GrandCode Tops Codeforces as Specialist Agent Models Advance and Web Risks Rise

AI High Signal Digest

2026-04-02

GrandCode Tops Codeforces as Specialist Agent Models Advance and Web Risks Rise

By AI High Signal Digest • April 2, 2026

GrandCode's live Codeforces wins, stronger specialist agent models such as Holo3 and GLM-5V-Turbo, and DeepMind's AI Agent Traps paper defined this cycle. The brief also covers new research on multi-agent systems and memory, fresh product releases, major industry moves, and policy-adjacent developments around sovereign AI and AI literacy.

Top Stories

Why it matters: The clearest signals this cycle were about agents getting better at coding and computer use, while the security risks around deploying them on the open web became much clearer.

GrandCode reached first place in live competitive programming

GrandCode ranked first in Codeforces Rounds 1087, 1088, and 1089, ahead of all human participants, including grandmasters [1]. The system is a Qwen-based multi-agent reinforcement learning stack that coordinates modules for hypothesis generation, solving, test generation, and summarization, then improves them with post-training and online test-time RL [1]. A comparison shared with the result shows how quickly this frontier has moved: OpenAI o3 was listed at 175th in April 2025, Gemini 3.1 Pro at 8th in February 2026, and GrandCode at 1st in March 2026 [1].

Impact: Competitive coding is increasingly becoming a live benchmark where agentic systems are posting top-tier results, not just strong offline demos [1].

Specialist agent models pushed deeper into computer use and visual coding

H Company launched Holo3, reporting 78.9% on OSWorld-Verified and claiming performance ahead of GPT-5.4 and Opus 4.6 at one-tenth the cost; weights are on Hugging Face and the API is live [2]. A separate model summary describes Holo3 as a Qwen3.5-based 35B A3B model with Transformers support and a free license [3]. Z.ai also released GLM-5V-Turbo, a vision-coding model that natively handles images, videos, design drafts, and document layouts, and can generate runnable code from screenshots and web interfaces [4, 5]. Z.ai says it leads benchmarks in multimodal coding, tool use, GUI agents, design-draft reconstruction, and visual code generation while keeping stable text-coding performance [6, 4, 7].

Impact: One analysis framed these releases as evidence that the agent stack is fragmenting into specialist layers for perception, planning, and execution rather than converging on a single general model [8].

DeepMind's AI Agent Traps paper reframed agent security

A new Google DeepMind paper introduces AI Agent Traps, a framework for adversarial content embedded in web pages and digital resources that targets autonomous agents [9]. The taxonomy covers six attack classes, including hidden instructions in HTML/CSS and memory attacks such as RAG poisoning and latent memory corruption [9]. The paper says hidden prompt injections can partially commandeer agents in up to 86% of scenarios, and latent memory poisoning can exceed 80% attack success with less than 0.1% data contamination [9].

> The attack surface is no longer just the model. It is every web page, every retrieved document, every piece of content the agent ingests at inference time. [9]

Impact: The security boundary for agents is shifting from model weights to the full environment they read and act on [9].

Research & Innovation

Why it matters: New research is focusing less on raw scale alone and more on how agents organize, remember, scale predictably, and act in the physical world.

Multi-agent design is getting both stronger and more constrained

One study found that self-organizing LLM agents spontaneously developed specialized roles and outperformed manually designed role assignments across 25,000 tasks with up to 256 agents [10]. It reports a 14% edge for sequential coordination over centralized approaches, more than 5,000 organically generated roles, and open-source models reaching 95% of closed-source quality at lower

cost [10]. A separate MIT theoretical result pulls in the opposite direction: when agents only subdivide shared context and do not receive new exogenous signals, delegated multi-agent planning is decision-theoretically dominated by a centralized Bayes decision maker with the same information [11]. That paper argues that splitting tasks across agents introduces irrecoverable information loss and that multi-agent setups help only when agents access genuinely different information sources [11].

Memory and context are becoming trainable subsystems

MemFactory proposes a unified framework that treats agent memory as a first-class, trainable component instead of separate storage, retrieval, and training systems [12]. It adds modular memory components, native GRPO integration for RL-based memory policy tuning, and reports up to 14.8% relative gains over baselines [12]. Separately, Baseten researchers built a 7M-parameter perceiver that compresses KV caches 8x while retaining 90%+ factual retention in a single forward pass, positioning it as an early step toward models that can extend working memory more efficiently [13].

Training science keeps getting tighter

The Delphi 1e23 run finished within 0.005 of a preregistered projected loss, even though the forecast was based on models more than 100x smaller at 3e20 FLOPs [14]. Related posts said Marin's scaling laws extrapolated at least 100x, though the run still showed loss spikes and bending curves that the team says it is trying to fix [15, 14]. Liquid AI's LFM2.5-350M, trained on 28T tokens with scaled RL, reported large jumps over LFM2-350M in instruction following, data extraction, and tool use [16]. A separate comment noted that this works out to roughly 100,000 tokens per parameter versus Chinchilla's cited optimum of 20 [17, 18].

Robotics agents are getting better evaluation loops

CaP-X was released as an open-source framework and benchmark for coding agents in robotics, where agents write code for perception and control, execute it on simulated and real robots, then iteratively improve reliability [19, 20]. The release includes a toolkit across perception, control, and visualization, a 187-task CaP-Gym benchmark, and CaP-RL results where a 7B open model improved from 20% to 72% success after 50 training iterations, with transfer to real robots [20].

Products & Launches

Why it matters: The shipping layer is moving quickly: labs are turning model advances into tools for coding, documentation, storage, and media workflows.

