SOFTWARE SUSHI JOURNAL / WASHI LEDGER · DESK RESEARCH

The Best Coding Model for the Mac Studio M5 Ultra 512 GB

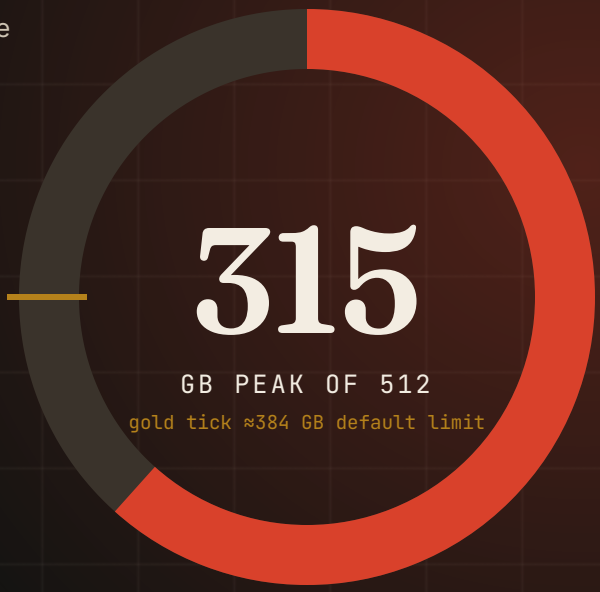
結

BEST PICK ON TODAY'S PUBLISHED EVIDENCE

GLM-5.3-Flash 8-bit.

Front-end or 3D-heavy work? **DeepSeek V4.1 Flash.**

Which open model to run on Apple's 512 GB Mac Studio, the benchmarks behind the pick, the speed measured on this machine, the setup we recommend, and how it compares with other hardware and the cloud. Every figure is sourced and rated for confidence.



92.0%	≈40 tok/s	≈2,150 tok/s	$22,299
SWE-bench Verified, Vals AI harness	8-bit decode on M5 Ultra 512 GB, 1K–32K	8-bit prefill, 8K–64K	Mac Studio M5 Ultra, 512 GB / 1 TB
確 HIGH	中 MEDIUM	中 MEDIUM	確 HIGH

Method: desk research of official specs, independent leaderboards and published, reproducible measurements. Rumors, predictions and unverified social-media claims are excluded. Software Sushi measured nothing for this report. Replaces our earlier report "Best Coding Models by Device." · softwaresushi.com

THE LEDGER

目次 Contents

結 00 The answer 03

一 01 The machine 04

二 02 The pick: GLM-5.3-Flash 05

三 03 Benchmarks: the models that fit, head to head 06

四 04 What fits in 512 GB 07

五 05 Measured speed on the M5 Ultra 08

六 06 M5 Ultra vs M3 Ultra 10

七 07 How it compares with other hardware 11

八 08 Recommended setup 12

九 09 Cost versus the API and the cloud 13

十 10 If you don't have an M5 Ultra 512 GB 15

外 11 Sidebar: a viral claim, checked 15

試 12 Gaps, and what we will test 16

典 – Sources 17

HOW TO READ THE SEALS

確 HIGH	official spec, public reproducible leaderboard, or two or more independent sources that agree.
中 MEDIUM	one independent measurement with its setup published.
社 VENDOR	the model or hardware maker's own number.
概 ESTIMATE	our arithmetic from cited figures.
未 NOT MEASURED ON M5 ULTRA	a setting or command documented elsewhere but not yet measured on this machine.

結 VERDICT

The answer

結 BEST PICK ON TODAY'S PUBLISHED EVIDENCE

GLM-5.3-Flash 8-bit.

On a Mac Studio with the M5 Ultra and 512 GB, the best-supported coding model in November 2026 is **GLM-5.3-Flash** (Z.ai, 320B total / 18B active, MIT license) in an **8-bit MLX build**, served with **oMLX**. It has the best independent agentic coding score of any open model that fits comfortably: **92.0%** on SWE-bench Verified in Vals AI's harness.[12] That comparison is against the older DeepSeek V4 Flash 0731 (88.8%); DeepSeek V4.1 Flash has no Vals SWE-bench score yet. At 8-bit it peaked at about **315 GB** and decoded at **36.6–40.4 tok/s** from 1K to 128K tokens of context, with prefill of about **2,100–2,180 tok/s** from 8K up, in a published community run on this exact configuration.[43, 44] 中 MEDIUM

If your work is front-end or 3D-heavy, pick DeepSeek V4.1 Flash. It wins 15 of 20 shared coding boards and Vals Terminal-Bench 2.1 (74.5% vs 62.9%), and the six-board coding composite is a tie (56.6 vs 56.2).[21] 確 HIGH

92.0%	≈40 tok/s	≈2,150 tok/s	$22,299
SWE-bench Verified, Vals AI harness[12] 確 HIGH	8-bit decode on M5 Ultra 512 GB, plain, 1K–32K; 36.6 at 128K[43] 中 MEDIUM	8-bit prefill, 8K–64K; a 128K prompt took 62 s[43] 中 MEDIUM	Mac Studio M5 Ultra, 512 GB / 1 TB[2] 確 HIGH

Five things to know

1. **The pick, on today's published evidence:** GLM-5.3-Flash 8-bit. It beats DeepSeek V4 Flash 0731 (88.8%) and Qwen3.8-27B (86.0%) on the same Vals board (DeepSeek V4.1 Flash has not been scored there yet); only the 754B GLM-5.3 (95.4%) and closed Claude Opus 5 (97.0%) score higher.[12] It is plain MIT.[17]
2. **The alternatives that fit: DeepSeek V4.1 Flash** for front-end and 3D work (it wins 15 of 20 shared coding boards and Vals Terminal-Bench 2.1; the six-board composite is a tie)[21]; the full **GLM-5.3** at 406 GB for maximum quality at 19–25 tok/s[5]; **Qwen3.8-27B** as a fast small helper (48 tok/s).[11] Kimi K3 and DeepSeek V4 Pro do not fit.[8, 9]
3. **Speed options:** plain 8-bit gives about 40 tok/s with maximum quality. A 4-bit build with multi-token prediction (MTP) reached 61.5–140 tok/s at 177–179 GB in another M5 Ultra 512 GB run, varying run to run.[45, 46] 中 MEDIUM
4. **The setup:** oMLX (Homebrew), an 8-bit MLX build (334.1 GB on disk), `omlx launch claude` for Claude Code, 64K+ context, lower reasoning effort for quick edits.[49, 50, 48, 34, 41]
5. **The cost:** counting output tokens only, the GLM-5.3-Flash API is 8–50× cheaper per token than owning the machine; in an input-heavy agent session the gap is about 2–8× (5–22× with API prompt caching). Buy it for privacy, offline work, fixed cost or a shared team box, not to save money. 概 ESTIMATE

Honest gap. The 8-bit M5 Ultra 512 GB speed comes from one community submitter on one platform (oMLX, Nov 14, 2026), with the settings published.[43, 44] No reviewer or second lab has reproduced it yet, so it is rated Medium. Software Sushi can film the first independent reproduction (section 12).

