JULY 17 → AUGUST 17, 2026

9 MODELS. 31 DAYS.

Every open-weight release that mattered — and what it costs against Claude and GPT.

VALS AI · SWE-BENCH VERIFIED

AN OPEN MODEL IS NOW #2 ON EARTH

96.4

deepseek-v4-pro — 0.6 points behind Claude Opus 5, at a twelfth of the price.

OPEN SOURCE NEWS YOU MISSED

August 2026

- kimi-k3
- qwen-3.8:27b
- laguna-s-2.1
- nemotron-3.5-lightning
- laguna-xs-2.1
- deepseek-v4-pro
- deepseek-v4-flash
- muse-glimmer
- ornith

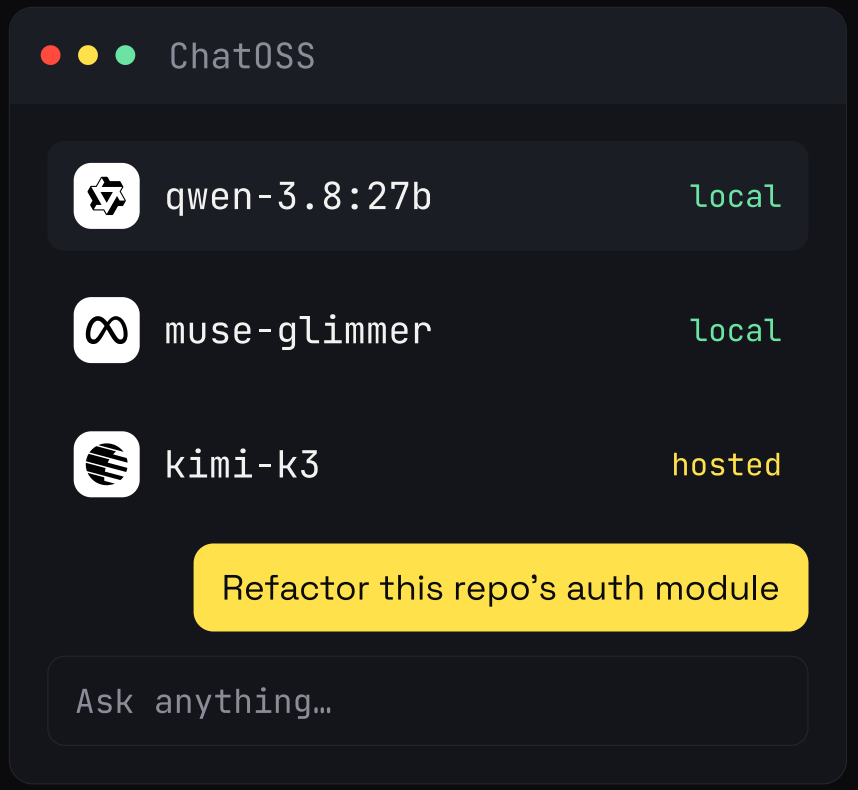
FOLLOW ALONG

TRY ALL NINE YOURSELF

Local weights or hosted routes, one window. No API keys.

ChatOSS.ai

Free · macOS · Windows · Linux



01 / 09 · MOONSHOT AI · JUL 16 SPECS

kimi-k3

The largest open-weight model to date — 1.56 TB of weights, native vision.

Text + vision | 1M context | Tool use

PARAMETERS 2.8T	ACTIVE / TOKEN 104B
MIXTURE OF EXPERTS 16 of 896	PRICE IN → OUT $2.80 → $14

1M context · cache read $0.30 · Kimi K3 License

01 / 09 · KIMI-K3

SCORES

TOP OPEN CODER

Independent tests in green, lab's own numbers in yellow · SWE-bench Pro not published

