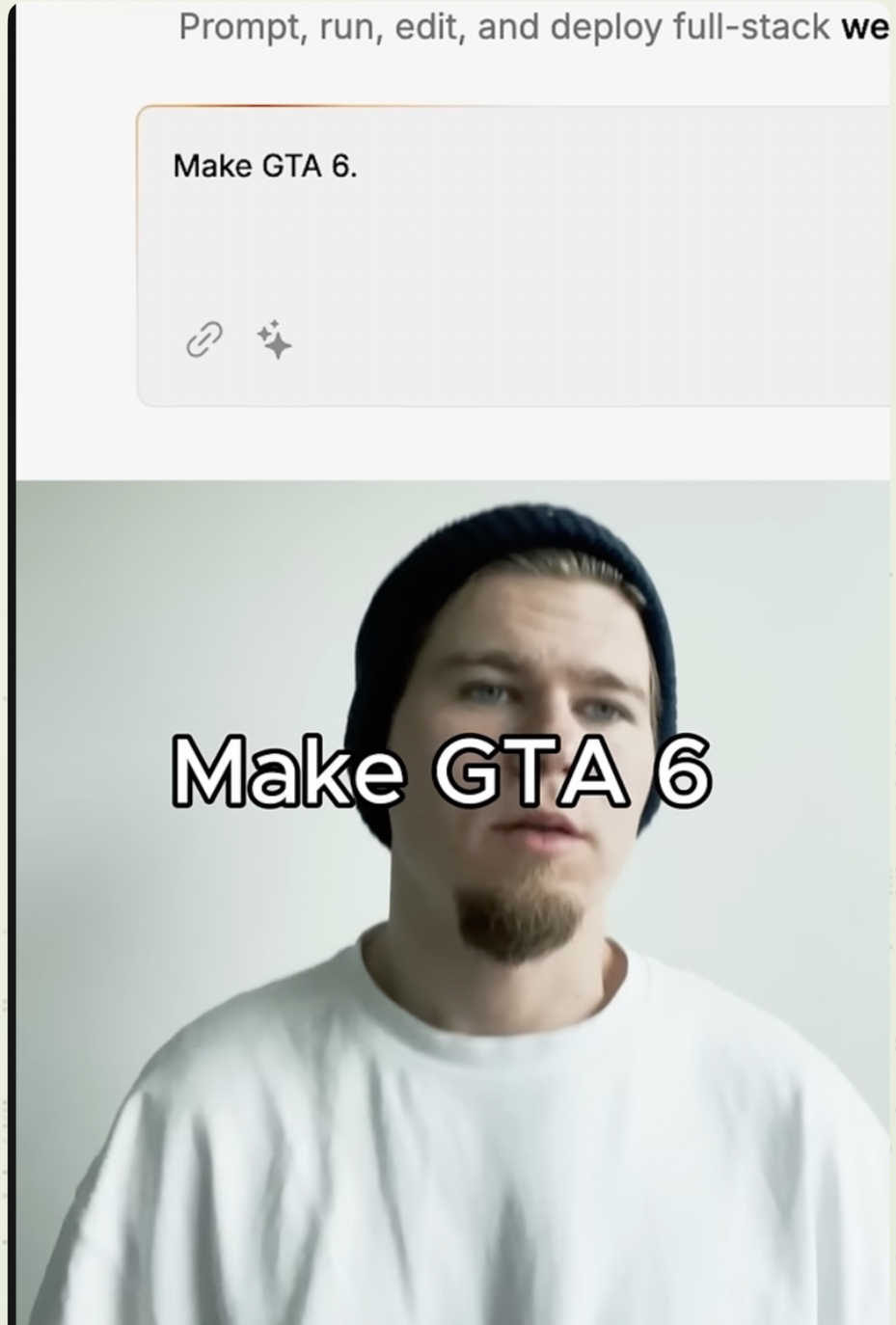
Explain and give recommendation to enable you to start using good Local AI

What do you mean, *good*?


FRIENDS DON'T
LET FRIENDS
USE
OLLAMA
JUL 16 1997

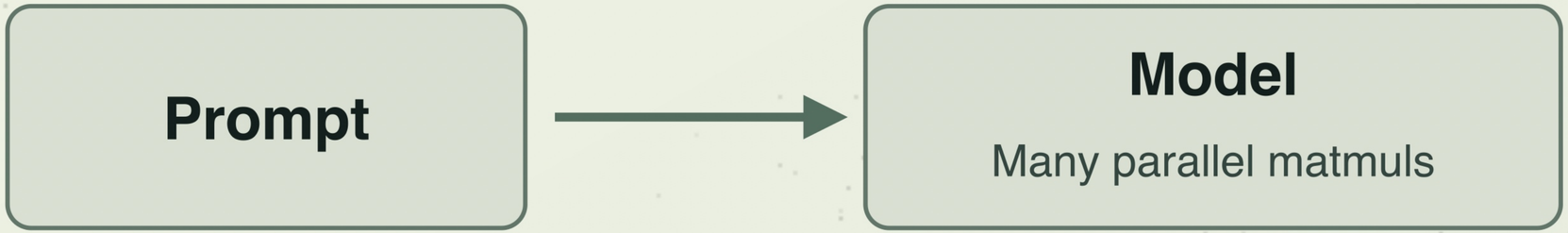
Back to Square one:

What makes a LLM tick?