Confidence labels: 確 HIGH official spec, public reproducible leaderboard, or two or more independent sources that agree. 中 MEDIUM one independent measurement with its setup published. 社 VENDOR the model or hardware maker's own number. 概 ESTIMATE our arithmetic from cited figures. 未 NOT MEASURED ON M5 ULTRA a setting or command documented elsewhere but not yet measured on this machine.

CHAPTER 01

The machine

SPEC	MAC STUDIO M5 ULTRA, 512 GB	WHY IT MATTERS	CONFIDENCE
Chip	36-core CPU, 80-core GPU with a Neural Accelerator in each core; the 512 GB option requires this chip [1, 2]	Compute sets prefill speed	確 HIGH
Memory / bandwidth	512 GB unified · 1.2 TB/s (M3 Ultra: 819 GB/s) [1, 11]	Capacity decides what fits; bandwidth sets decode speed	確 HIGH
GPU-usable memory	About 75% by default (≈384 GB on 512 GB); raise iogpu.wired_limit_mb for bigger models [7, 8, 49]	8-bit GLM-5.3-Flash fits under the default; GLM-5.3 does not	確 HIGH 概 ESTIMATE for 384
Price and availability	$22,299 (512 GB / 1 TB), orderable since Oct 28, 2026; $29,799 with 16 TB [2]	Fixed cost	確 HIGH
Power	9 W idle, 385 W maximum (Apple, 512 GB / 16 TB config); 480 W PSU [3, 1]; about 45 W in single-user LLM generation (one reviewer) [6]	Quiet, desk-friendly	確 HIGH / 中 MEDIUM

Three numbers decide the experience

- **Fit:** a mixture-of-experts (MoE) model must hold *every* expert in memory, though each token reads only the active ones. GLM-5.3-Flash has 320B parameters in total and 18B active. [16]
- **Decode (tok/s):** how fast the model writes. Bound mostly by memory bandwidth.
- **Prefill (prompt tok/s):** how fast it reads your repository before answering. Bound by compute. Coding agents re-send tens of thousands of tokens on every turn, so prefill is what you feel. [11]

How we rate evidence

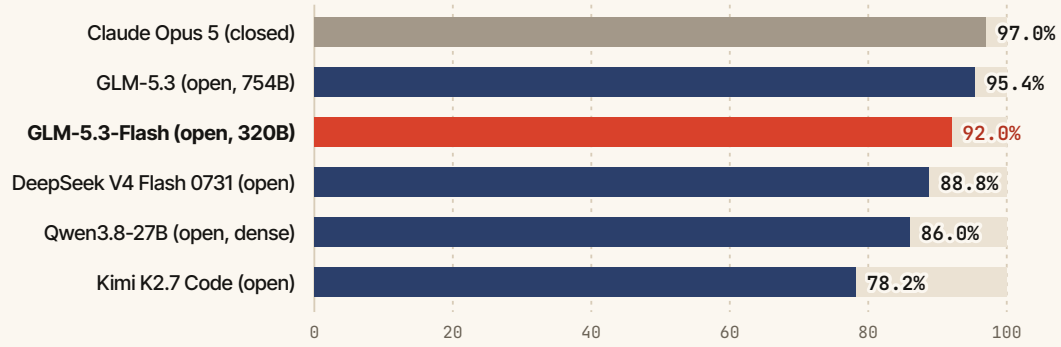
Harness and settings change results. Z.ai reports GLM-5.3-Flash at 84.3 on Terminal-Bench 2.1 with its own harness; Vals AI's independent run scores it 62.9%. [16, 21] Speculative decoding (MTP, EAGLE, DFlash) can double decode speed on code, so every speculative run is marked. We compare models only within one leaderboard.

CHAPTER 02

The pick: GLM-5.3-Flash

ITEM	VALUE	CONFIDENCE
Release	Aug 26, 2026, open weights on Hugging Face the same day [16, 19]	確 HIGH
Architecture	320B total, 18B active MoE; 45 layers; hybrid linear plus sparse attention [16, 18]	確 HIGH
Context	1M tokens [18, 33]	確 HIGH
Max output	128K per Z.ai, 163,840 per the HF card (sources disagree) [18, 33]	中 MEDIUM
License	**MIT** (the full 754B GLM-5.3 uses a custom license) [17, 19]	確 HIGH
API price (for comparison)	$0.15 in / $0.50 out per million tokens [22, 33]	確 HIGH
Mac runtimes	MLX via mlx-vlm or oMLX (mlx-lm lacks the glm5_next architecture); llama.cpp GGUF also works [40, 28, 31]	確 HIGH

SWE-bench Verified, one harness (Vals AI)



確 HIGH One independent harness, read Nov 16, 2026. [12] DeepSeek V4.1 Flash has no Vals SWE-bench result yet; V4 Flash 0731 is shown. All open models listed fit in 512 GB except where section 4 says otherwise.

Why it is the pick on today's evidence: it is the highest-scoring open model on Vals SWE-bench that fits at 8-bit with room to spare (DeepSeek V4.1 Flash has not been scored on that board, so this could change), it is MIT-licensed, and it has the most published Mac measurements of any model in its class. GLM-5.3 scores higher (95.4%) but runs at half the speed (section 5) and needs a raised memory limit (sections 4 and 8).