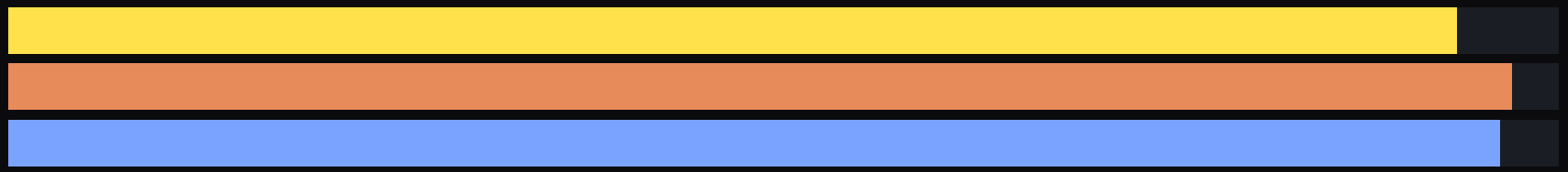
01 / 09 · KIMI-K3

VS CLAUDE & GPT

HALF THE PRICE, SAME SCORES

kimi-k3 · Jul 16 | Claude Opus 5 · Jul 24 | GPT-5.6 Sol · Jul 9

SWE-BENCH VERIFIED — TESTED BY VALS AI, EXCEPT SOL (OPENAI'S FIGURE)



OUTPUT PRICE PER 1M TOKENS

Kimi K3 and Opus 5 tested by Vals AI; Sol's 96.2 is OpenAI's own figure · Terminal-Bench 2.1: K3 88.3, Sol 88.8

02 / 09 · DEEPSEEK · AUG 13

SPECS

deepseek-v4-pro

GA build of the 1.6T flagship. MIT weights, no vision, cache reads at a third of a cent.

Text only | 1M context | 3 reasoning levels

PARAMETERS 1.6T	ACTIVE / TOKEN 49B
ARCHITECTURE MoE	PRICE IN → OUT $0.66 → $1.98

1M context · cache $0.0036 · peak rate $1.32 → $3.96

#2 ON VERIFIED, LAST ON TERMINAL

Same model, same test, two testers — a 33-point gap. Always ask who ran it.

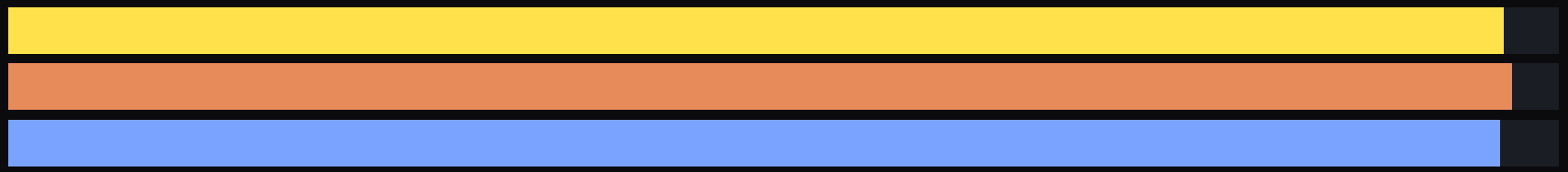
02 / 09 · DEEPSEEK-V4-PRO

VS CLAUDE & GPT

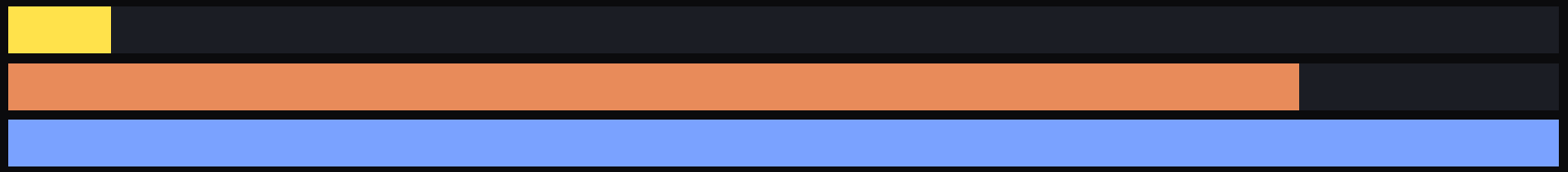
A TWELFTH OF THE PRICE

deepseek-v4-pro · Aug 13 | Claude Opus 5 · Jul 24 | GPT-5.6 Sol · Jul 9

SWE-BENCH VERIFIED — TESTED BY VALS AI, EXCEPT SOL (OPENAI'S FIGURE)



OUTPUT PRICE PER 1M TOKENS



Both closed models still beat it on agentic coding — SWE-bench Pro: Opus 5 79.2, Sol 64.6

03 / 09 · ALIBABA QWEN · AUG 14 SPECS

qwen-3.8:27b

Apache 2.0, text + image + video in, and it fits on one 24GB GPU after quantization.