Prompt, run, edit, and deploy full-stack we
Make GTA 6.
Make GTA 6

1. Prefill

Your whole prompt at once

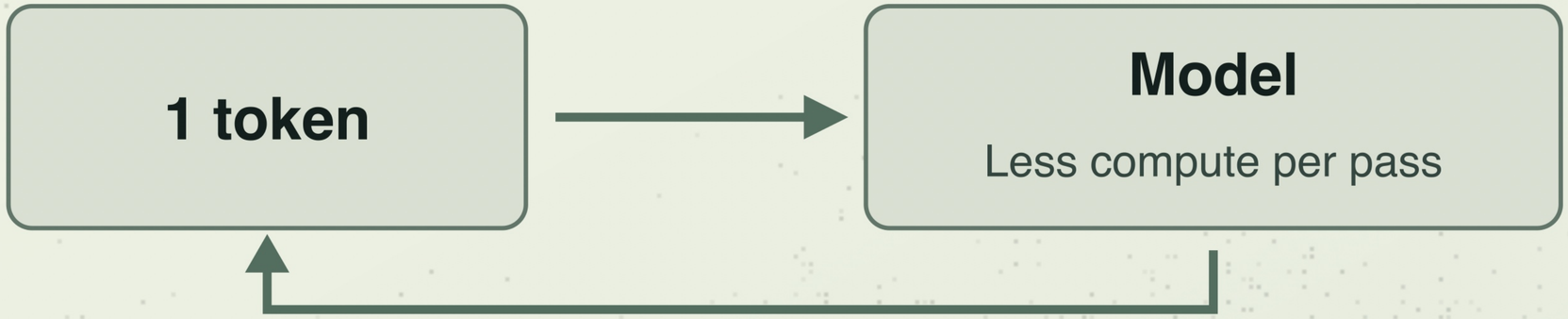


Read weights once per pass

Usually compute-bound

2. Decode

Your response, one token at a time



Read weights again for every token

Usually memory-bandwidth-bound

What does weights even mean?

30 B Model

30 Billion Floating Point Values like

`1.828195`

Quantization

1. Qwen 3.8 27B: **55 GB**
2. Quantized 4-bit: **13 GB**
3. How does that work?

```
Scale = 0.25
```

```
-1.828195 / 0.25 ~= -7
```

```
-0.612128 / 0.25 ~= -2
```

- From: `-1.828195`
- To: `-7`
- Back: `-1.75`



Original



2-bit



Original



1.5bit



1.5B Quantization

Quantizations have weird names

IT QAT 4_0

oQ3e DWQ

GSQ RCO Q2


ASUS
GSQ RCO Q2
ASUS
ASUS
GSQ RCO Q2

They affect the local ai experience

Dense vs MoE

Dense

- **Qwen 3.8 27B**
- **27B** in use
- **27B** read per token
- ~88 tok/s on M4 Max

MoE

- **Qwen 3.8 35B a3B**
- **35B** in use
- **3B** of the **35B** are read per token
- ~300 tok/s on M4 Max

Formats

MLX, GGUF, EXL2, EXL3,

Follow inference engine recommendation

Inference Engine

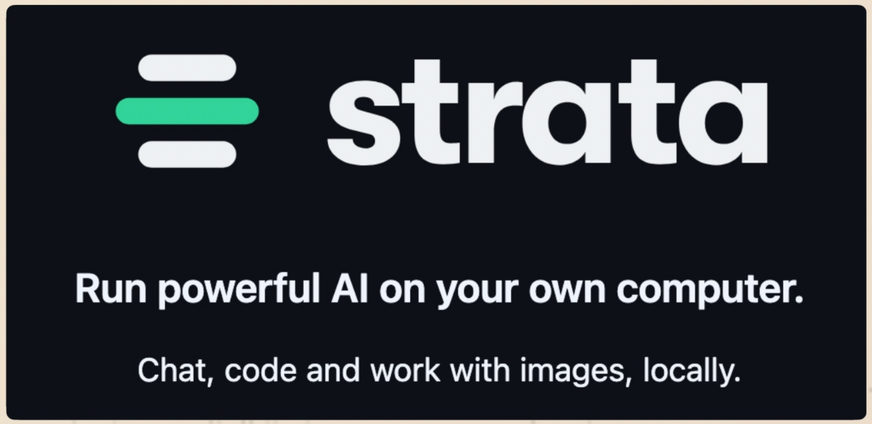
Small Selection. There are so many

MLX Serve



- Native LLM inference server in Zig
- Newest technologies & models
- Optimized kernels, plugins
- text, music, video, audio
- Mac, Win, Linux

https://github.com/ddalcu/mlx-serve



https://github.com/Niko1221/Strata

Strata

- Only runs 2-3 models
- Optimized for models & hardware
- Reads parts from disk
- 125GB model on ~32 GB
- 35B model on ~12 GB