CHAPTER 03

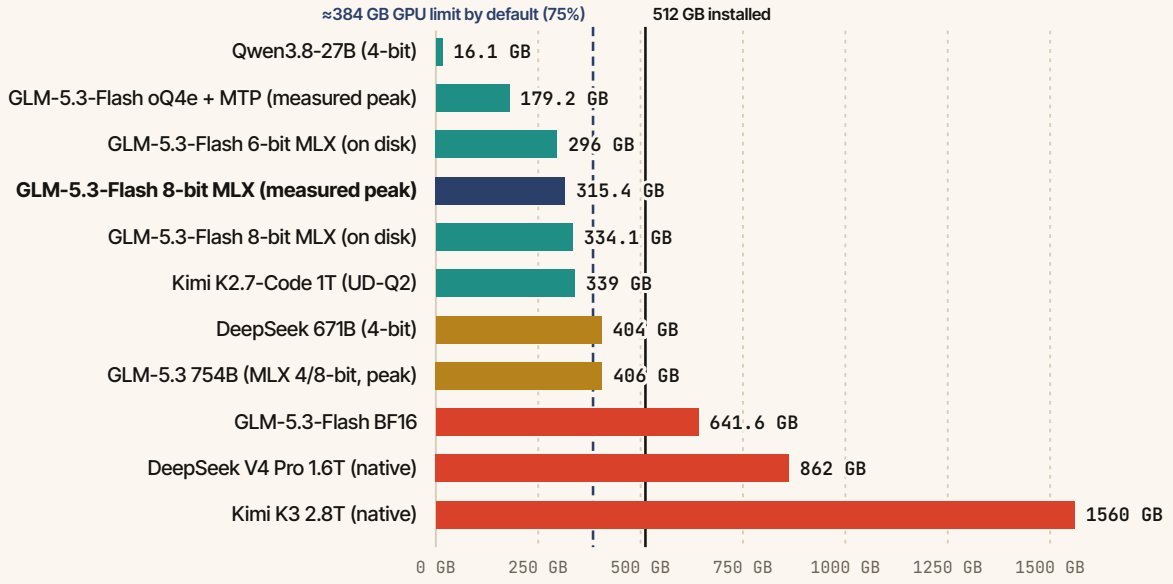
Benchmarks: the models that fit, head to head

BENCHMARK (WHO RAN IT)	GLM-5.3-FLASH	DEEPSEEK V4.1 FLASH	GLM-5.3	QWEN3.8-27B
SWE-bench Verified (Vals) [12]	**92.0%**	n/a (V4 Flash: 88.8%)	95.4%	86.0%
LiveCodeBench (Vals) [20]	80.5%	n/a	80.5%	**84.0%**
Terminal-Bench 2.1 (Vals) [21]	62.9%	**74.5%**	—	—
Terminal-Bench 4.0 (Artificial Analysis) [21, 22]	**32.8%**	26.8%	—	5.6% (xhigh)
Aider Polyglot, pass@2 [23]	42.8%	43.0%	**46.5%**	—
Vibe Code Bench (Vals) [21]	30.8%	**84.7%**	—	—
Design Arena, 3D (Elo) [21]	1293	**1339**	—	—
Coding boards won, head to head [21]	5 of 20	**15 of 20**	—	—
Six-board coding composite (LLMBenchmark) [21]	56.2	56.6	—	69.7

Reading the table: GLM-5.3-Flash leads the agentic, repository-level work that coding agents do (SWE-bench, Terminal-Bench 4.0). DeepSeek V4.1 Flash leads front-end, 3D and Vals Terminal-Bench 2.1. Composites disagree (LLMBenchmark's rates Qwen3.8-27B highest; Artificial Analysis rates GLM-5.3-Flash well above it [22]), so we rank on individual boards. 社 VENDOR Z.ai's own launch table is not used for ranking. [16]

CHAPTER 04

四 What fits in 512 GB



Indigo: the pick. Teal: fits under the default GPU limit. Gold: fits only after raising the wired limit, with less room for context. Vermilion: does not fit one machine. Peaks are measured MLX peaks on 512 GB machines; disk sizes from model cards and Unsloth tables.[43, 45, 50, 51, 5, 8, 28, 9] 概 ESTIMATE for the 384 GB line (75% of 512 GB).

— Quantization: why 8-bit

BUILD	SIZE	QUALITY SIGNAL	CONFIDENCE
8-bit MLX	334.1 GB on disk; 312.6–315.4 GB MLX peak on M5 Ultra[50, 43]	Builder's quality anchor; 8-bit has been indistinguishable from BF16 on their other models[50]	中 MEDIUM
6-bit MLX	≈296 GB (orcarouter OrcaSAQ build)[51]	+0.1% perplexity vs 8-bit (pipenetwork build)[50]	中 MEDIUM
Mixed 4/8-bit MLX	≈176 GB peak[11]	+3.2% perplexity vs 8-bit[50]	中 MEDIUM
UD-Q4_K_XL GGUF	≈200 GB	92.2% top-1 agreement with BF16 (3-bit 86.3%, 2-bit 78.3%)[28]	確 HIGH

With 512 GB there is no reason to give up quality on the default model: the 8-bit build peaked at 315.4 GB, under the ≈384 GB default GPU limit, and peak memory rose only 2.8 GB from a 1K to a 128K prompt.[43] That leaves ≈197 GB of the 512 GB installed, but only ≈69 GB of GPU-usable headroom at the default limit (≈61 GB against the run's 322.9 GB system wired peak). Running a second large model or a very large KV cache next to it means raising `iogpu.wired_limit_mb`.

概 ESTIMATE

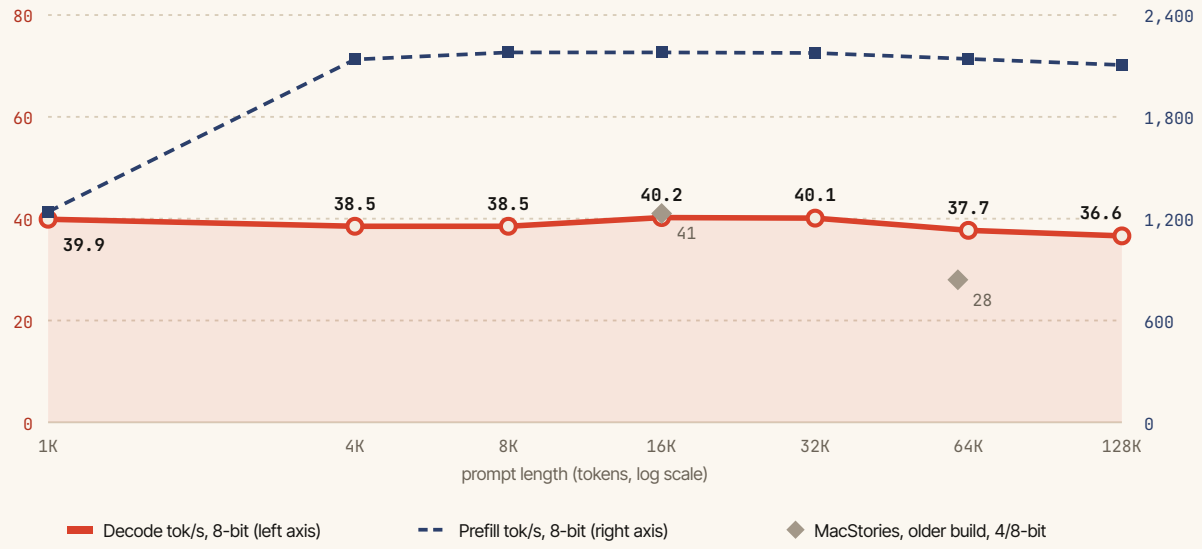
8-bit build caveat: the pipenetwork card reports two numerical bugs in mlx-vlm's early glm5_next port (a missing SwiGLU clamp and an mHC dtype cast) and ships a patched runtime; use a runtime that includes those fixes.[50] We could not confirm which oMLX releases include them.