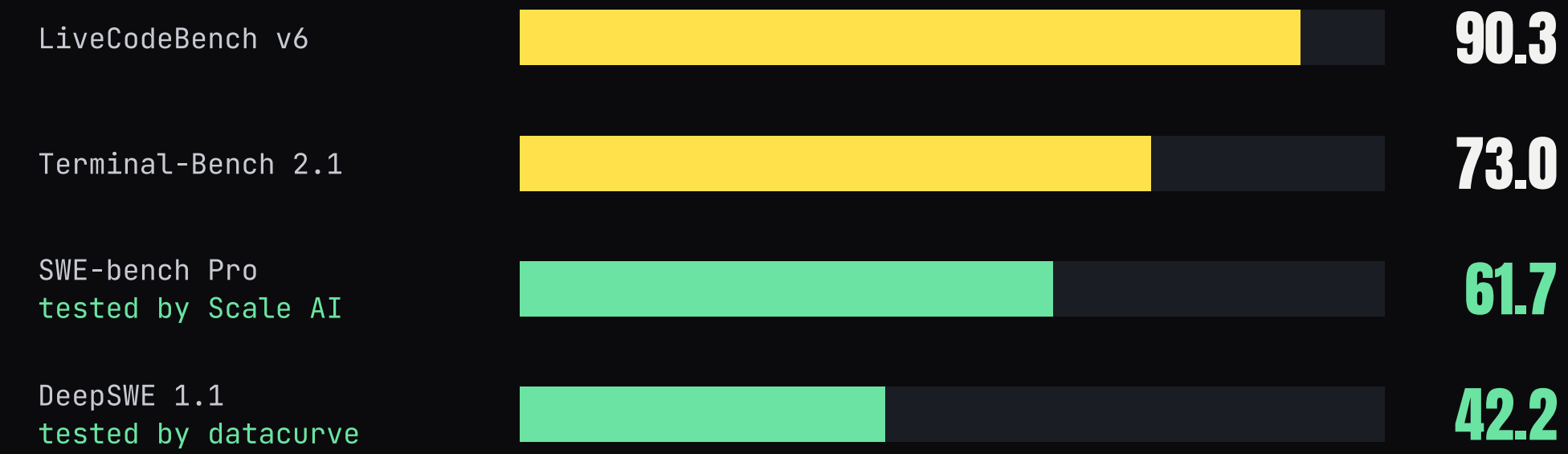
Text + image + video | 262K context | Computer use