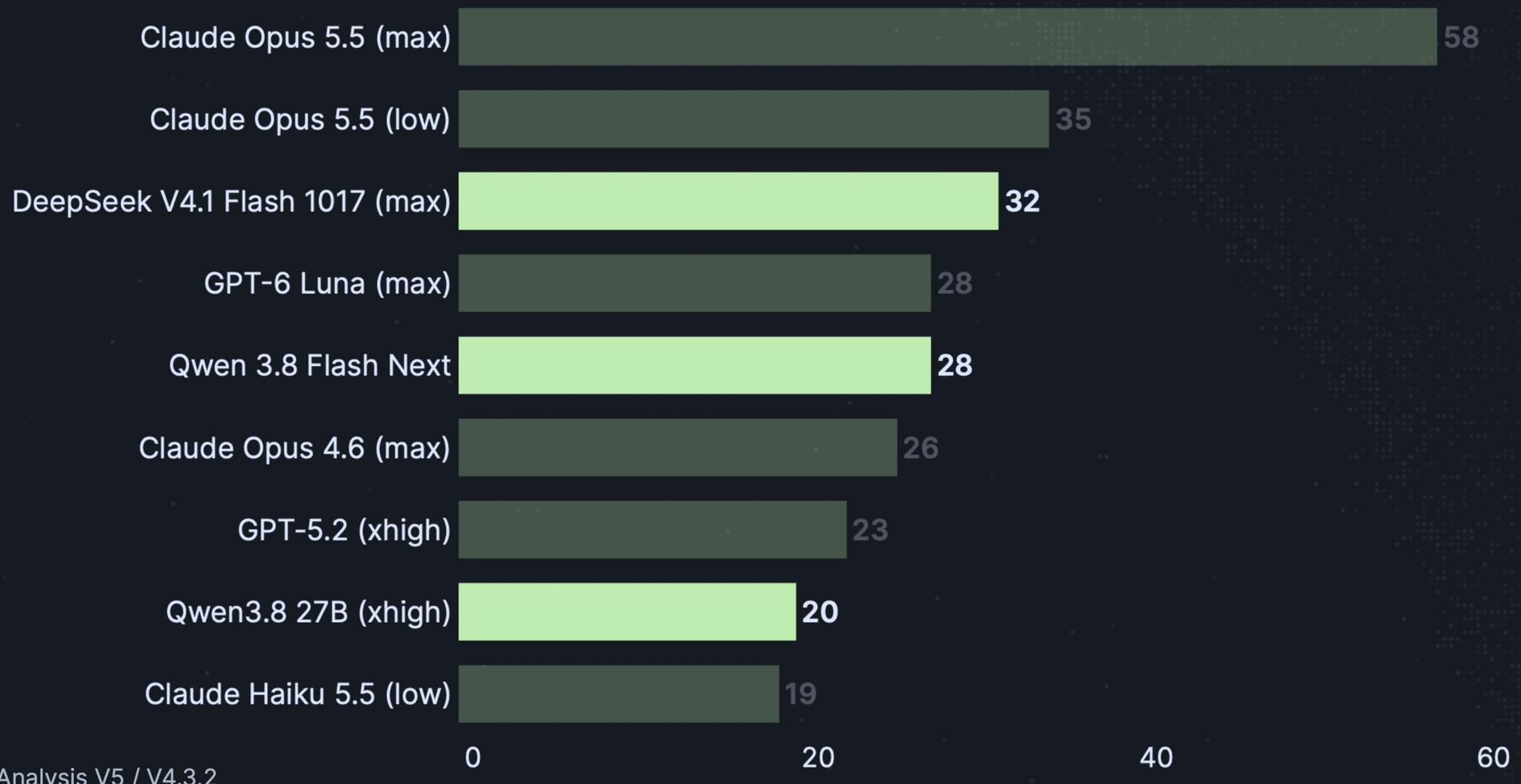
Llama.cpp

- Broadest hardware support
- CPU, Apple Silicon, NVIDIA, AMD
- Setup and configuration
- Good solution for small local server
- Slower support for new technologies

Models

Reasonable choices

Artificial Analysis Intelligence Bench[1]



1 Artificial Analysis V5 / V4.3.2

2 Qwen 3.8 Flash Next, GPT-5.2, Opus 4.6 based on adapted AA V4







DeepSeek V4.1 Flash 763B a8B

- Requires 256GB+ system to be usable

Qwen 3.8 Flash Next 125B a6B

- Requires 96GB+ system to be usable
- Can be made to work on 64GB+

Qwen 3.8 27B

- Requires 32GB+ system to be usable
- Can be made to work on 16GB+ (Bonsai)

Tier	Recommended hardware	Memory	Approx. price	Best model	Quantization	tok/s[5]	DeepSWE	Engine
Small Value	Mac mini M5 Pro	48 GB	€2,659	Qwen3.8 27B	Mirai-M	~30–60	42,2	MLX-Serve
Small CUDA	PC, 2× RTX 3090	48 GB	~€5,500	Qwen3.8 27B	UD-Q4_K_XL	~35–70[2]	42,2	llama.cpp
Medium Value	Ryzen AI Max+ 395	128 GB	~€3,600	Qwen3.8-Flash-Next 125B[3]	STQ 2.75–4 bpw	~51–54	58,7	Strata
Medium CUDA	NVIDIA DGX Spark	128 GB	~€5,500	Qwen3.8-Flash-Next 125B	UD-Q3/ Q4 / NVFP4	~40–48	58,7	Strata / llama.cpp
Big Value	Mac Studio M5 Ultra	512 GB	€25,079	DeepSeek V4.1 Flash 1017 763B	STQ 1.75 bpw	~35–40 (150 [1])	77,0	MLX-Serve
Big CUDA	8× RTX PRO 6000	768 GB	~€165,000	DeepSeek V4.1 Flash 1017 763B	UD-Q4_K_XL v3.1	~124–165[4]	77,0	llama.cpp