未 NOT MEASURED ON M5 ULTRA

CHAPTER 05

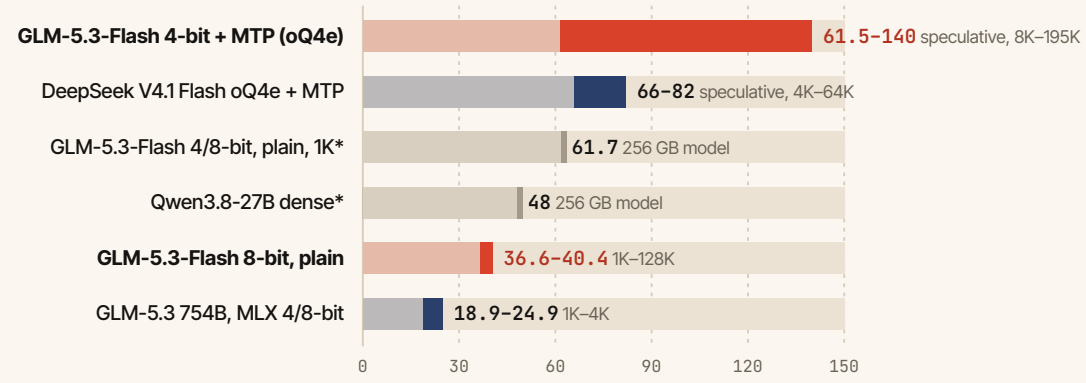
五 Measured speed on the M5 Ultra

GLM-5.3-Flash 8-bit on M5 Ultra 512 GB, by prompt length



MEDIUM oMLX v0.8.0rc1, macOS 27.2, no speculative decoding, code context, one submitter, Nov 14, 2026. [43] Low prefill at 1K reflects warm-up (a second session measured 613.5 tok/s at 1K and 1,457 at 4K). [44] Gray diamonds: MacStories' 4/8-bit run on the 256 GB model with an earlier oMLX build, 41 tok/s at 16K and 28 at 61K, including reasoning tokens. [11]

Every coding model measured on an M5 Ultra (decode tok/s)



Dark segment: measured range; light segment: zero to the low end. * measured on the 256 GB model (same chip and bandwidth). Speculative rows use MTP. Sources: 45, 46, 4, 27, 11, 43, 5.

MODEL AND BUILD ON M5 ULTRA	DECODE	PREFILL	MEMORY	CONFIDENCE
GLM-5.3-Flash 8-bit MLX, plain (512 GB)	39.9–40.4 at 1K–32K; 37.7 at 64K; 36.6 at 128K	2,105–2,180 tok/s at 8K–128K; 128K TTFT 62.3 s	312.6–315.4 GB [43, 44]	中 MEDIUM
GLM-5.3-Flash oQ4e + MTP (512 GB)	61.5–140.0 (16K: 138.9–140.0; 195K: 61.5)	2,166–2,299 tok/s at 8K–195K	176.5–179.2 GB [45, 46]	中 MEDIUM speculative
GLM-5.3-Flash 4/8-bit, plain (256 GB)	61.7 at 1K (oMLX); 41 at 16K, 28 at 61K (MacStories, incl. reasoning)	1,107 at 16K; 61K prompt in 62.5 s	≈176 GB [27, 11]	確 HIGH (two sources)
DeepSeek V4.1 Flash oQ4e + MTP (512 GB)	66–82	1,239–1,308 at 4K–64K	— [4]	中 MEDIUM speculative
GLM-5.3 754B, MLX 4/8-bit (512 GB)	24.9 at 1K; 18.9 at 4K	590.8 tok/s	406 GB [5]	中 MEDIUM
Qwen3.8-27B dense (256 GB)	48	1,701 at 6K	small [11]	確 HIGH

— Serving a team: batching

The same 8-bit session measured aggregate decode of 41.8 tok/s with 2 parallel requests, 55.0 with 4 and **85.9 with 8** (2.15×). [43] 中 MEDIUM A shorter session reported higher scaling (159.4 tok/s at 4 requests) [44]; we use the conservative figure until it is reproduced. On an M3 Ultra, 8 requests gave 79.5 tok/s. [26]

— Context length

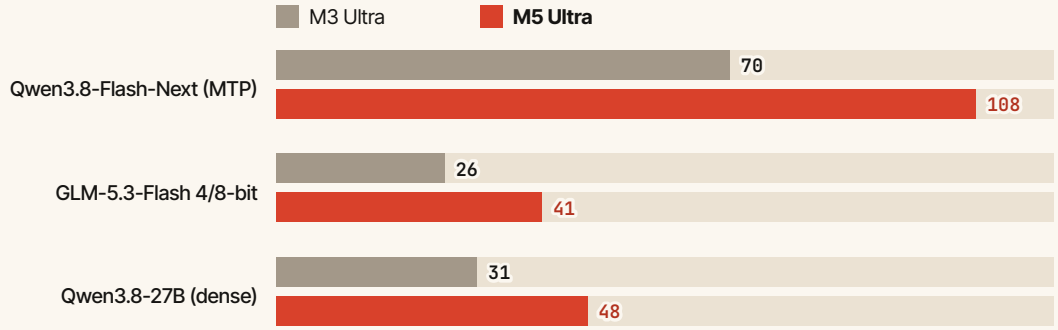
The newer oMLX build held 8-bit decode within 9% from 1K to 128K (40.2 → 36.6). MacStories' earlier build lost about a third between 16K and 61K (41 → 28). [43, 11] Keep the runtime current, and treat long-context speeds from older builds as a floor.

