

Dayward

daywardai.com

Four ways to keep your data private

On business plans, “not used for training” is now the norm. The real choice is how much you run yourselves.

1 Fully local, on-prem

Open models on a Mac Studio or DGX.
Nothing leaves the building.

Tradeoff: hardware cost and upkeep are yours.

2 Frontier models, your cloud

Azure OpenAI, AWS Bedrock or Google Vertex. No training on your data.

Tradeoff: review retention and abuse monitoring.

3 Enterprise chat plans

ChatGPT, Claude or Gemini Enterprise, or M365 Copilot. Fastest rollout.

Tradeoff: your data sits with the vendor.

4 Hybrid, Apple PCC-style

Local by default. Only the hard queries go out, to a stateless private cloud.

Tradeoff: more moving parts to manage.

From your pocket to the server closet.

What actually runs on each Apple device.

iPhone & iPad

Proofread, summarize and search your manuals. On the device, offline.



• iPhone 18 Pro

CHIP

A20 Pro

MEMORY

12GB

BANDWIDTH

~115GB/s

• iPad Pro

CHIP

M5

MEMORY

up to 16GB

BANDWIDTH

153GB/s

RUNS WELL

Fast enough for proofreading and summaries.

Apple Foundation Model **BUILT IN**

3 billion parameters (AFM 3 Core), ships with iOS
Summarizes 2,000 words in ~7 sec (iPhone 17 Pro)

“Custom-built in collaboration with Google and its Gemini models” (Apple)

RUNS OK

Fine for asking questions of your own manuals, offline, with the source page cited.

webAI Frontline (iPad)

webAI’s own ColVec1.1 search model (4B or 8B)
A 34,000-page jet manual set on one iPad (webAI)

iPhone memory and bandwidth aren’t published by Apple: MacRumors; ~115GB/s calculated from Apple’s “50% higher”. iPad: apple.com. Model: Apple. Speed: Peter Frieze, iPhone 17 Pro, iOS 27 beta. Frontline: webAI, Aug 27, 2026.

MacBook Air

Use it on a flight. Client files stay on the laptop.



CHIP
M5

MEMORY
up to **32GB**

BANDWIDTH
153GB/s

ALSO LOCAL Underdog.ai A private 27B assistant; a 7.9GB version fits 16GB Macs. Invite-only.

RUNS WELL

Fast enough for chat and light coding.

Qwen3.5-35B-A3B

35 billion parameters, 3 billion active • 4-bit
~59 tokens per second in short bursts (fanless)

About Gemini 2.5 Pro level (Mar 2025), per Epoch AI

RUNS OK

Fine for making images on the go.

Qwen-Image 2.1

7 billion parameters (image generator)
~4 min per 1024px image (32GB M5 MacBook Pro)

MacBook Pro

Use it on a flight. Client files stay on the laptop.



CHIP

up to **M5 Max**

MEMORY

up to **128GB**

BANDWIDTH

up to **614GB/s**

RUNS WELL

Fast enough for chat and coding.

gpt-oss-120b

117 billion parameters, 5.1 billion active · native 4-bit
~72 tokens per second at 4K tokens of context

About OpenAI o1 level (Dec 2024), per Epoch AI

RUNS OK

Fine for batch jobs: reviewing a whole codebase or
a folder full of client contracts.

Qwen3.8-Flash-Next

125 billion parameters, 6 billion active · 4-bit
Reads 100K tokens in ~2 min, then ~61 per second
Beats Gemini 3.1 Pro (Feb 2026) on Artificial Analysis

RUNS (BARELY)

Kimi K3, 2.8 trillion parameters: ~1 token per second from four SSDs. Bragging rights only.

Specs: apple.com, Oct 21, 2026. Measured on M5 Max, 128GB: oMLX; MTPLX; Deltafin (authors' own test). Quality (full-precision models): Epoch AI; Artificial Analysis.

Mac mini

An office AI server: private Q&A over contracts and docs.



CHIP

up to **M5 Pro**

MEMORY

up to **64GB**

BANDWIDTH

up to **307GB/s**

RUNS WELL

Fast enough for live chat and coding.

Qwen3.6-35B-A3B

35 billion parameters, 3 billion active · 8-bit
~65 tokens per second at 16K context (M5 Pro)

Close to OpenAI o3 (Apr 2025) on Artificial Analysis*

RUNS OK

Fine for overnight batch jobs, like summarizing a whole folder of PDFs by morning.

Qwen3.8-27B

27 billion parameters, all active · 8-bit
~18 tokens per second, steady to 64K (M5 Pro)

Beats Gemini 3.1 Pro (Feb 2026) at max reasoning

Specs: apple.com, Oct 21, 2026. Speeds: oMLX community benchmarks, M5 Pro 64GB (oMLX logs the chip, not the Mac model). Quality: Artificial Analysis (full-precision models); *o3's score is its estimate.

Mac Studio

An office AI server: private Q&A over contracts and docs.



CHIP

M5 Ultra

MEMORY

up to **512GB[†]**

BANDWIDTH

1.2TB/s

RUNS WELL

Fast enough for live chat and coding.

DeepSeek V4.1 Flash 1017

552 billion parameters, 16 billion active · 2-bit*
~48-69 tokens per second out to 64K (256GB)

RUNS OK

Fine overnight: ingest a whole document archive
and summarize every inbox by morning.

GLM-5.3 **PROJECTED**

753 billion parameters · 4-bit · needs 512GB
~22 tokens per second, projected from M3 Ultra

[A top-3 open model on Artificial Analysis](#)

Specs: apple.com, Oct 21, 2026; [†]ships late Oct. Speeds: oMLX (80-core M5 Ultra 256GB; M3 Ultra). Projection: Dayward.
Quality: Artificial Analysis. *2-bit quantization generally gives lower-quality results than 4-bit or full precision.