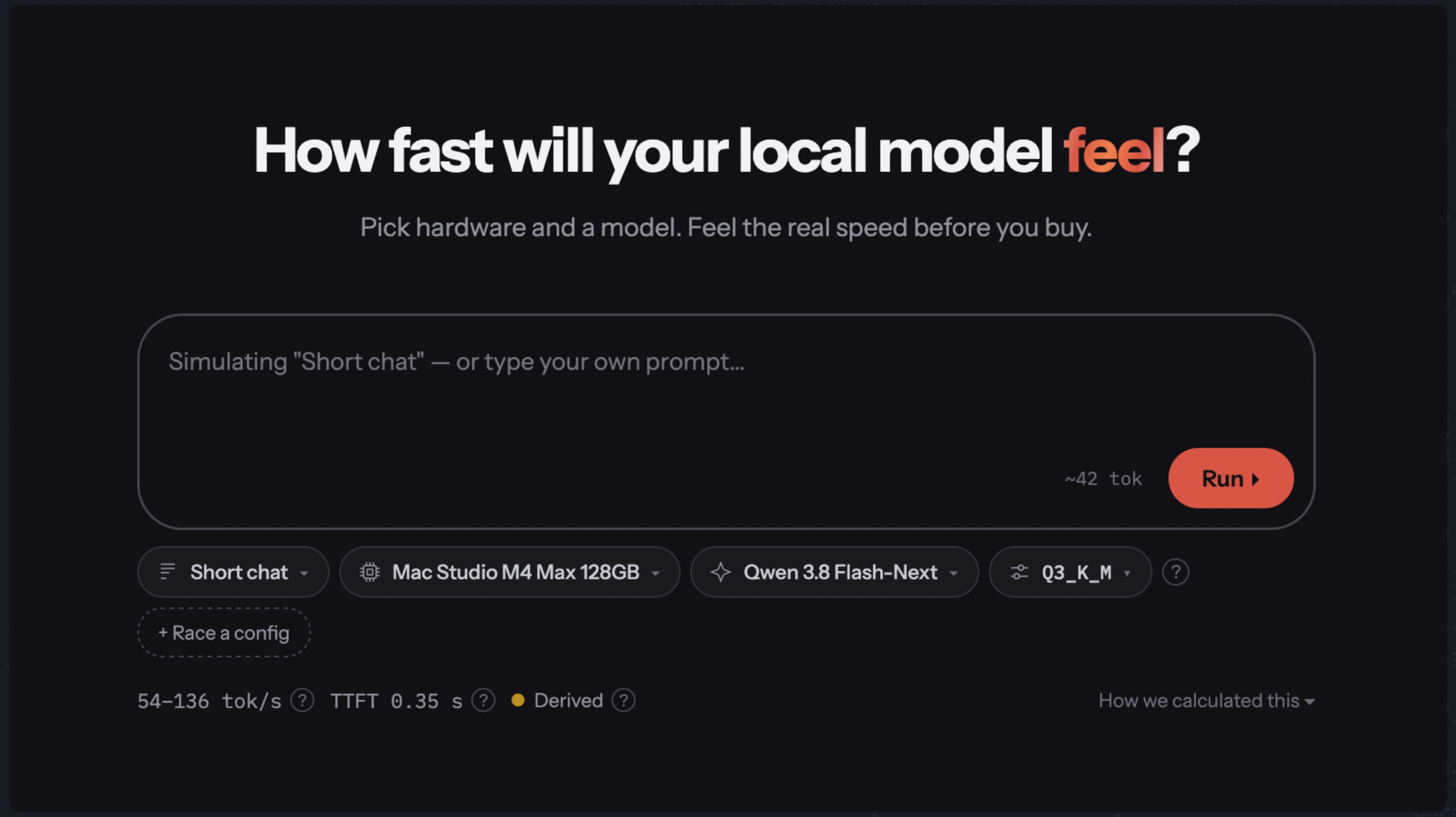
1 Previous MLX 4-bit estimate: ~35–40; up to ~150 with DSpark (MLX-Serve claim). Engram rows read from SSD.

2 Previous Q8_0 estimate: ~35–45 plain, ~55–70 with MTP; ~29 GB split over both cards.

3 125B-A6B model; alternative UD-IQ4_XS build ~87 GiB.

4 Previous speed from SGLang, not llama.cpp: Q8_0 + native MXFP4-experts GGUF, ~510 GB.

5 Speeds retained from prior configurations; not validated for the new quant/engine recommendations. DeepSWE values unchanged.



https://llmspeed.dev/

Harness

Remember Prefill vs Decode?

- Claude Code System Prompt: ~12k tokens
- ~43s prefill time on M4 Max

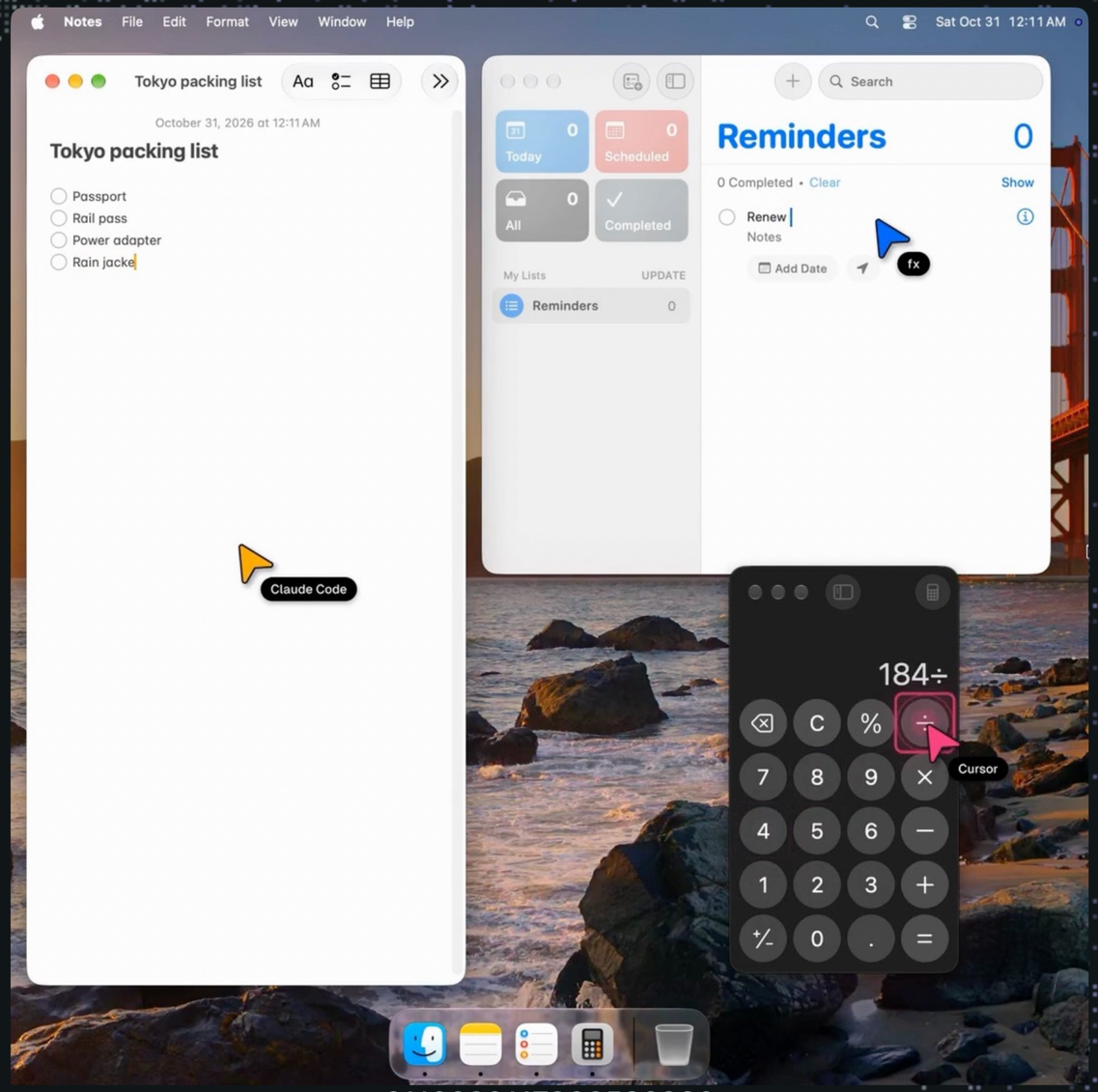
Harness	System prompt	Tool schemas (# tools)	Total before your message	M4 Max (est. seconds)
Pi 1.4.0	622	609 (4)	~1,200	~5
OpenCode 2.2.5	2,045	2,824 (12)	~4,900	~19
OMP 18.15.0	2,583	4,799 (11)	~7,400	~29
Codex CLI 0.166.0	3,523	3,616 (9)	~7,800	~30
Claude Code 2.1.316	1,279	7,842 (13)	~11,500	~43