六

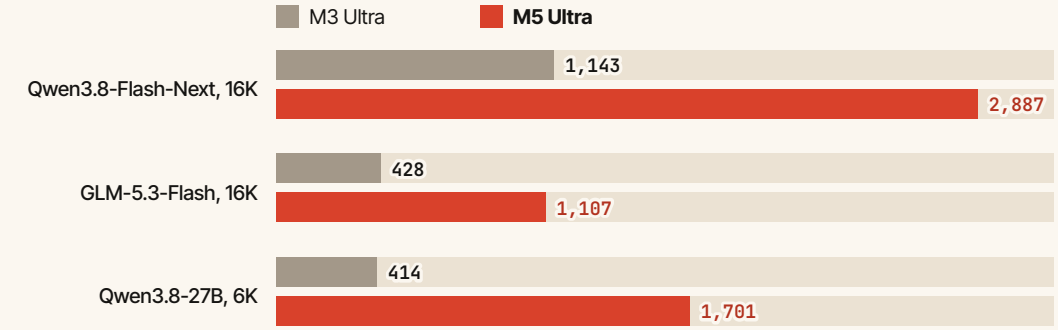
CHAPTER 06

M5 Ultra vs M3 Ultra

Decode, tokens/s (same software, same prompts)



Prefill, tokens/s



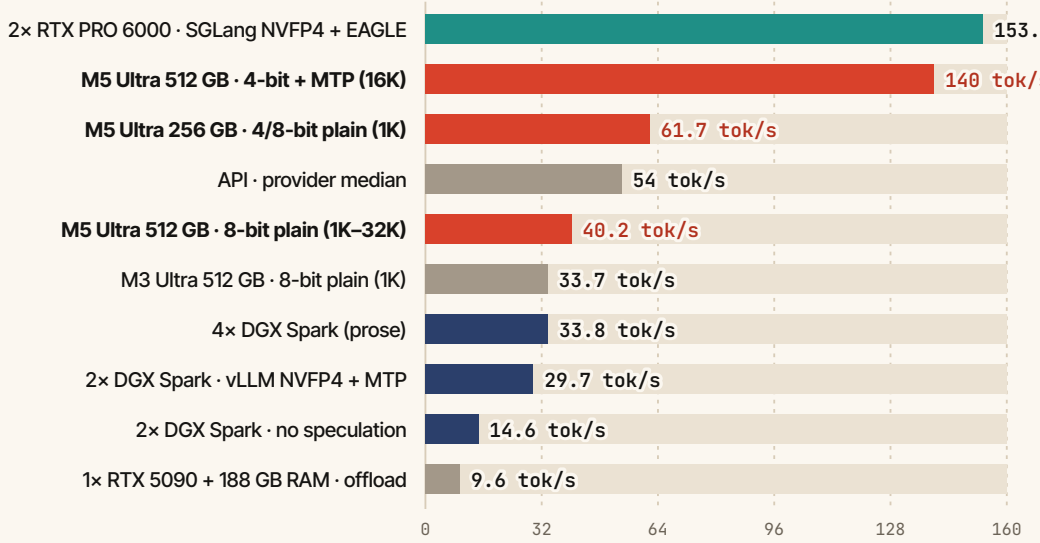
確 HIGH MacStories ran the same oMLX build, model files and prompts on an M3 Ultra 512 GB and an M5 Ultra 256 GB: decode +54–58%, prefill 2.5–4.1×. [11] The llama.cpp standard test agrees on direction (119 vs 64 tok/s decode; 4,911 vs 1,488 prefill), on different builds. [10]

GLM-5.3-FLASH 8-BIT, OMLX COMMUNITY RUNS	M3 ULTRA 512 GB (V0.7.3)	M5 ULTRA 512 GB (V0.8.0RC1)	CHANGE
Decode at 1K	33.7 tok/s [26]	39.9–40.2 tok/s [43, 44]	+18–19%
Prefill at 4K	542.5 tok/s [26]	1,457–2,138 tok/s [44, 43]	2.7–3.9×
Peak memory	314.0 GB [26]	313.4 GB [43]	same

概 ESTIMATE Different submitters and oMLX versions, so this shows direction, not a controlled test. The consistent message across sources: the M5 Ultra's largest gain is prefill (reading the repository), which is what coding agents spend most of their time on.

CHAPTER 07

How it compares with other hardware



GLM-5.3-Flash single-stream decode. Vermilion: M5 Ultra. Speculative runs are labeled. Sources: 13, 45, 27, 22, 43, 26, 24, 29, 30, 31.

Two RTX PRO 6000 cards are the fastest measured GLM-5.3-Flash box (153.5 tok/s with EAGLE)[13], but at NVFP4, not 8-bit, and with no published system price. Two DGX Sparks reach 14.6 tok/s plain and 26.5–29.7 with MTP, reproduced by three groups.[29, 30, 32] The highest two-node figure we found is 32.9 tok/s: Context Studios' own run on 2× ASUS Ascent GX10 (the same GB10 chip) with an EXL3 4-bit build in vLLM.[25] The M5 Ultra 512 GB is the only single quiet box with a published 8-bit run.

CHAPTER 08

八 Recommended setup

Commands below come from the cited docs. Items marked 未 NOT MEASURED ON M5 ULTRA are documented but have not been measured on this machine.

— 8.1 Install and serve (oMLX)

zsh — mac-studio

```
# 1. Install oMLX (Homebrew route from the oMLX guide)
brew tap jundot/omlx https://github.com/jundot/omlx
brew install jundot/omlx/omlx
# 2. Download an 8-bit MLX build (334.1 GB on disk)
hf download pipenetwork/GLM-5.3-Flash-MLX-8bit --local-dir ~/models/GLM-5.3-Flash-MLX-8bit
# 3. Serve every model in ~/models on localhost:8000 (OpenAI + Anthropic APIs)
omlx serve --model-dir ~/models
# 4. Start Claude Code against it (sets ANTHROPIC_BASE_URL and model tiers)
omlx launch claude
```

Sources: install, serve and SSH tunnel from the oMLX guide [49]; Anthropic `/v1/messages` endpoint and Claude Code launcher from oMLX [47, 48]; build and size from the model card. [50] The measured run used an oMLX build labeled GLM-5.3-Flash-MLX-8bit [43]; whether it is byte-identical to the pipenetwork repo is not stated. 未 NOT MEASURED ON M5 ULTRA

— 8.2 Settings that matter

SETTING	RECOMMENDATION	WHY / SOURCE	CONFIDENCE
Quantization	8-bit for quality; oQ4e + MTP when speed matters more	8-bit ≈40 tok/s; 4-bit + MTP 61.5–140 [43, 45]	中 MEDIUM
Context	Give agents 64K+ tokens; oMLX's Claude Code launcher enforces at least 48K	Ollama's Claude Code guide asks for 64K+ [34, 49]	確 HIGH
Reasoning effort	Default is max; use low or high for quick edits	Saves tokens and time [41]	確 HIGH 未 NOT MEASURED ON M5 ULTRA
GPU memory limit	Not needed to load the 8-bit build: 315.4 GB MLX peak (322.9 GB system wired peak) is under the ≈384 GB default, leaving ≈61–69 GB. Needed for GLM-5.3 (406 GB), or for a second large model or very large KV cache: `sudo sysctl iogpu.wired_limit_mb=<MB>` , reset at reboot	Peak from [43]; command from [49, 7]	中 MEDIUM 概 ESTIMATE
Remote use	From a laptop: `ssh -L 8000:localhost:8000 user@mac-studio` , then point the agent at 127.0.0.1:8000	oMLX guide [49]	確 HIGH 未 NOT MEASURED ON M5 ULTRA
Team use	Let several agents share the server; 8 parallel requests ≈86 tok/s aggregate	Batching results [43]	中 MEDIUM

