Sakana Pushes Orchestration, Adobe Proves AI Monetization, and Apple Details AFM 3

AI High Signal Digest

2026-06-22

Sakana Pushes Orchestration, Adobe Proves AI Monetization, and Apple Details AFM 3

By AI High Signal Digest • June 22, 2026

Sakana made a strong case for model orchestration as a frontier layer, Adobe showed rare AI revenue scale with healthy margins, and Apple outlined the model architecture now powering AI across its devices and cloud. The brief also covers security implications, training-system advances, and new agent products.

Top Stories

Why it matters: today's clearest signals were about where frontier capability is moving—into orchestration layers, platform-scale monetization, and security-sensitive use cases.

- **Sakana launched Fugu, a multi-agent orchestration system exposed through a single model API.** The company says **Fugu Ultra** matches Fable and Mythos performance while avoiding export-control risk, and says the system works by dynamically routing across a swappable pool of models rather than relying on one frontier model [1, 2]. That makes this more than a model release: it is a bet that orchestration itself is becoming a core frontier layer.
- **Adobe posted one of the strongest AI monetization readouts in software.** Q2 revenue reached **$6.62B** with **36%** net margins, while AI-first ARR tripled year over year to more than **$500M**. Firefly alone reached **$300M** ARR with roughly **50%** QoQ growth, Acrobat AI Assistant paid users grew more than **150%**, and freemium MAUs rose to **850M** from **700M** a year ago [3, 4, 5]. Adobe said it is absorbing GenAI compute costs while expanding profitability [3].
- **A widely shared claim about Anthropic's Mythos sharpened the AI-security debate.** Mark Warner said NSA/Cyber Command leader-

ship told him Mythos "broke into almost all of our classified systems, not in weeks, but in hours," but the Economist author who relayed the quote later said it should not be read literally and likely depended on Mythos being used with other tools under particular conditions [6, 7, 8]. Even with that caveat, the reaction centered on a broader point: AI attackers bring effectively unlimited time and patience, which some argue means companies will need offensive agents testing their own systems continuously [9].

Research & Innovation

Why it matters: the most useful technical progress today came from better systems design, not just bigger base models.

- **Apple's AFM 3 shows how Apple is pushing capability under device constraints.** The new family includes five models of up to **20B** parameters for iPhones, Macs, and Apple's cloud [10]. One key technique stores most parameters in flash and activates only **1–4B** of the 20B for a task; another elastically scales the number of active experts with request difficulty [10]. Reported gains include text-to-speech quality rising from **3.87 to 4.15 MOS**, dictation wins of **44.7% vs 17.6%**, and **+10%** response satisfaction with **+14%** math performance for Cloud Pro over Cloud [10].
- **Huawei described a 6x Muon training speedup on Ascend clusters.** On a **512-card** setup training a **100B+ MoE** model, optimizer step time fell from **2700ms to 450ms** through redundancy removal across compute, communication, memory scheduling, and replica execution [11]. The post singled out DP de-redundancy, communication-free Muon for expert weights, matrix fusion, and replica de-redundancy as the main levers [11].
- **CMU's V-pretraining offered a smaller-data route to better reasoning.** The method uses a small labeled feedback set to train a task designer that shapes self-supervised targets, lifting Qwen2.5-0.5B's GSM8K Pass@1 from **22.20** to **29.60** without directly supervising the learner [12].

Products & Launches

Why it matters: new releases are increasingly aimed at concrete workflows in media generation, coding, and agent access.

- **MaineCoon** is a real-time audio-visual model focused on social interaction. Posts cited **22B** parameters, up to **47.5 FPS** on a single H100, cost below **$0.001/second**, and streaming generation for **1000s+ seconds** with continuous alignment across audio, motion, expression, and visuals [13, 14]. Its inference stack uses auxiliary models to manage cache and lookahead buffers [13]. Early access is at mainecoon.tech [14].

- **Seed 2.1 Pro Preview** ranked **#8** in Code Arena: Frontend with a score of **1539**, on par with Opus 4.6, and landed in the top 10 across five of seven subcategories. Public release is expected in a few weeks [15].
- **Sakana's Fugu** is live to try at sakana.ai/fugu [1].

Industry Moves

Why it matters: companies are pairing model strategy with domain distribution and purpose-built inference infrastructure.

- **Harvey is building a legal foundation model series** aimed at delivering frontier intelligence affordably and securely while letting firms and governments own specialized versions of their models [16]. Its agentic system is designed for long-running legal matters, with control over tools, sub-agents, and escalation to frontier models or human partners [16].
- **Together AI and 5C are deploying NVIDIA GB300 NVL72 systems** for inference and reasoning at scale, combining high-density compute, advanced cooling, and AI-optimized storage with Pegatron, Vertiv, and VAST Data [17, 18].

Policy & Regulation

Why it matters: access policy is becoming part of how labs govern frontier capability.

- **Anthropic is rolling out identity verification for "certain capabilities" through Persona** [19]. A related post said U.S. users are being asked for government ID to access Fable, alongside broader pressure for digital identity systems in the U.S., UK, and EU [20].

Quick Takes

Why it matters: these smaller items still point to near-term shifts in model releases and agent products.

- A **"claude-sonnet-5"** slug appeared on an Anthropic partner provider, hinting at a near-term release [21].
- **DeepSeek** has created a new **Harness** group for agentic products including a desktop agent app and CLI, and is hiring across research, engineering, and product [22, 23].
- **Codex** users are pushing multi-step testing loops that generate user stories, test them, fix issues, and re-test across hundreds of flows [24].
- **Nous Research's Hermes Agent** passed **1,500 GitHub contributors** [25].

Sources

1. X post by @SakanaAILabs
2. X post by @hardmaru
3. X post by @fchollet
4. X post by @fchollet
5. X post by @fchollet
6. X post by @kimmonismus
7. X post by @shashj
8. X post by @kimmonismus
9. X post by @scottastevenson
10. X post by @TheTuringPost
11. X post by @ZhihuFrontier
12. X post by @dl_weekly
13. X post by @fchollet
14. X post by @catnips_ai
15. X post by @arena
16. X post by @gabepereyra
17. X post by @togethercompute
18. X post by @5CGroupAI
19. X post by @0xIlyy
20. X post by @TomLikesRobots
21. X post by @synthwavedd
22. X post by @Lentils80
23. X post by @tianyi
24. X post by @tomosman
25. X post by @Teknium