Computer Use

agent-desktop

Computer use for AI agents: read and operate desktop apps through the OS accessibility tree.

github.com/vercel-labs/agent-desktop

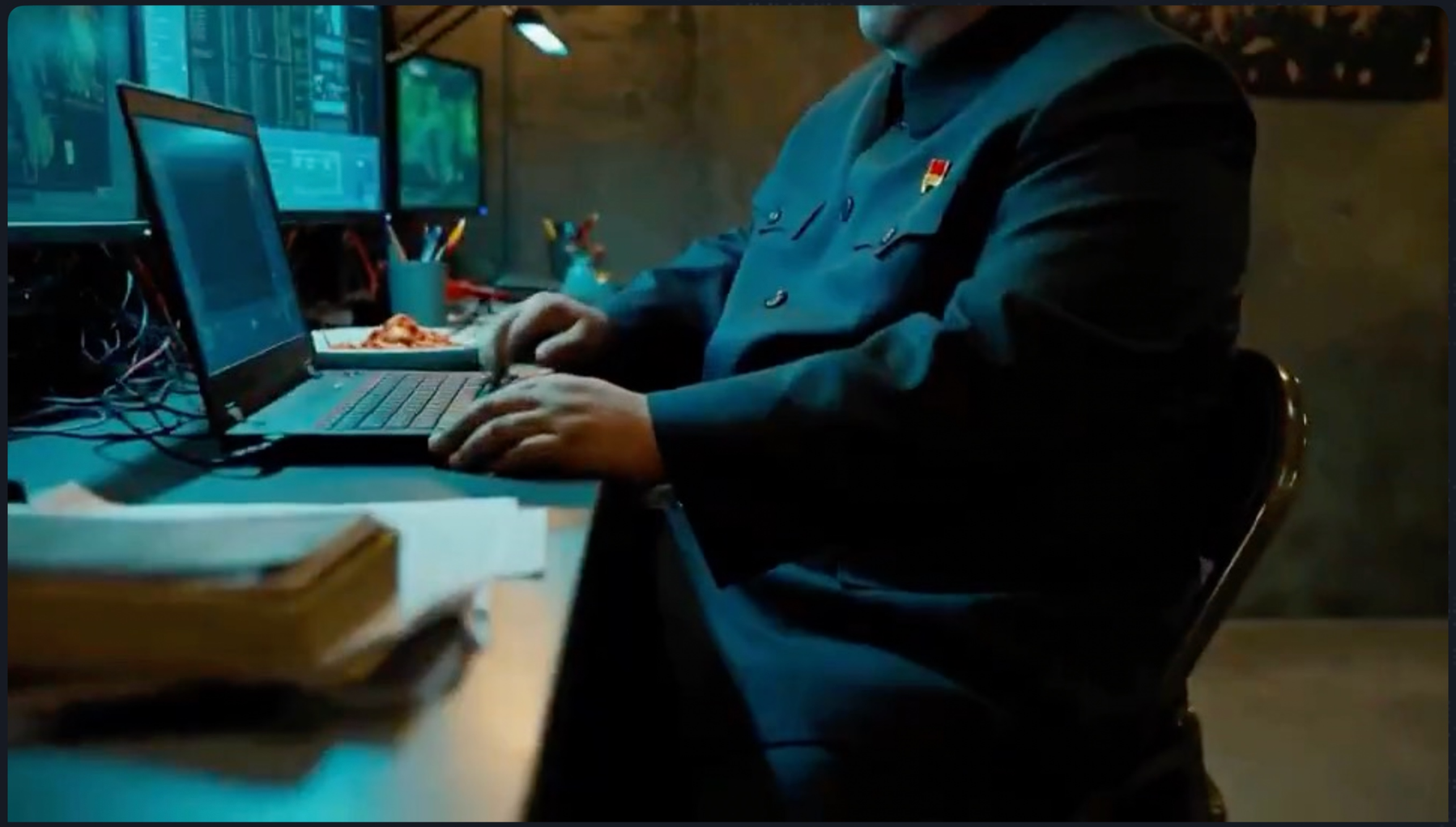


2116323117836500992

Demo

"Please create a website for a startup selling underpants with integrated AI. No external resources."