Several Macs, one model

Mac Studios linked over Thunderbolt 5. Use it for team coding agents.



CHIP

M5 Ultra

MEMORY

up to **2TB**

LINK

80Gb/s

RUNS WELL

Fast enough for live chat and coding.

GLM-5.3-Flash

320 billion parameters, 18 billion active • 4-bit
~114 tokens per second on 2 M5 Ultra Mac Studios
Reads prompts at ~1,600 tokens per second

Beats Gemini 3.1 Pro (Feb 2026) on Artificial Analysis

RUNS OK

Fine for overnight team jobs: ingest a whole archive and draft every summary by morning.

Muse Spark Open

816 billion parameters, 42 billion active • 4-bit
25 tokens per second on 2 M5 Ultras, 37 on 4
Reads an 8K-token prompt in 5-8 seconds

Beats Gemini 3.1 Pro (Feb 2026) on Artificial Analysis

2TB = four 512GB Studios (calculated). 80Gb/s per Thunderbolt 5 cable; bond several. Speeds: Ash Hart (TensorFold); Apple ML Research (MLX, 8K-token prompts). All on M5 Ultra. Four 256GB Studios cost \$43,196 (calculated); the 512GB model isn't priced yet. Quality: Artificial Analysis.

WHY THE MAC, IN ONE METRIC

Per dollar, Mac memory bandwidth wins

Token generation is limited by memory bandwidth, so \$ per GB/s is a useful yardstick.



Mac Studio, M5 Ultra

NVIDIA DGX Spark

NVIDIA DGX Station

\$4.6

per GB/s

\$5,499 · 96GB at 1.2TB/s

\$25.5

per GB/s

\$6,950 · 128GB at 273GB/s

\$13–22

per GB/s

~\$93K–\$159K · 252GB at 7.1TB/s

IN PRACTICE

Nearly 2× Spark's speed

GLM-5.3-Flash: ~114 tokens per second on 2 M5 Ultra Studios vs 60+ reported on 2 Sparks

THE CATCH

Bandwidth isn't everything

NVIDIA leads on compute, so it reads prompts faster. Long first prompts and many users at once are the Mac's weak spot: at 256K tokens, an M5 Ultra takes 108 seconds to reply, 94% of it reading the prompt.

Prices as of Oct 20, 2026: apple.com; NVIDIA store; DGX Station OEM listings. Idea: Alex Cheema, X, Oct 18, 2026. Speeds as posted on X, Oct 2026: Ash Hart (@ashxhart, Mac, TensorFold 4-bit); @MiaAI_lab (Spark, setup not stated). Long prompt: MacStories, Federico Viticci (oMLX).

CREDITS

Thanks to everyone we mentioned

Their benchmarks, research and products made this deck.



ALL SOURCES

PEOPLE

Alex Cheema
[@alexocheema](#)

Ash Hart
[@ashxhart](#)

Mia
[@MiaAI_lab](#)

Peter Friese
[@peterfriese](#)

Ivan Fioravanti
[@ivanfioravanti](#)

antirez
[@antirez](#)

Jun Kim
[@jundotkim](#)

Enes Cingöz
[github.com/enescingoz](#)

Youssef Al Toukhi
[github.com/youssefal](#)

Chris Ringrose
[github.com/chrisringrose](#)

TOOLS & BENCHMARKS

oMLX
[omlx.ai](#)

MTPLX
[mtplx.com](#)

TensorFold
[tensorfold.dev](#)

Deltafin
[gavamedia · argonautlabsai](#)

mac-llm-bench
[github.com/enescingoz](#)

MLX
[github.com/ml-explore](#)

ModelFit
[@modelfit_io](#)

EXO Labs
[@exolabs](#)

Epoch AI
[@EpochAIResearch](#)

RESEARCH & REPORTING

Artificial Analysis
[@ArtificialAnlys](#)

Apple · ML Research
[@Apple](#)

MacRumors
[@MacRumors](#)

TechCrunch
[@TechCrunch](#)

MacStories
[@macstoriesnet](#)

Federico Viticci
[@viticci](#)

Context Studios
[contextstudios.ai](#)

The Register
[@TheRegister](#)

Hardware Busters
[hwbusters.com](#)

BetterClaw
[betterclaw.io](#)

MODELS

Qwen
[@Alibaba_Qwen](#)

Z.ai
[@Zai_org](#)

DeepSeek
[@deepseek_ai](#)

Kimi
[@Kimi_Moonshot](#)

OpenAI
[@OpenAI](#)

AI at Meta
[@AIatMeta](#)

Google
[@Google](#)

PRODUCTS & PLATFORMS

webAI
[@thewebAI](#)

Underdog.ai
[underdog.ai · @UnderdogAI](#)

Sigil Wen
[@0xSigil · Underdog.ai](#)

NVIDIA
[@nvidia](#)

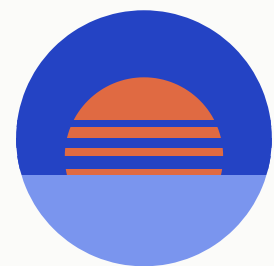
Microsoft
[@Microsoft](#)

AWS
[@awscloud](#)

Anthropic
[@AnthropicAI](#)

Hugging Face
[@huggingface](#)

Scan the code for every source, with links: [daywardai.com/deck](#)



Dayward

daywardai.com



Sources & credits
daywardai.com/deck