PARAMETERS 27.8B	ACTIVE / TOKEN all of it
ARCHITECTURE Dense	PRICE IN → OUT $0.40 → $3

262K context · Apache 2.0 · free on your own GPU

SCORES

FRONTIER SCORES ON ONE GPU

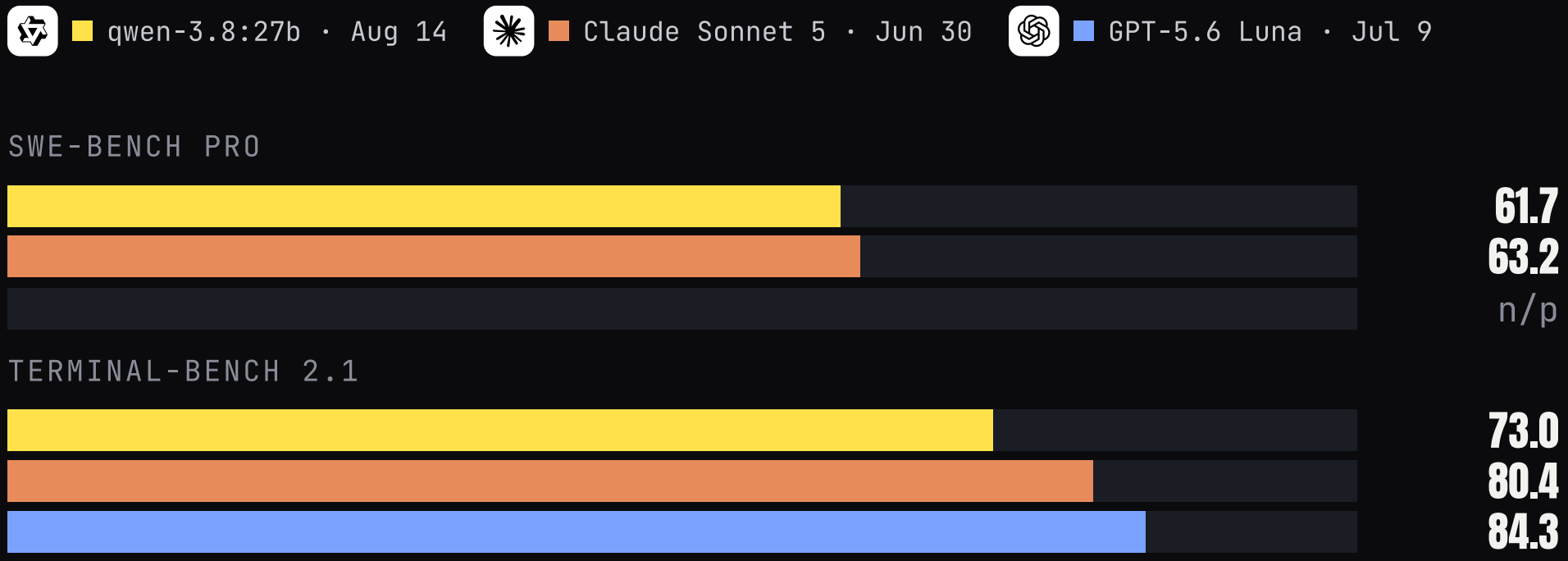


Independent tests in green · DeepSWE score tripled over Qwen3.6-27B

03 / 09 · QWEN-3.8:27B

VS CLAUDE & GPT

1.5 POINTS OFF SONNET 5



Output per 1M: Qwen $3 · Sonnet 5 $10 · Luna $1.20 — or $0 running local

deepseek-v4-flash

Same weights as the preview, re-post-trained. Six of 256 experts fire per token.

Text only | 1M context | Tool use

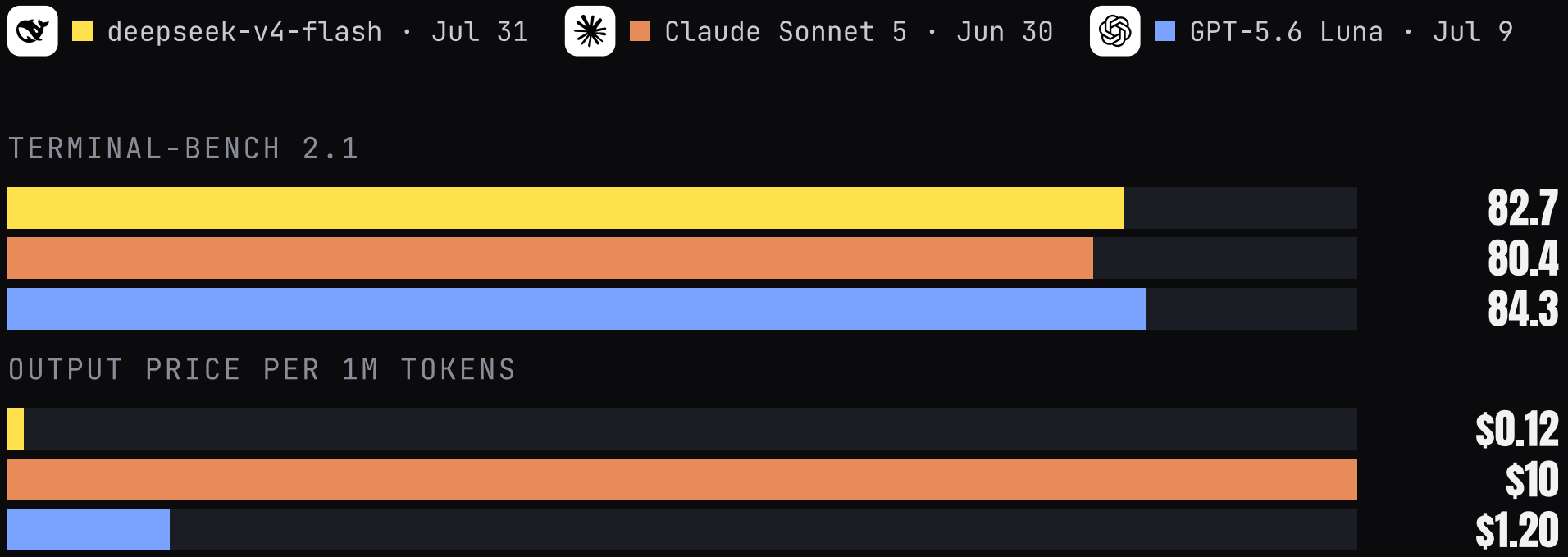
PARAMETERS 284B	ACTIVE / TOKEN 13B
ARCHITECTURE MoE	PRICE IN → OUT $0.06 → $0.12

1M context · MIT · GGUF available · 156GB RAM floor

04 / 09 · DEEPSEEK-V4-FLASH

VS CLAUDE & GPT

BEATS SONNET 5, 80× CHEAPER



DeepSWE 54.4 · AA Intelligence 50 — above the far larger V4 Pro

SPECS

laguna-s-2.1

Coding-first MoE with thinking interleaved between tool calls — 6.8% of the model fires per token.

