Mythos Safeguards, DeepSeek Price Cuts, and Google's Agent Expansion

AI High Signal Digest

2026-05-23

Mythos Safeguards, DeepSeek Price Cuts, and Google's Agent Expansion

By AI High Signal Digest • May 23, 2026

Anthropic linked stronger cyber capabilities to tighter safeguards, DeepSeek pushed frontier pricing lower with a permanent V4 Pro cut, and Google expanded into persistent agents and conversational video. Also inside: efficient agent research, notable product launches in speech and security, and two meaningful policy moves.

Top Stories

Why it matters: today's biggest signals were tighter safety limits on frontier cyber models, faster price compression at the frontier, and broader consumer AI productization.

- **Anthropic tied frontier cyber capability to tighter release controls.** The company said Project Glasswing and its partners have already found more than 10,000 high- or critical-severity vulnerabilities in essential software, and warned that the software industry will need to adapt to the volume that models like Claude Mythos Preview can find [1, 2]. Anthropic also said it wants "far stronger safeguards" before a general release of Mythos-class models [3]. In external benchmark posts, Mythos was reported to beat GPT-5.5 on SWE-bench Pro, HLE, UK AISI cyber ranges, and exploit benchmarks [4].
- **DeepSeek made its V4 Pro price cut permanent.** First-party pricing is now $0.435 per 1M input tokens, $0.87 output, and $0.0036 cached input, with a blended price around $0.18 per 1M [5]. Artificial Analysis said the move puts V4 Pro on the Intelligence Index vs. cost Pareto frontier, with index cost around $268 versus $892 for Gemini 3.1 Pro Preview, $3,357 for GPT-5.5, and $5,117 for Claude Opus 4.7 [5].

- **Google expanded beyond chat into persistent agents and conversational video.** Gemini Spark is framed as a 24/7 personal AI agent for recurring tasks, new skills, and end-to-end workflows [6]. Gemini Omni lets users create and edit video through natural language, with custom avatars, multimodal inputs, and physics-aware scene consistency [7, 8, 9, 10, 11].

Research & Innovation

Why it matters: the most useful research updates were about lowering agent cost, improving training efficiency, and measuring model quality more honestly.

- **Agent workflows may be getting compiled into weights.** A paper highlighted by DAIR says full agentic workflows can be distilled into a model at roughly 100x lower inference cost while preserving near-frontier task quality [12].
- **Introspective X Training targets better performance per FLOP.** The method annotates data with model-generated critiques up front and reported up to 2.8x FLOP efficiency plus 5–10 point gains, especially in math and code, across training stages on 8B models [13].
- **A new eval paper argues loss is a weak proxy for reasoning quality.** "Forecasting Downstream Performance of LLMs With Proxy Metrics" says cross-entropy poorly predicts downstream reasoning performance and proposes proxy metrics over expert reasoning traces instead [14].

Products & Launches

Why it matters: new launches focused less on standalone chat and more on developer infrastructure, speech quality, and security tooling.

- **Google's Managed Agents + Interactions API** gives an agent a secure hosted Linux sandbox for code execution and memory management through a single API flow [15].
- **Cartesia's Sonic-3.5** took the top spot on the Artificial Analysis Speech Arena with Elo 1,218, ahead of Inworld Realtime TTS 1.5 Max and Gemini 3.1 Flash TTS; it supports 42 languages and 500+ voices [16].
- **Perplexity open-sourced Bumblebee**, a read-only scanner for macOS and Linux that checks developer machines for risky packages, extensions, and AI tool configs, and can trigger deeper scans when new supply-chain risks emerge [17, 18].

Industry Moves

Why it matters: capital and partnerships continue to flow toward labs and deployments that can turn model capability into durable distribution.

- **DeepSeek's financing round is advancing alongside an open-source stance.** Posts citing Bloomberg said the company is moving ahead with a roughly $10B financing round, while founder Liang Wenfeng told investors DeepSeek has no interest in short-term commercialization and will remain open-source [19, 20].
- **Google DeepMind expanded its partnership with Singapore** to help deploy AI at scale in scientific discovery, pandemic preparedness, and healthcare [21].
- **AI media production is moving into core operating budgets.** One major e-commerce company said it now generates 75% of its visual media with AI using Runway, reproducing projects that once cost $800K for under $10K and shifting a $5–6M annual production budget toward AI workflows [22].

Policy & Regulation

Why it matters: governments and regulators are starting to shape who can own AI assets and what AI marketing claims can survive scrutiny.

- **China halted Meta's planned acquisition of Manus**, signaling tighter state control over strategically important AI technology and disrupting the common startup strategy of relocating abroad for Western capital and partnerships [23].
- **The FTC settled active-listening AI marketing claims.** Cox Media Group and two other firms will pay nearly $1 million over allegations they deceived customers about an AI-powered ad-targeting service [24].

Quick Takes

Why it matters: a few smaller updates sharpened the picture on provenance, search efficiency, world models, and robotics.

- **SynthID** is expanding to more partners, and Google added new AI-content detection paths in Gemini App and Search [25].
- **MiniMax Agent** switched its default search from Serper to Perplexity, cutting tool calls 45%, token usage 42%, and total cost 27% while raising pass rate 2% [26].
- **Project Genie** can now turn Google Maps Street View locations into interactive worlds for eligible Google AI Ultra subscribers [27, 28].
- **Figure** said a recent logistics deployment ran continuously for 200 hours without failure [29].

Sources

1. X post by @AnthropicAI
2. X post by @AnthropicAI
3. X post by @scaling01
4. X post by @scaling01
5. X post by @ArtificialAnlys
6. X post by @Google
7. X post by @Google
8. X post by @Google
9. X post by @Google
10. X post by @Google
11. X post by @Google
12. X post by @dair_ai
13. X post by @rajammanabrolu
14. X post by @arkil_patel
15. X post by @_philschmid
16. X post by @ArtificialAnlys
17. X post by @perplexity_ai
18. X post by @perplexity_ai
19. X post by @kimmonismus
20. X post by @luluyilun
21. X post by @GoogleDeepMind
22. X post by @c_valenzuelab
23. X post by @DeepLearningAI
24. X post by @FTC
25. X post by @GoogleDeepMind
26. X post by @MiniMaxAgent
27. X post by @GoogleDeepMind
28. X post by @GoogleDeepMind
29. X post by @adcock_brett