AI DAILY LEARNING · DAY 110 · JULY 16, 2026

The Open-Weight Counter-Punch

GLM-5.2, DeepSeek V4 & Qwen 3.6 Landed Days After Sonnet 5, Grok 4.5, and GPT-5.6 — Here's What the Free Weights Actually Buy You

GLM-5.2, DeepSeek V4, and Qwen 3.6 all shipped inside the same ten-day window as Sonnet 5, Grok 4.5, and GPT-5.6 — and on paper, the open-weight side is winning: cheaper tokens, permissive licenses, benchmarks that beat or tie the paid flagships. The catch is that no open model wins the same benchmark twice. GLM-5.2 owns long-horizon coding, DeepSeek V4 owns competitive coding and the lowest price, and Qwen 3.6 owns the single-GPU deployment case — which means "closing the gap" depends entirely on which gap you mean.

Day 109 closed with a promise: the open-weight counter-punch. Today we find out what it's actually worth. 110 days into this series we've watched frontier capability compress toward a near-tie (Day 80), watched three governments converge on the same companion-AI guardrails independently (Day 108), and watched three labs reprice within the same ten days (Day 109); today the axis flips from what the closed labs charge to what the open ones can actually do at a fraction of that price — and where they still can't.

1/6	86x	744B / 40B	2x
GLM-5.2's cost to match or beat GPT-5.5 on long-horizon coding benchmarks	How much cheaper DeepSeek V4 Pro's output tokens are than Claude Opus 4.7's	GLM-5.2's total vs. active parameters per token (Mixture-of-Experts)	DeepSeek V4's new peak-hour price multiplier, live from its official mid-July launch

1) Three open-weight labs, three different bets

None of this month's open-weight releases is trying to be the same thing. Z.ai's **GLM-5.2** ships under an MIT license with roughly 744 billion total parameters and about 40 billion active per token — a Mixture-of-Experts design built specifically for multi-hour, open-ended engineering work. On FrontierSWE, the benchmark that tracks long-running coding projects rather than single-shot problems, GLM-5.2 scores 74.4 against DeepSeek V4 Pro's 29.0, and independent tests have it matching or beating GPT-5.5 on several coding benchmarks at roughly one-sixth the cost. **DeepSeek V4 Pro** takes the opposite bet: it dominates competitive, single-shot coding

— LiveCodeBench 93.5% (the global #1), Codeforces 3206 — and ties Claude Opus 4.7 on SWE-bench Verified, all at $0.435 in / $0.87 out per million tokens, a fraction of GLM-5.2's own already-cheap pricing. **Qwen 3.6**, Alibaba's actually-open release (as opposed to the closed Qwen 3.7 Max), skips the leaderboard chase entirely — it's a compact MoE built to run on a single GPU with strong tool-calling and vision, aimed at teams who need to self-host rather than teams chasing a benchmark screenshot.

2) Where the gap is closing — and where it isn't

“Open-weight is catching up” is true and also the wrong frame. None of the three touches the closed frontier — Opus 4.8, GPT-5.6's Sol, Grok 4.5 — on hard multi-step reasoning or factual recall; DeepSeek V4 explicitly loses ground on GPQA Diamond and SimpleQA-Verified, the benchmarks that measure exactly that. What open weights have done is eat the middle. GLM-5.2's long-horizon coding score and DeepSeek V4's competitive-coding score both land in the same territory OpenAI's Terra, Anthropic's Sonnet 5, and xAI's Grok 4.5 are fighting over — the high-volume, not-that-hard workloads that make up most real usage, which Day 109's market signal already flagged as the actual battleground. The gap isn't closing at the top; it's closing exactly where the price war is happening, which is also exactly where most production traffic actually lives. A team routing most of its calls to a mid-tier model already has three new, radically cheaper options to route to — they just don't all do the same job.

3) The catch: open weights don't mean free to run

“Open-weight” describes the license, not the electricity bill. GLM-5.2's 744 billion total parameters have to sit in GPU memory whether or not a given token activates all of them — the Mixture-of-Experts design only computes through about 40 billion of them per token, which is why it's cheaper to run than a dense model that size, but you still need enough VRAM to hold the whole thing before you save a cent. Self-hosting at that scale means racks, not a laptop; most teams end up paying a hosting provider anyway, which puts you back on someone else's price list. DeepSeek underlined the point itself: its official V4 launch this month introduces peak-hour pricing — double the rate during 9am–12pm and 2pm–6pm local time — the same surge logic a power utility uses, on a model whose entire pitch was being the cheap option. Open weights changed who can compete on capability. They didn't repeal the economics of running a 700-billion-parameter model.