Text only | 1M context | Thinking toggle

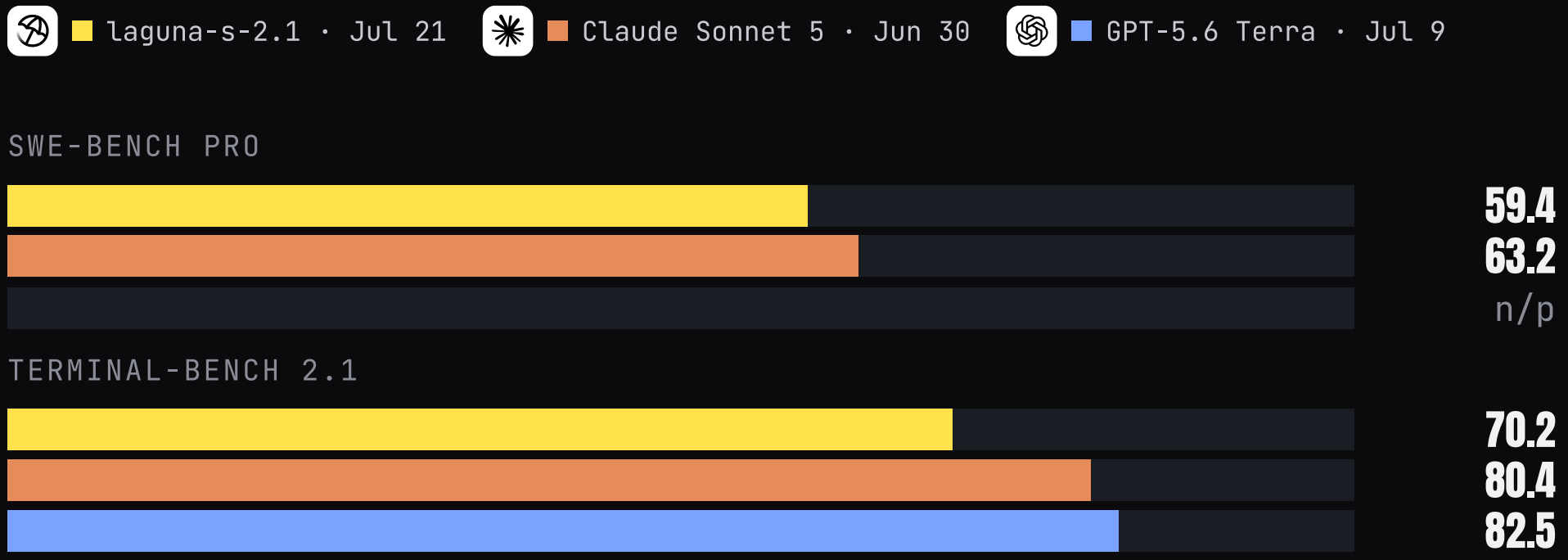
PARAMETERS 118B	ACTIVE / TOKEN 8B
ARCHITECTURE MoE	PRICE IN → OUT $0.09 → $0.18

1M context · OpenMDW-1.1 · free 256K tier on OpenRouter

05 / 09 · LAGUNA-S-2.1

VS CLAUDE & GPT

CLOSE ON CODE, 55× CHEAPER



Terminal-Bench drops to 60.4 with thinking off · output $0.18 vs $10 and $12

06 / 09 · META · AUG 10

SPECS

muse-glimmer

Meta's first Apache 2.0 model since Llama — dense, agent-first, distilled from Muse Spark.