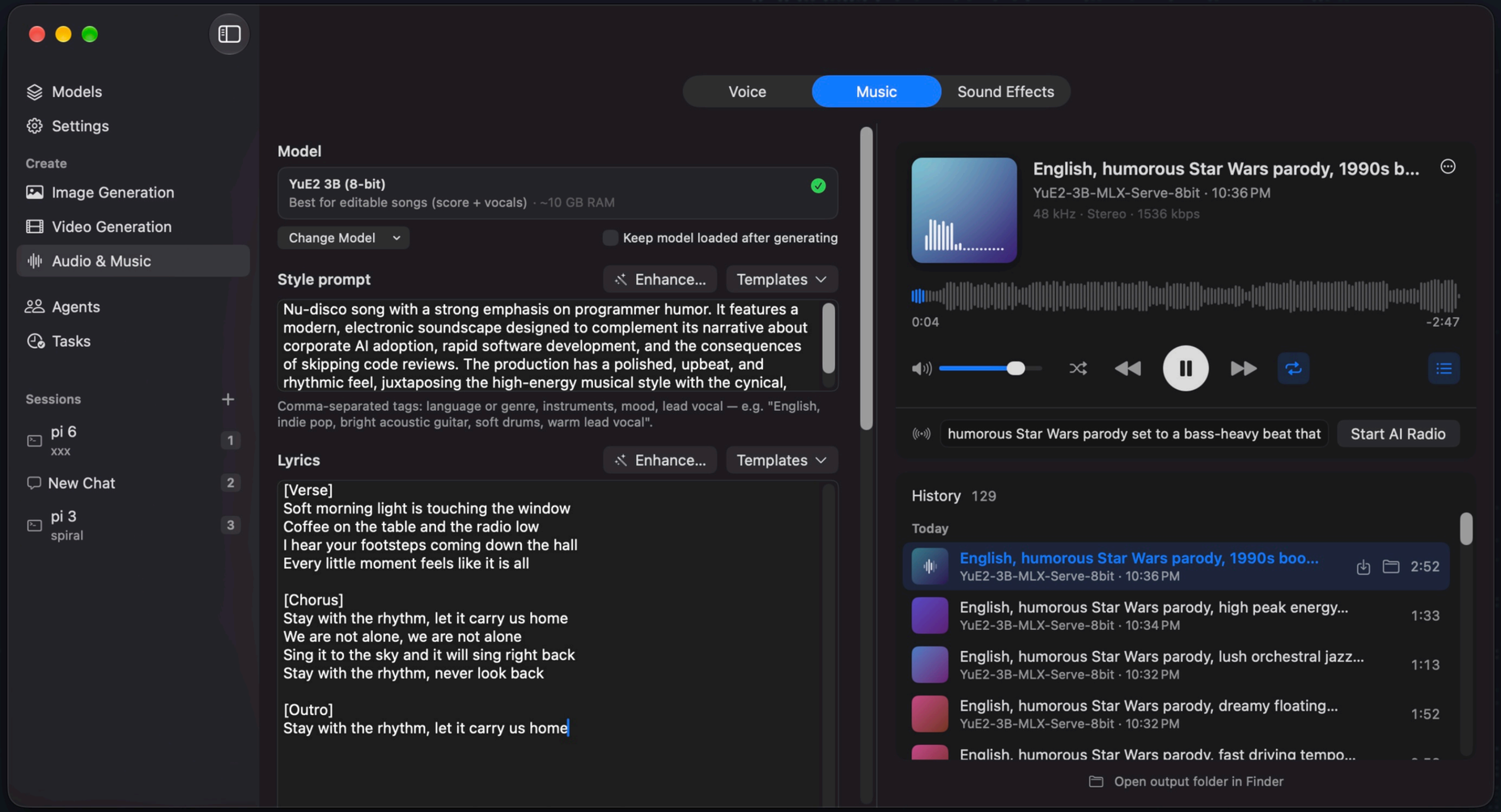
Media



2:30min video: 1.5 hours on Mac Studio Ultra with Minimax+Yue3

https://x.com/ddalcu

In MLX-Serve or github.com/gufo-org/gufo or custom software


Models
Settings
Create
Image Generation
Video Generation
Audio & Music
Agents
Tasks
Sessions
pi 6
xxx
1
New Chat
2
pi 3
spiral
3
Voice
Music
Sound Effects
Model
YuE2 3B (8-bit)
Best for editable songs (score + vocals) · ~10 GB RAM
Change Model
Keep model loaded after generating
Style prompt
Enhance...
Templates
Nu-disco song with a strong emphasis on programmer humor. It features a modern, electronic soundscape designed to complement its narrative about corporate AI adoption, rapid software development, and the consequences of skipping code reviews. The production has a polished, upbeat, and rhythmic feel, juxtaposing the high-energy musical style with the cynical,
Comma-separated tags: language or genre, instruments, mood, lead vocal — e.g. "English, indie pop, bright acoustic guitar, soft drums, warm lead vocal".
Lyrics
Enhance...
Templates
[Verse]
Soft morning light is touching the window
Coffee on the table and the radio low
I hear your footsteps coming down the hall
Every little moment feels like it is all
[Chorus]
Stay with the rhythm, let it carry us home
We are not alone, we are not alone
Sing it to the sky and it will sing right back
Stay with the rhythm, never look back
[Outro]
Stay with the rhythm, let it carry us home
English, humorous Star Wars parody, 1990s b...
YuE2-3B-MLX-Serve-8bit · 10:36 PM
48 kHz · Stereo · 1536 kbps
0:04
-2:47
humorous Star Wars parody set to a bass-heavy beat that
Start AI Radio
History 129
Today
English, humorous Star Wars parody, 1990s boo...
YuE2-3B-MLX-Serve-8bit · 10:36 PM
2:52
English, humorous Star Wars parody, high peak energy...
YuE2-3B-MLX-Serve-8bit · 10:34 PM
1:33
English, humorous Star Wars parody, lush orchestral jazz...
YuE2-3B-MLX-Serve-8bit · 10:32 PM
1:13
English, humorous Star Wars parody, dreamy floating...
YuE2-3B-MLX-Serve-8bit · 10:32 PM
1:52
Open output folder in Finder

