Anthropic's J-Space, Tencent's Hy3, and the Shift From Compute to Data

AI News Digest

2026-07-07

Anthropic's J-Space, Tencent's Hy3, and the Shift From Compute to Data

By AI News Digest • July 7, 2026

Anthropic unveiled a new interpretability result inside Claude, Tencent and Meituan added fresh momentum to the open-model and domestic-chip race, and new benchmarks showed agents improving on bounded tasks while still struggling with long workflows. A second theme ran through the day: AI's next constraints are increasingly being framed around data coverage and memory architecture, not just raw compute.

Anthropic put interpretability at the center of the day

A new J-space claim goes beyond output-level auditing

Anthropic said a new interpretability technique reveals a J-space inside Claude, an internal region it compares to a global workspace in neuroscience [1, 2]. The company said watching this space makes it possible to observe silent reasoning steps, hidden goals, and situational awareness even when those do not appear in the outward response [3, 4, 5].

> The J-space lets us read, audit, and shape what Claude is actively thinking about—useful tools for keeping models trustworthy as they grow more capable. [6]

Anthropic also said deleting the J-space preserves fluent language and fact recall but weakens multi-step reasoning [7]. It published a paper and a Neuronpedia interactive demo for open-weight models [6, 8].

Why it matters: This is a notable step from evaluating models only by outputs toward auditing and intervening on internal state [7, 6].

Open models and domestic stacks kept gaining ground

Tencent pushed a large open model into the market

Tencent Hunyuan released Hy3, a 295B MoE model it described as best in its size class and competitive with trillion-scale flagships. It positioned the model as reliable and affordable for agentic use cases, released it under Apache 2.0 for commercial use, and offered a free API for two weeks [9].

A Tencent researcher separately described the jump from Hy2 to Hy3 as a major step in reasoning, agentic capability, anti-hallucination, reliability, and product experience [10].

Why it matters: Another major lab is treating openness and commercial usability as part of its frontier-model strategy, not as a secondary release path [9].

China also logged a new domestic-chip milestone

ChinAI reported that Meituan released LongCat-2.0, describing it as the first trillion-parameter model trained entirely on a 50,000 Chinese-chip cluster [11]. Separately, JD said its Oxygen AI Item Center processes hundreds of millions of item updates per day on Huawei Ascend NPUs and supports a catalog with tens of billions of SKUs [12].

Why it matters: The domestic-chip story is not only about training claims anymore; it is also about operating large AI systems in production at national e-commerce scale [11, 12].

Agent benchmarks improved, but reliability is still uneven

Bounded digital tasks moved up fast

The Remote Labor Index rose from 2.5% in October 2025 to 16.1% in July 2026 on end-to-end freelance work, with GPT-5.5 at 6.3%, Opus 4.8 at 8.3%, and Fable 5 at 16.1% across tasks such as 3D design, video ads, and floor plans [12]. Separately, Fable wrote a GPU kernel that achieved an 18.71x speedup over an optimized PyTorch baseline on RTX PRO 6000 Blackwell, using a single cooperative kernel launch per decoded token [12].

Dario Amodei also said the updated Sonnet 3.5 now reaches about 50% on SWE-bench, versus roughly 3-4% earlier in the year [13].

Why it matters: Recent gains are showing up clearly on bounded, scoreable software and freelance tasks—not just in chat demos [12, 13].

Long-horizon computer use remains far from dependable

OSWorld 2.0 raised the difficulty bar to 108 long-horizon computer tasks with a median human completion time of 1.6 hours, spanning tools such as Slack, AWS, GitLab, and professional-service portals [12]. The strongest tested setting,

Claude Opus 4.8 with maximum thinking and batched tool calls, reached only 20.6% binary accuracy, with performance dropping sharply as tasks became longer and more stateful [12].

Why it matters: Agents are improving on narrow, economically useful work, but the reliability gap is still large once tasks stretch across many steps, hidden state, and changing requirements [12].

The bottleneck conversation is moving beyond raw compute

Data coverage is starting to look like the next hard constraint

New commentary argued that the field is moving from a compute-limited regime into a data-limited one as useful public internet text approaches exhaustion at roughly 300 trillion tokens and hard RL tasks also begin to run dry [14]. The same analysis estimated external data spending at roughly $7 billion per year today, rising past $100 billion annually by 2030, with model differentiation increasingly driven by exclusive data and custom RL tasks [14].

It further argued that pretraining plus RL may already be enough for much economically valuable work, making workflow coverage, edge cases, and tacit knowledge the new bottleneck [14].

Why it matters: If that framing is right, the next moat looks less like raw compute and more like access to proprietary or purpose-built data [14].

Memory architecture is becoming a competitive lever for inference

Ben Thompson said frontier labs and hyperscalers are intensely focused on lowering memory requirements and argued that agentic inference could push toward disaggregated memory systems, including standalone memory racks that offload context beyond HBM limits [15]. John Carmack made a similar hardware-side case, arguing that deterministic inference access patterns make hybrid flash/HBM systems plausible and noting that NAND flash is more than 100x cheaper per GB than HBM [16].

Why it matters: The scaling debate is broadening from training clusters to inference architecture and deployment economics [15, 16].

Sources

1. X post by @AnthropicAI
2. X post by @AnthropicAI
3. X post by @AnthropicAI
4. X post by @AnthropicAI
5. X post by @AnthropicAI

6. X post by @AnthropicAI
7. X post by @AnthropicAI
8. X post by @AnthropicAI
9. X post by @TencentHunyuan
10. X post by @ShunyuYao12
11. ChinAI #365: Around the Horn (26th episode)
12. Import AI 464: Fable writes GPU kernels; AI automation; and analog computation
13. #452 – Dario Amodei: Anthropic CEO on Claude, AGI & the Future of AI & Humanity
14. X post by @willdepue
15. Did Memory Makers Overplay Their Hand? | Sharp Tech with Ben Thompson
16. X post by @ID_AA_Carmack