Text + vision | 128K context | Tool use

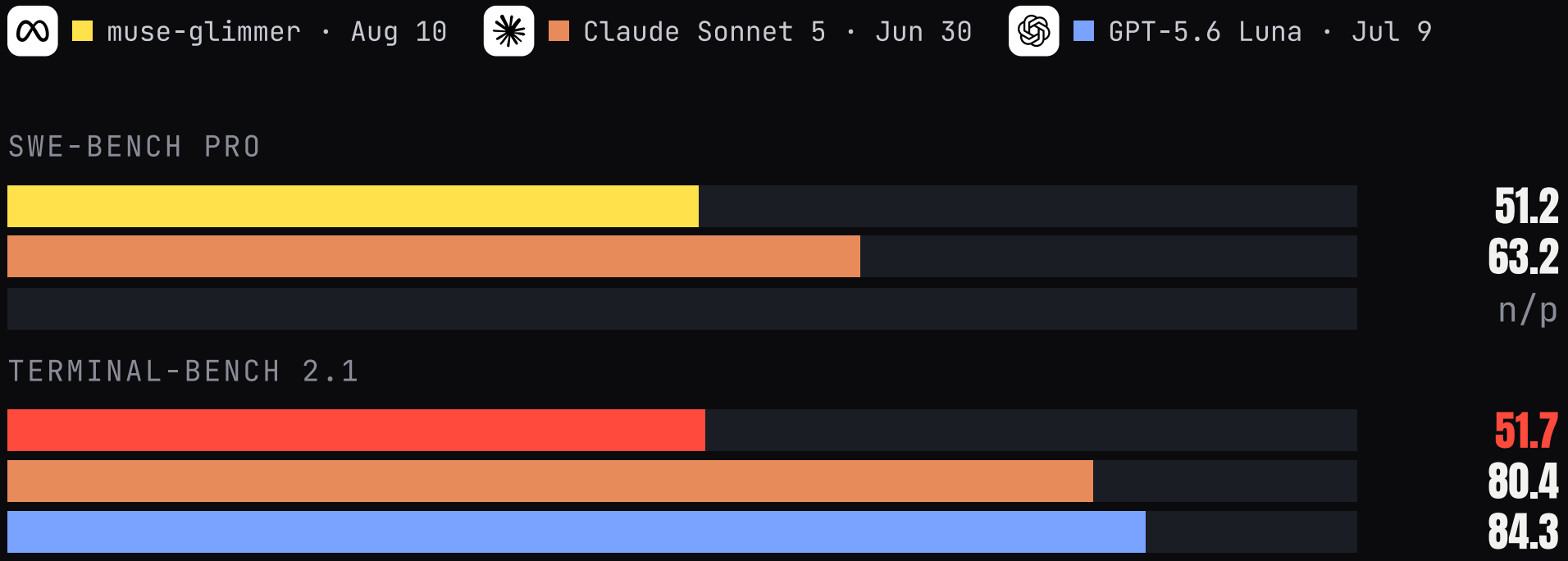
PARAMETERS **30B**	ACTIVE / TOKEN **all of it**
ARCHITECTURE **Dense**	PRICE IN → OUT **$0.30 → $1.20**

128K context · Apache 2.0 · ~17GB at 4-bit · runs on a 64GB Mac

06 / 09 · MUSE-GLIMMER

VS CLAUDE & GPT

GREAT AGENT, WEAK TERMINAL



SWE-bench Verified 76.0 · AIME 94.7 · MCP-Atlas 75.5 — the agent suites are where it wins

07 / 09 · NVIDIA · AUG 11

SPECS

nemotron-3.5-lightning

Weights, data and recipes all open. Built for tool calls and retries — up to 4× the output speed of its size class.

Text only | 1M context | Built for tool calls

PARAMETERS	30B
ACTIVE / TOKEN	3B
ARCHITECTURE	Mamba + MoE
PRICE IN → OUT	$0.08 → $0.20

1M context · OpenMDW-1.1 · free tier on OpenRouter · Jetson to DGX

07 / 09 · NEMOTRON-3.5-LIGHTNING

VS CLAUDE & GPT

CHEAP GRUNT WORK, NOT CODING

nemotron-3.5 · Aug 11 | Claude Sonnet 5 · Jun 30 | GPT-5.6 Luna · Jul 9

SWE-bench Verified 51.6 · PinchBench 85.4 · AA Intelligence 24, level with gpt-oss-120b

ORN ornith

Trained to write its own agent scaffold first, then solve inside it. Four MIT sizes, self-host only.