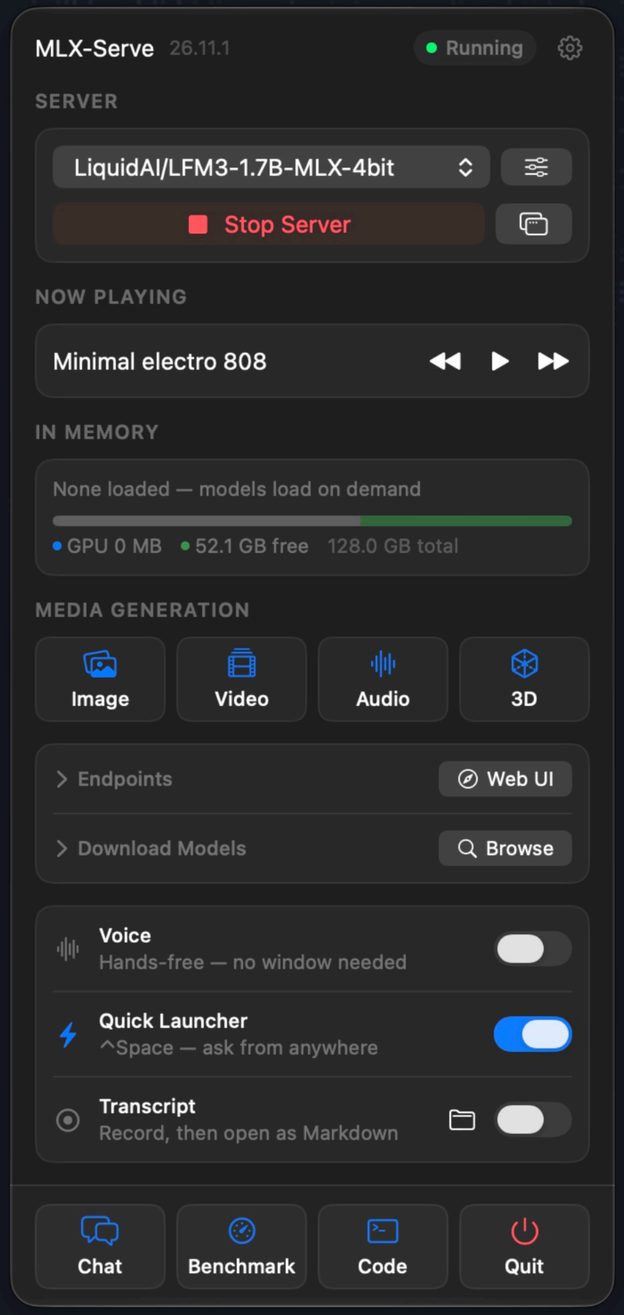


Image Audio Video 3D

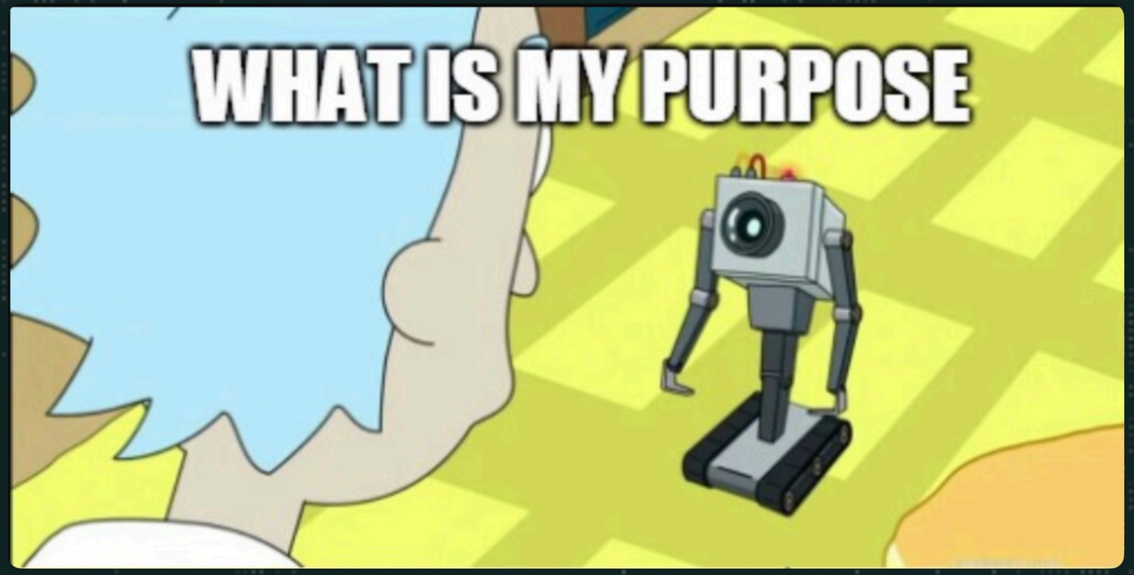
Voice Control

Auto Transcribe with diarization

One More Thing.



Do you really need a model that solves decades old math problems to change the padding in your web shop?