- **Arcee AI** released Trinity-Large-Thinking on the Arcee API with open weights on Hugging Face under Apache 2.0, aimed at developers and enterprises that want models they can inspect, post-train, host, distill, and own [21].
- **Claude Code** added NO_FLICKER mode, an experimental terminal renderer that Anthropic says most internal users already prefer and that supports mouse events; it is enabled with `CLAUDE_CODE_NO_FLICKER=1 claude` [22]. Claude Code is also available in the Claude mobile app on iOS and Android, with session handoff to the local CLI [23, 24].
- **OpenAI Codex** got a Linear plugin designed to keep the ticket and the work in sync [25].
- **Together AI** open-sourced 12 agent skills for Claude Code and Codex so coding agents can use Together's SDK patterns, model IDs, and API calls without copying docs by hand [26].
- **LangChain** embedded Chat LangChain directly in its docs, grounding answers in the full docs, knowledge base, and open-source code [27]. **Hugging Face** also launched Storage Buckets for Spaces, letting teams mount persistent storage volumes directly inside Spaces [28].
- **WAN 2.7-Image** is now available on fal with features including realistic faces, color-palette extraction, multilingual text rendering, and interactive editing [29].

Industry Moves

Why it matters: Capital, partnerships, and infrastructure constraints are shaping where AI can actually scale.

- **The Information** reported that **OpenRouter** is raising $120M at a $1.3B valuation led by CapitalG; the company gives developers access to 300+ models through one API and is reportedly already at $50M+ ARR [30].
- A **Business Insider** report said OpenAI's **Stagecraft** project uses 3,000-4,000 freelancers, paid at least $50 per hour, to create ChatGPT training materials across 439 occupations ranging from commercial pilots to HR specialists [31, 32]. The stated goal is to "map economically relevant tasks and evaluate the model's capabilities," and the work runs through Handshake AI [31].
- **Cohere** expanded its partnership with **EnsembleHP** to build what it describes as the healthcare industry's first revenue-cycle-management-native LLM, purpose-built for complex financial workflows in healthcare operations [33].
- A Bloomberg-linked note said half of US data centers planned for 2026 are expected to be delayed or canceled because of shortages in transformers, switchgear, and batteries, while US manufacturing capacity remains insufficient and imports are needed [34]. A separate macro view of the AI stack argued that after $350B in revenue growth, semis still capture 79%

of profits, infra 14%, and apps 7% [35].

Policy & Regulation

Why it matters: The policy-adjacent updates this cycle were less about new laws and more about sovereign AI cooperation, public trust, and AI literacy.

- **Sakana AI** signed an MoU with **French Current AI**, with the French AI Ambassador signing on France's behalf during a visit to Sakana AI [36]. The agreement covers international cooperation on the AI stack and contributions to the Global South, with the stated goal of helping establish a sovereign AI ecosystem alongside France and other partner countries [36].
- **Anthropic** published a report covering 80,508 Claude users across 159 countries and 70 languages on what people want from AI, what they have already gotten from it, and what they fear [37].
- **Google Research** expanded AI Quests, a gamified AI literacy experience built with the Stanford Accelerator for Learning, to eight additional languages including Spanish and Malay [38].

Quick Takes

Why it matters: These smaller signals help track where evaluation, retrieval, and developer tooling are moving next.

- **Arena** added Pareto frontier charts across Text, Vision, Search, Document, and Code leaderboards to show performance versus blended token price; on the current Text frontier, Google DeepMind had five models, with xAI and DeepSeek at two each [39].
- **Kaggle** introduced Standardized Agent Exams so agents can register for an exam, solve it, and join a leaderboard [40].
- **YC-Bench** was introduced as a benchmark for whether an agent can run a simulated startup over a one-year horizon spanning hundreds of turns [41, 42].
- **Tinker** added longer context windows for select models: 128k for Kimi K2.5 and GPT-OSS-120B, and 256k for Nemotron 3 Super 120B and Qwen3.5 397B [43].
- **Qdrant** reported that adding hard negatives to sparse-embedding training improved search relevance by 28% over BM25 on real benchmarks [44]. In a follow-up on specialization versus generalization, it reported 28% in-domain gains and 8-10% cross-domain gains, but failure out of domain because of overfitting [45].
- **SkyPilot** added native support for VAST Data storage so AI workloads can mount large datasets directly instead of waiting for data copying before training starts [46].

Sources

1. X post by @deep_reinforce
2. X post by @hcompany_ai
3. X post by @mervenoyann
4. X post by @Zai_org
5. X post by @Zai_org
6. X post by @Zai_org
7. X post by @Zai_org
8. X post by @LiorOnAI
9. X post by @omarsar0
10. X post by @dair_ai
11. X post by @omarsar0
12. X post by @omarsar0
13. X post by @baseten
14. X post by @percyliang
15. X post by @WilliamBarrHeld
16. X post by @liquidai
17. X post by @awnihannun
18. X post by @awnihannun
19. X post by @letian_fu
20. X post by @DrJimFan
21. X post by @arcee_ai
22. X post by @bcherny
23. X post by @bcherny
24. X post by @_catwu
25. X post by @OpenAIDevs
26. X post by @togethercompute
27. X post by @LangChain
28. X post by @_akhaliq
29. X post by @fal
30. X post by @steph_palazzolo
31. X post by @TheRundownAI
32. X post by @TheRundownAI
33. X post by @cohere
34. X post by @AkshatRathi
35. X post by @apoorv03
36. X post by @SakanaAILabs
37. X post by @dl_weekly
38. X post by @GoogleResearch
39. X post by @arena
40. X post by @osanseviero
41. X post by @arankomatsuzaki
42. X post by @arankomatsuzaki
43. X post by @tinkerapi
44. X post by @qdrant_engine

45. X post by @qdrant_engine
46. X post by @skypilot_org