China's AI Buildout, Copilot's Reset, and Fable 5's Return

AI High Signal Digest

2026-07-05

China's AI Buildout, Copilot's Reset, and Fable 5's Return

By AI High Signal Digest • July 5, 2026

China's AI infrastructure economics, Microsoft's Copilot consolidation, and Fable 5's return led the day. The brief also covers Sakana's ICML research, new multimodal product launches, Moonshot's research-first strategy, and Alibaba's reported Claude Code ban.

Top Stories

Why it matters: the biggest signals today were about infrastructure, distribution, and models proving themselves in live use.

- **China's AI compute push is scaling fast.** One market analysis said the national computing power network could attract **Rmb7tn** of investment in 2026, with roughly **Rmb2tn ($300bn)** of data-center spending over five years. A typical GW-scale campus is modeled as **50%+ inference**, but domestic chips still trail on performance: Nvidia is still seen holding **55%** overall share, and Huawei 910B/910C servers were said to produce only **1/6 to 1/3** of an H800's daily token output. [1]
- **Microsoft is merging consumer and enterprise Copilot into one app.** The August-targeted overhaul reportedly adds AI coding tools, paid AutoPilot agents, and add-ons like Copilot Cowork after cutting features customers were not using. Copilot had **20M paying users** by April, up from **15M** in January, but still trails ChatGPT's **50M+** paid subscribers. [2]
- **Fable 5 is back in public testing.** Arena said the model has returned to Battle Mode and Agent Mode and had previously ranked **#1** in Agent Arena; separate posts pointed to strong 3D-generation demos across 60+ hard tasks and one case where it chose propensity score matching in a

retention analysis without being asked. [3, 4, 5]

Research & Innovation

Why it matters: the strongest technical work focused on memory, efficiency, and better training data rather than just larger scale.

- **Sakana AI brought a broad ICML slate.** Its 11 papers span multi-agent coordination, sparse LLMs, test-time scaling, long-term memory, and agent benchmarks. Highlights included **FwPKM** at about **75%** 5-needle NIAH accuracy at 128K context, **Doc-to-LoRA** for internalizing documents into model weights, and **TwELL** sparse kernels with **20%+** speedups on billion-parameter models. [6, 7, 8, 9]
- **EfficientRollout targets RL's biggest time sink.** The paper says rollout generation consumes nearly **70%** of LLM RL training time; its quantized self-drafter, roofline-based switch, and adaptive draft length produced up to **19.6%** faster rollout generation and **12.7%** faster training steps. [10]
- **DolphinMath aims to make high-quality math data abundant.** QuixiAI released a generator for unlimited math problems from elementary to postgraduate level, with mechanically correct step-by-step solutions for pretraining, SFT, and RL. [11, 12]

Products & Launches

Why it matters: new launches are turning multimodal performance and retrieval infrastructure into usable tools.

- **Google launched Nano Banana 2 Lite and Gemini Omni Flash.** Nano Banana 2 Lite was priced at **$0.034 per 1K images** with four-second generation, while Gemini Omni Flash was priced at **$0.10 per second** for developer video generation and conversational editing. [13]
- **LlamaIndex released Index v2 for agentic retrieval.** It exposes retrieve, read, grep, and find APIs so agents can navigate evolving knowledge bases; the legal-kb reference app adds project-scoped knowledge bases, visual citations, version control, and data export. [14]
- **Dreamina Seedance 2.5 is coming to CapCut.** CapCut said it supports seamless generation and editing, up to **50 multimodal references**, and **30-second** scenes across web, desktop, and mobile. [15]

Industry Moves

Why it matters: labs are differentiating through go-to-market choices and infrastructure strategy as much as model quality.

- **Moonshot AI is staying research-first.** Its enterprise chief said Kimi will rely on partners for last-mile deployment instead of building a heavy

services team; the company is reportedly raising **$2B** at about a **$30B** valuation, expanding via AWS, and says its KV-cache hit rate is above **90%**. [16]
- **OpenAI's reported Jalapeño chip points to the silicon race.** Posts said OpenAI unveiled its first custom AI chip, while a claimed **nine-month** design-to-tape-out timeline drew skepticism; a separate comment argued that at OpenAI's scale, owning silicon is now necessary. [17]

Policy & Regulation

Why it matters: model-access restrictions are starting to shape internal software policy at major companies.

- **Alibaba is reportedly banning Claude Code at work starting July 10.** The company classified Anthropic's coding agent as high-risk software after reports it contained checks for China-linked users; Anthropic already bars Chinese companies and foreign entities they own from using its models, and Alibaba is directing staff to its own Qoder tool. [18]

Quick Takes

Why it matters: smaller updates still show where tooling, evals, and training methods are moving.

- A team said it distilled **2.3M** Claude Fable 5 reasoning traces into **Qwen3-4B** and open-sourced the weights. [19]
- Newer Claude models were reported to fail Pi's edit tool more often than older versions, especially on tasks close to but not exactly on training distribution. [20, 21]
- A practical MCP paper said tool-selection accuracy falls below **90%** after **10-30 tools**, while MCP itself adds little latency. [22]
- **OctoTools** was accepted as an **ACL 2026 Oral**, positioning its training-free tool-use framework for wider attention. [23]

Sources

1. X post by @pequityresearch
2. X post by @kimmonismus
3. X post by @arena
4. X post by @arena
5. X post by @_catwu
6. X post by @SakanaAILabs
7. X post by @SakanaAILabs
8. X post by @SakanaAILabs
9. X post by @SakanaAILabs
10. X post by @VukRosic99

11. X post by @QuixiAI
12. X post by @QuixiAI
13. X post by @dl_weekly
14. X post by @llama_index
15. X post by @capcutapp
16. X post by @TechBuzzChina
17. X post by @thursdai_pod
18. X post by @kimmonismus
19. X post by @waterloo_intern
20. X post by @mitsuhiko
21. X post by @dbreunig
22. X post by @TheTuringPost
23. X post by @lupantech