8.3 Wiring coding agents

TOOL	HOW IT CONNECTS TO THE STUDIO	WATCH OUT FOR	CONFIDENCE
Claude Code	`omlx launch claude`, or ANTHROPIC_BASE_URL=http://127.0.0.1:8000 plus model-tier variables [48, 49]	Claude Code is tuned for Claude; expect weaker tool-call reliability from other models [49]	確 HIGH
Aider, Cline, OpenHands	Any OpenAI-compatible URL: http://localhost:8000/v1 [47, 38, 39]	OpenHands needs 22K+ context; Cline: turn on Compact Prompt [39, 38]	確 HIGH
Cursor	Override OpenAI Base URL with a **public HTTPS** endpoint [36, 37]	Requests pass through Cursor's servers: not private or offline [36]	確 HIGH
Ollama	GLM-5.3-Flash is a **cloud-only** tag; no local Ollama build [35]	Use oMLX, mlx-vlm or llama.cpp for local	確 HIGH

CHAPTER 09

九 Cost versus the API and the cloud

SCENARIO	INPUTS 概 ESTIMATE	RESULT
$22,299 at GLM-5.3-Flash API prices	$0.50 per million output tokens ($0.15 input, $0.03 cached input) [2, 22]	≈44.6 billion output tokens
$22,299 at Claude Opus 5 (Anthropic API)	$25 per million output ($5 input) [52]	≈0.89 billion output tokens
$22,299 vs AWS 8× H200	$63.296/hr (p5en.48xlarge) [15]	≈352 instance-hours (≈14.7 days 24/7)
M5 Ultra 512 GB, 3 years, workday use	8 h/day, 250 days/yr, 40–62 tok/s, 385 W max at 18.19¢/kWh [3, 42]	≈$17–26 per million output tokens
M5 Ultra 512 GB, 3 years, 24/7	≈$613/yr power at full load, 40–62 tok/s	≈$4.1–6.4 per million output tokens

— A worked agent session (input included)

The 8–50× figure above counts **output tokens only**. Coding agents resend the repository context on every turn, so input tokens dominate the API bill. 概 ESTIMATE Assumptions: 40 turns, an average 50K tokens of context resent per turn (2.0M input tokens) and 1,500 tokens written per turn (60K output). Local speed from the measured 8-bit run (≈2,150 tok/s prefill, ≈40 tok/s decode), with no prompt-cache reuse on the Mac. [43]

PER SESSION	GLM-5.3-FLASH API	M5 ULTRA 512 GB, WORKDAY USE	M5 ULTRA 512 GB, 24/7
Cost basis	$0.15 in / $0.50 out per million; cached input $0.03 [22]	$3.79 per hour (3-year amortization, 8 h × 250 days, 385 W)	$0.92 per hour (3-year amortization, 24/7, 385 W)
Time	—	15.5 min prefill + 25 min decode = 40.5 min	40.5 min
Cost, no caching	$0.30 input + $0.03 output = **$0.33** (input is 91%)	**$2.56** (7.7× the API)	**$0.62** (1.9× the API)
Cost, 90% of input cached	**$0.11**	22× the API	5.4× the API

概 ESTIMATE Counting input narrows the gap from 8–50× (output only) to about 2–8× without API caching, or about 5–22× with it. The Mac's fast prefill is why: it reads resent context at ≈2,150 tok/s instead of paying per token. Local prefix caching would shorten the Mac's time further; we have not measured it.

Bottom line: per token the GLM-5.3-Flash API is 8–50× cheaper than owning the box on output tokens alone, and about 2–22× cheaper once input tokens are counted. Buy the M5 Ultra when code cannot leave the building, when you want a fixed cost with no rate limits, when you work offline, or when a small team shares one machine. Against Claude Opus 5 at $25 per million output tokens (the model behind the 97.0% score), a heavily used box does cost less per token, but it runs a model that scores 92.0% rather than 97.0% on Vals SWE-bench. [12] Token counts include GLM's reasoning tokens; power uses Apple's maximum rating, so real power cost is lower.

十

CHAPTER 10

If you don't have an M5 Ultra 512 GB

YOUR HARDWARE	PICK	MEASURED	CONFIDENCE
M5 Ultra 256 GB	GLM-5.3-Flash 4-bit (≈176 GB)	61.7 tok/s at 1K; 41 at 16K [27, 11]	確 HIGH
M3 Ultra 512 GB	GLM-5.3-Flash 8-bit	33.7 tok/s at 1K; 542.5 prefill at 4K [26]	中 MEDIUM
2× RTX PRO 6000 (192 GB)	GLM-5.3-Flash NVFP4 (SGLang, EAGLE)	153.5 tok/s [13]	中 MEDIUM
2–4× DGX Spark	GLM-5.3-Flash NVFP4 (vLLM, MTP)	26.5–29.7 tok/s on 2× (NVFP4 + MTP); 32.9 on 2× GX10 (EXL3 4-bit); 33.8–58.4 on 4× [29, 30, 25, 24]	確 HIGH / 中 MEDIUM
128 GB Mac (M5 Max)	gpt-oss-120b	79.1 tok/s [14]	中 MEDIUM
16–32 GB GPU or laptop	Qwen3.8-27B	59 tok/s on RTX 5090; 86.0% Vals SWE-bench [11, 12]	確 HIGH
No hardware	GLM-5.3-Flash API	$0.15 / $0.50 per million; ≈54 tok/s [22]	確 HIGH

外

CHAPTER 11

Sidebar: a viral claim, checked

CLAIM (X, SEP 21, 2026)	WHAT THE SOURCES SAY	VERDICT
"The M5 Ultra is slower than two DGX Sparks: 28 vs 39 tok/s, 1,016 vs 1,700 prefill" (@MiaAI_lab)	The 28 and 1,016 are MacStories' figures for a 61K-token prompt; at 16K MacStories shows 41 tok/s and 1,107 prefill. [11] No independent 2× Spark run reaches 39 tok/s: published runs span 14.6–32.9, where 29.7 is the best NVFP4 + MTP result and 32.9 is Context Studios' 2× GX10 run (same GB10 chip, EXL3 4-bit, vLLM). [29, 30, 25] The newer 8-bit M5 Ultra 512 GB run holds 36.6 tok/s and 2,105 prefill at 128K. [43]	**Misleading**: real numbers, wrong comparison

Our earlier report, “Best Coding Models by Device,” contains the full claim-by-claim check of this account. Removed as unverifiable: self-reported 271 / 114 tok/s TensorFold figures for two M5 Ultras, and a GLM 5.5 size rumor.