THE ONE RULE

The gap that's closing is the one between mid-tier closed models and top-tier open ones — not the one between open models and the actual frontier. Pick GLM-5.2, DeepSeek V4, or Qwen 3.6 for the specific job each one wins, not because “open-weight” sounds like a category you've already evaluated.

Model	Price (in / out per 1M)	Best at	License
GLM-5.2 (Zhipu / Z.ai)	$1.40 / $4.40	Long-horizon coding (SWE-bench Pro 62.1%)	MIT (open weight)
DeepSeek V4 Pro	$0.435 / $0.87 (2x at peak hours)	Competitive coding (LiveCodeBench 93.5%)	Open weight
Qwen 3.6	Runs on 1 GPU	Tool-calling & vision, self-hosted deployment	Open weight
Claude Sonnet 5 (ref.)	$2 / $10 (intro, through Aug 31)	Agentic coding & tool use	Closed
GPT-5.6 Terra (ref.)	$2.50 / $15	High-volume business tasks	Closed

VIRAL / OPEN-SOURCE SPOTLIGHT

DeepSeek V4: the cheap model that just invented surge pricing

DeepSeek V4 has spent 2026 as the AI world's go-to answer to “what's the cheapest model that's actually good at coding” — V4 Pro's list price of $0.435 in / $0.87 out per million tokens is roughly 35x cheaper on input and 86x cheaper on output than Claude Opus 4.7, while still leading LiveCodeBench globally and tying Opus 4.7 on SWE-bench Verified. That reputation is why this month's detail landed oddly: DeepSeek's official V4 release, arriving mid-July after three months as a preview, ships with peak-hour API pricing — calls placed 9am–12pm or 2pm–6pm local time cost double the off-peak rate. It's the first time an open-weight lab has borrowed a pricing mechanic straight from electricity grids, and it's a quiet admission that even the cheapest model on the leaderboard has a real, variable cost to serve at scale. For any team that built a workflow around DeepSeek's flat, ultra-low price, the fix is boring but effective: batch what you can into the off-peak window, and stop assuming “open-weight” means “price never changes.”

$0.435 / $0.87	93.5%	2x
V4 Pro's off-peak list price per 1M tokens, in / out	V4 Pro's LiveCodeBench score, #1 globally	The new peak-hour surcharge, live from V4's official mid-July launch

MARKET SIGNAL

Day 108 tracked a UN-backed scientific panel warning that AI capability is advancing faster than anyone can guarantee its safety. This month's grading brought a number to that warning: the Future of Life Institute's 2026 AI Safety Index scored nine labs on risk management, transparency, and whether they keep their safety pledges, and not one cleared a B. Anthropic topped the table — with a C+. OpenAI and Google DeepMind scored a C. Meta scored a D+. xAI, DeepSeek, and Mistral all failed outright. Put next to today's topic, that's an uncomfortable pairing: the two labs setting the pace on open-weight price and performance, DeepSeek and xAI, are also the two flunking the industry's own safety scorecard. Cheap and fast is not the same axis as safe, and right now nothing in the pricing race is pricing that difference in.

PRACTICAL TAKEAWAYS

1. Match the model to the job, not the leaderboard

GLM-5.2 wins long-horizon, multi-file coding; DeepSeek V4 wins single-shot competitive coding at the lowest price; Qwen 3.6 wins when you need to self-host on modest hardware. Picking “the best open-weight model” off an aggregate score will hand you the wrong one for at least two of your three most common tasks.

2. Budget for peak-hour pricing before it becomes the norm

DeepSeek V4 just added a 2x peak-hour surcharge to a lab whose entire reputation was flat, rock-bottom pricing. If your workflow leans on “cheap and always available,” build in a batching or off-peak-scheduling option now, because the labs racing to the bottom on price are starting to claw margin back through timing instead.

3. Weigh the safety score, not just the price, when picking a vendor

The Future of Life Institute's index puts DeepSeek and xAI at a failing grade for governance and safety practice, even as they set the pace on cost and speed. If you're routing production traffic to whichever model is cheapest this week, know that you may also be routing it to whichever lab is least accountable this week.

TOMORROW

Tomorrow — Day 111: with six-plus price tiers now live across three closed labs and three open-weight ones, the natural next move isn't picking a model — it's picking a router. We look at the model-routing layer emerging to automate exactly the “cheapest-tier-that-clears-the-bar” decision this series keeps recommending by hand, and whether the router itself becomes the product these labs are actually competing to power.