Text only | 256K context | Writes its own scaffold

FLAGSHIP	397B
ALSO SHIPS	35B · 31B · 9B
ACTIVE / TOKEN	not disclosed
PRICE	self-host

256K context · MIT · the 35B GGUF runs on 8GB VRAM

08 / 09 · ORNITH-1.0-397B

VS CLAUDE & GPT

BEATS OPUS 4.8 IN THE TERMINAL

ORN ornith-397b · Jun 25 | Claude Opus 4.8 · May 28 | GPT-5.6 Sol · Jul 9

All Ornith numbers self-reported, replication pending · closed pair costs $25–30 per 1M out

09 / 09 · POOLSIDE · JUL 2

SPECS

laguna-xs-2.1

The small sibling: 3B active, FP8 KV cache, and it fits on a 36GB Mac.

Text only | 256K context | Thinking toggle

PARAMETERS 33B	ACTIVE / TOKEN 3B
SWE-BENCH VERIFIED 70.9	PRICE IN → OUT $0.06 → $0.12

256K context · OpenMDW-1.1 · Ollama and llama.cpp

09 / 09 · LAGUNA-XS-2.1

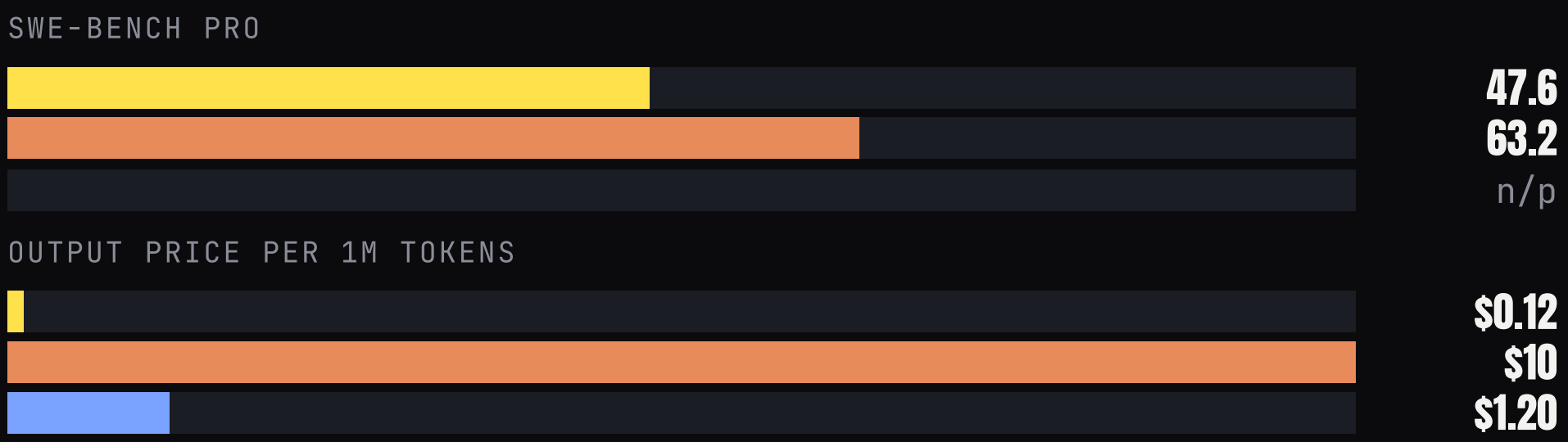
VS CLAUDE & GPT

75% OF SONNET, 1% OF THE COST

laguna-xs-2.1 · Jul 2

Claude Sonnet 5 · Jun 30

GPT-5.6 Luna · Jul 9



SWE-bench Multilingual 63.1 · Terminal-Bench 2.1 not published for this build

OUTPUT PRICE PER 1M TOKENS · CHEAPEST = 1×

THE WHOLE MONTH, ONE CHART

Bars show how many times more expensive than the cheapest · yellow = open weights · Ornith is self-host only

IF I HAD TO PICK THREE

WHAT I'D ACTUALLY RUN

qwen-3.8:27b

One GPU · 61.7 SWE-bench Pro on 24GB, Apache 2.0

deepseek-v4-flash

Agent fleets · 82.7 Terminal-Bench at $0.12 per 1M out

kimi-k3

Hardest work · 93.4 Verified on Vals AI, $0.30 cache read

All three run in ChatOSS — ChatOSS.ai

NEXT MONTH

ALREADY QUEUED UP

01 Meta says Muse Spark 1.2 weights are coming

02 DeepSeek's peak pricing lands in real bills

03 Independent replays of every August claim

SUBSCRIBE FOR SEPTEMBER

ChatOSS.ai