試

CHAPTER 12

Gaps, and what we will test

- **One source for the 8-bit M5 Ultra speed.** All 8-bit M5 Ultra 512 GB numbers come from one oMLX community submitter (two sessions).[43, 44] Reviewers (MacStories, others) tested the 256 GB model at 4/8-bit.
- **Speculative 4-bit numbers swing widely** (61.5–140 tok/s) between runs and prompt lengths.[45, 46]
- **No M5 Ultra runs** for Kimi K2.7-Code or GLM-5.2; DeepSeek V4.1 Flash and GLM-5.3 have one platform each.[4, 5]
- **DeepSeek V4.1 Flash** has no Vals SWE-bench or LiveCodeBench result yet.
- **Agent-level speed** (minutes per real task in Claude Code) has not been published for this machine.
- **Runtime correctness:** which oMLX releases include the glm5_next numerical fixes is not documented.[50]

> **Software Sushi test plan (to film):** on a Mac Studio M5 Ultra 512 GB, run oMLX with the 8-bit GLM-5.3-Flash build; record the built-in benchmark at 1K–128K (reproducing the community run), then a real Claude Code task on a small repository with a timer, tok/s and memory on screen. Publish the recipe and raw logs. If the numbers differ, we publish ours.

典

BIBLIOGRAPHY

Sources

[1] Apple — Mac Studio technical specifications — https://www.apple.com/mac-studio/specs/

[2] MacRumors — Mac Studio with M5 Ultra now available with 512GB for $22,299 (Oct 28, 2026) — https://www.macrumors.com/2026/10/28/mac-studio-m5-ultra-512gb-ram/

[3] Apple Support — Mac Studio power consumption and thermal output — https://support.apple.com/en-us/102027

[4] oMLX Community Benchmarks — DeepSeek-V4.1-Flash oQ4e-mtp on M5 Ultra 512GB (Nov 9, 2026) — https://omlx.ai/benchmarks/performance?cursor=WyJwZXJmb3JtYW5jZTpoaWRlX3NwZWNwcmVmaWxsPTEiLCIyMDI2LTExLTA5IDE2OjQxOjQzliw2MTAxNjIsMiwyLDEwXQ%3D%3D&hide_specprefill=1

[5] oMLX Benchmark — GLM-5.3-MLX-mixed-4_8bit-mtp on M5 Ultra 512GB — https://omlx.ai/benchmarks/performance/gteyri10

[6] senn-tech — M5 Ultra: first measured local LLM numbers — https://senn-tech.com/en/blog/m5-ultra-erste-messwerte-lokale-llms

[7] llamaperf — M5 Ultra Mac Studio for local LLMs: early measurements (cites MacStories) — https://llamaperf.com/blog/m5-ultra-mac-studio-local-llm

[8] Pinggy — Self-Hosting LLMs on the 512GB M5 Ultra Mac Studio: What Actually Fits (Sep 1, 2026) — https://pinggy.io/blog/self_hosting_llms_on_512gb_m5_ultra_mac_studio/

[9] Dr. Chaos — The 512GB M5 Ultra Mac Studio for Local AI and Security (Nov 6, 2026) — https://www.drchaos.com/blog/mac-studio-m5-ultra-512gb-local-ai-security

[10] llama.cpp Apple Silicon thread (discussion 4167), as summarised by llamaperf: same llama-bench test on M3/M5 Ultra — https://llamaperf.com/blog/m5-ultra-mac-studio-local-llm

[11] MacStories — M5 Ultra Mac Studio Review: The Dream Mac for Local AI Agents (Sep 21, 2026) — https://www.macstories.net/stories/m5-ultra-mac-studio-review-the-dream-mac-for-local-ai-agents/

[12] LLMBenchmark.ai — SWE-bench Verified (Vals AI) leaderboard, read Nov 16, 2026 — https://llmbenchmark.ai/benchmarks/vals-swe-bench

[13] Helix — GLM-5.3-Flash on 2, 4 or 8 RTX PRO 6000 GPUs (Sep 14, 2026) — https://helix.ml/blog/glm53-flash-tp2-rtx-pro-6000

[14] llamaperf — M5 Max 128GB local LLM performance (community reports) — https://llamaperf.com/gpu/m5-max-128gb

[15] CloudCostDB — What a GPU actually costs on EC2, and when Bedrock tokens win (Nov 6, 2026) — https://cloudcostdb.com/blog/gpu-hour-economics/

[16] Z.ai — GLM-5.3-Flash: Frontier Intelligence, Flash Cost (launch post, Aug 26, 2026) — https://z.ai/blog/glm-5.3-flash

[17] Hugging Face — zai-org/GLM-5.3-Flash LICENSE file (MIT) — https://huggingface.co/zai-org/GLM-5.3-Flash/raw/main/LICENSE

[18] Z.ai developer docs — GLM-5.3-Flash/FlashX — https://docs.z.ai/guides/vlm/glm-5.3-flash

[19] Digital Applied — GLM-5.3's weights are out. The licence is not MIT (Aug 28, 2026) — https://www.digitalapplied.com/blog/glm-5-3-weights-bespoke-license-not-mit

[20] LLMBenchmark.ai — LiveCodeBench (Vals AI) leaderboard, read Nov 16, 2026 — https://llmbenchmark.ai/benchmarks/vals-livecodebench

[21] LLMBenchmark.ai — DeepSeek V4.1 Flash vs GLM 5.3 Flash (87 shared benchmarks, read Nov 16, 2026) — https://llmbenchmark.ai/compare/deepseek-v4-1-flash-vs-glm-5-3-flash

[22] Artificial Analysis — GLM-5.3-Flash vs Qwen3.8 27B (prices, output speed, Terminal-Bench 4.0), read Nov 16, 2026 — https://artificialanalysis.ai/models/comparisons/glm-5-3-flash-vs-qwen3-8-27b-medium

[23] Vibecoder Journal — Aider Polyglot leaderboard (updated Nov 15, 2026) — https://vibecoderjournal.com/benchmarks/aider-polyglot/