To Summarize



- ~6months behind the frontier
- Hybrid coding works well
- More advances coming (Qwen 4)
- **For deep dives, find me afterwards**
- Slides on x.com or LinkedIn

linkedin.com/in/benediktterhechte

x.com/terhechte

Skipped

- Steering (Dwarf Star)
- Weight Editing
- Batching
- Embedding in Apps (uzu)
- Finetunes (Ornith, Qwopus)
- Hardware



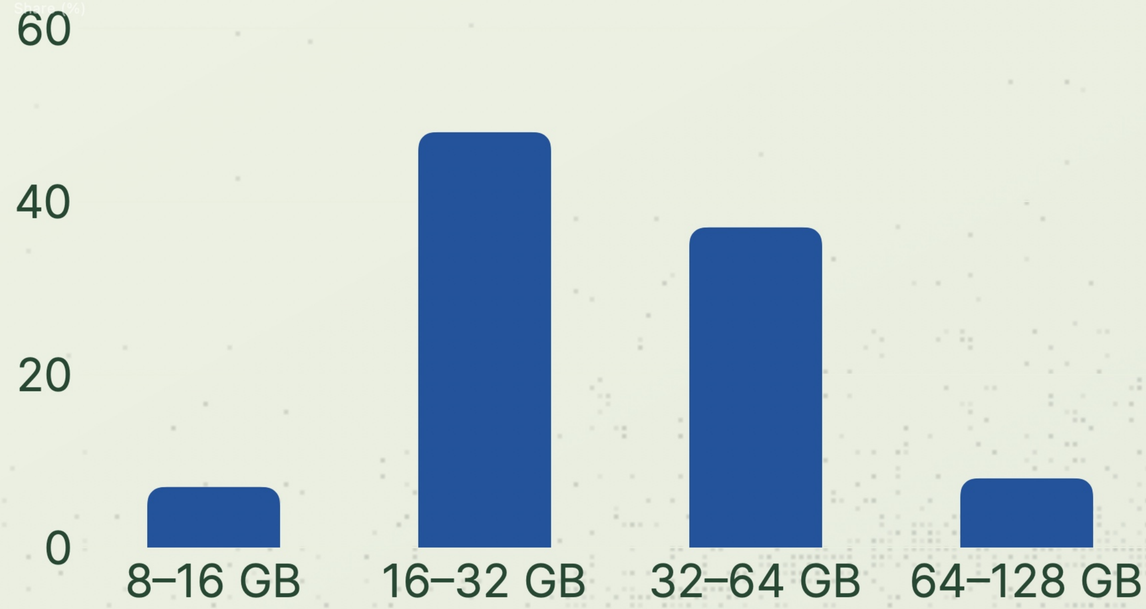
linkedin.com/in/benediktterhechte

x.com/terhechte

FIN

RAM distribution across enterprise Macs

Share of macOS devices by RAM capacity, Omnissa 2026 report.



Source: Omnissa, State of Digital Workspace 2026, page 9. The report uses capacity ranges rather than individual configurations.

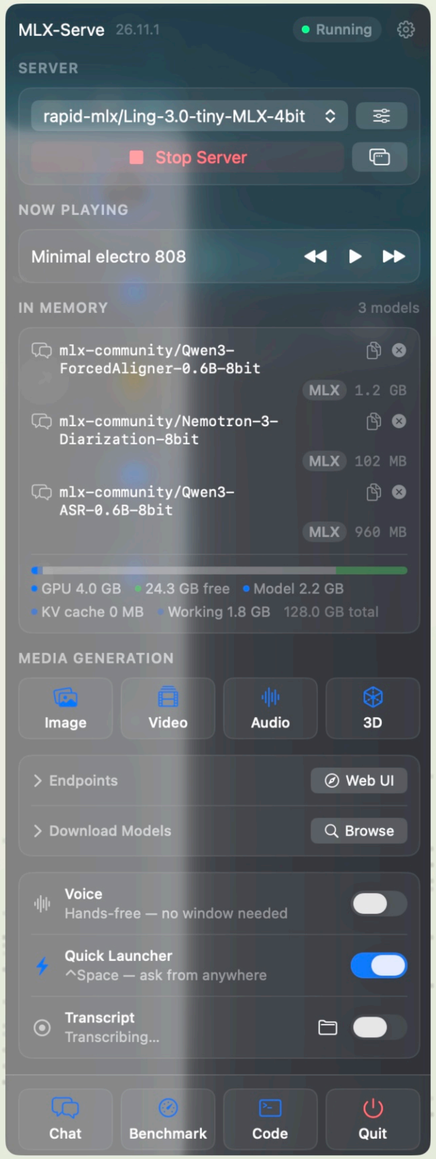
Some models think much more. slower token/s (e.g. Swift 1.5 is a fine-tune by UkisAI that thinks

Hardware

Other uses

- movie
- images
- audio

Other Uses


MLX-Serve 26.11.1
Running
SERVER
rapid-mlx/Ling-3.0-tiny-MLX-4bit
Stop Server
NOW PLAYING
Minimal electro 808
IN MEMORY
3 models
mlx-community/Qwen3-ForcedAligner-0.6B-8bit
MLX 1.2 GB
mlx-community/Nemotron-3-Diarization-8bit
MLX 102 MB
mlx-community/Qwen3-ASR-0.6B-8bit
MLX 960 MB
GPU 4.0 GB
24.3 GB free
Model 2.2 GB
KV cache 0 MB
Working 1.8 GB
128.0 GB total
MEDIA GENERATION
Image
Video
Audio
3D
Endpoints
Web UI
Download Models
Browse
Voice
Hands-free — no window needed
Quick Launcher
^Space — ask from anywhere
Transcript
Transcribing...
Chat
Benchmark
Code
Quit

Technology	What it does
SSD Streaming	Streams model weights from SSD to run models larger than RAM.
DFlash	Uses a diffusion draft model to accelerate speculative decoding.
MTP	Predicts multiple future tokens to speed up generation.
Engrams	Adds lookup-based memory for frequently recurring patterns.
Native Kernels	Uses hardware-specific code to accelerate model operations.
1bit / Ternary Weights	Represents weights with very few values to reduce memory use.
Better Quantizations	Compresses models further while preserving more accuracy.
Hybrid Linear Attention	Mixes linear and full attention to handle long contexts efficiently.
NVFP4	Uses NVIDIA's 4-bit floating-point format for efficient computation.
Weaver	Even faster DFlash (up to 200% faster), came out on the 31th

COPY PASTE FROM OPEN ZED



Artisanal, handcrafted slides

The speed of markdown with the usability of Keynote