[24] measuredlocally.com — M5 Ultra Mac Studio vs four DGX Sparks: the same model (Oct 3, 2026) — https://measuredlocally.com/mac-studio-m5-ultra-vs-sparks

[25] Context Studios — Mac Studio M5 Ultra for local AI, with own 2× GX10 measurements (updated Oct 12, 2026) — https://www.contextstudios.ai/blog/mac-studio-m5-ultra-local-ai-guide

[26] oMLX Benchmark — GLM-5.3-Flash MLX 8-bit on M3 Ultra 512GB — https://omlx.ai/benchmarks/performance/194n2ytm

[27] LLMCheck — GLM-5.3-Flash on M5 Ultra (lists oMLX plain run, 61.7 tok/s at 1K) — https://llmcheck.net/models/glm-53-flash-on-m5-ultra/

[28] Unsloth — GLM-5.3-Flash: how to run locally (quantization KLD table) — https://unsloth.ai/docs/models/glm-5.3-flash

[29] DevelopersIO (Classmethod) — GLM-5.3-Flash NVFP4 on 2× DGX Spark with vLLM (Sep 2026) — https://dev.classmethod.jp/articles/dgx-spark-2node-glm-5-3-flash-nvfp4-vllm/

[30] LimeChain — GLM-5.3-Flash on 2× DGX Spark, vLLM path (GitHub) — https://github.com/LimeChain/glm-5.3-flash-2xdgx-spark/blob/main/docs/vllm-path.md

[31] Graphometer — GLM-5.3-Flash on one desktop (RTX 5090 + CPU offload) — https://graphometer.ai/glm53-flash/

[32] Hugging Face discussion — 2× DGX Spark vLLM config, 20–23 tok/s with MTP — https://huggingface.co/zai-org/GLM-5.3-Flash/discussions/52

[33] Apidog — What is GLM-5.3-Flash? (specs, Artificial Analysis speed, list price) — https://apidog.com/blog/glm-5-3-flash-what-is/

[34] Ollama docs — Claude Code integration (Anthropic-compatible API, 64K+ context) — https://docs.ollama.com/integrations/claude-code

[35] Ollama library — glm-5.3-flash (cloud tag only) — https://ollama.com/library/glm-5.3-flash

[36] Cursor forum — local LLM needs a public HTTPS endpoint (staff answer) — https://forum.cursor.com/t/how-can-i-use-a-local-llm-on-my-desktop-ai-computer/152419

[37] ngrok docs — Use a local LLM in Cursor with Ollama — https://ngrok.com/docs/gateway/examples/cursor-local-llm

[38] Cline docs — Local models (Ollama, LM Studio, compact prompt) — https://docs.cline.bot/running-models-locally/overview

[39] OpenHands docs — Run local LLMs (LM Studio, Ollama, vLLM, SGLang) — https://docs.openhands.dev/openhands/usage/llms/local-llms

[40] Hugging Face — mlx-community/GLM-5.3-Flash-4bit (190 GiB; mlx-lm lacks glm5_next, use mlx-vlm) — https://huggingface.co/mlx-community/GLM-5.3-Flash-4bit

[41] Hugging Face — zai-org/GLM-5.3-Flash model card (supported serving frameworks) — https://huggingface.co/zai-org/GLM-5.3-Flash

[42] U.S. EIA — Electric Power Monthly, Table 5.3 (average residential price, 2026) — https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_03

[43] oMLX Community Benchmark — GLM-5.3-Flash-MLX-8bit on M5 Ultra (80c) 512 GB, 1K–128K session, oMLX v0.8.0rc1, no speculative decoding (Nov 14, 2026) — https://omlx.ai/benchmarks/performance/2qyzw3l3

[44] oMLX Community Benchmark — GLM-5.3-Flash-MLX-8bit on M5 Ultra (80c) 512 GB, 1K–16K session with batching (Nov 14, 2026) — https://omlx.ai/benchmarks/performance/0irgrjqs

[45] oMLX Community Benchmark — GLM-5.3-Flash-oQ4e-mtp on M5 Ultra (80c) 512 GB, 16K–195K, MTP on (submitted Nov 15, 2026, 8:30 PM ET / Nov 16, 01:30 GMT) — https://omlx.ai/benchmarks/performance/pdy8azbq

[46] oMLX Community Benchmark — GLM-5.3-Flash-oQ4e-mtp on M5 Ultra (80c) 512 GB, 8K–32K, MTP on (submitted Nov 15, 2026, 9:04 PM ET / Nov 16, 02:04 GMT) — https://omlx.ai/benchmarks/performance/6neo79pc

[47] oMLX — product page (OpenAI and Anthropic /v1/messages endpoints) — https://omlx.ai/

[48] oMLX source — Claude Code integration (`omlx launch claude`) — https://github.com/jundot/omlx/blob/a86a5587/omlx/integrations/claude.py

[49] jacar.es — oMLX on Mac: install, tune and benchmark a local LLM server (Homebrew install, wired limit, SSH tunnel) — https://jacar.es/en/how-to-install-and-tune-omlx-on-m5-max-128-gb/

[50] Hugging Face — pipenetwork/GLM-5.3-Flash-MLX-8bit (334.1 GB on disk, perplexity vs 8-bit, runtime fixes) — https://huggingface.co/pipenetwork/GLM-5.3-Flash-MLX-8bit

[51] Hugging Face — orcarouter/GLM-5.3-Flash-MLX (2- to 6-bit OrcaSAQ builds; 6-bit ≈296 GB; runs under mlx-vlm) — https://huggingface.co/orcarouter/GLM-5.3-Flash-MLX

[52] Anthropic — Claude pricing (API list prices; Claude Opus 5: $5 input / $25 output per million tokens, listed under legacy models), read Nov 16, 2026 — https://www.anthropic.com/pricing

All pages accessed Nov 14–16, 2026 (America/New_York); the benchmark, price and oMLX sources behind the main figures were re-checked on Nov 16, 2026. oMLX dates are the site's submission times, converted to America/New_York.

結

SOFTWARE SUSHI · PRIVATE AI, BUILT ON YOUR HARDWARE

Run private coding agents on *your own* AWS account or hardware.

The figures in this report are public and sourced. If you want GLM-5.3-Flash or another open model serving your team's coding agents on a Mac Studio, your GPUs or your AWS account, Software Sushi can plan and deploy it with you.

Book a discovery call →

AI/ML consulting

Software Sushi on AWS Marketplace

READ THE INTERACTIVE ARTICLE

softwaresushi.com/blog-details/best-coding-model-m5-ultra-512gb

WEBSITE

softwaresushi.com

SOFTWARE SUSHI
SCALABLE AI SOLUTIONS

© 2026 Software Sushi · Research report, November 16, 2026