# Distillation Gets Faster, Open Agents Self-Improve, and AI Supply Risks Turn Geopolitical

AI High Signal Digest

2026-04-13

## Distillation Gets Faster, Open Agents Self-Improve, and AI Supply Risks Turn Geopolitical

*By AI High Signal Digest • April 13, 2026*

This brief covers a major distillation advance in TRL, the emergence of self-improving open agents around Hermes, a medical-monitoring use case from Penn, and new signals on distribution, compliance, and compute risk across the AI industry.

### Top Stories

*Why it matters:* This cycle's biggest developments were about leverage and constraints: better ways to compress frontier models, open agents that iterate on themselves, a concrete medical-monitoring use case, and external limits on compute supply.

#### 1) Frontier distillation got materially more practical

TRL's rebuilt on-policy distillation trainer now supports teacher models above 100B parameters and is reported to run more than 40x faster than naive implementations through buffer and payload optimizations [1, 2]. In the cited example, the team distilled Qwen3-235B into a 4B student and gained 39+ points on AIME25 [2]. The same setup is presented as usable across Llama, Qwen, and Gemma model families [1]. Blog: TRL distillation trainer [2].

**Impact:** This is a concrete post-training efficiency story: very large teacher models can now be used to produce much smaller students with less engineering overhead than before [1, 2].

### 2) The open-agent stack is moving toward self-improvement

NousResearch's **hermes-agent-self-evolution** repo automates a loop where the system reviews past task histories, identifies errors and root causes, proposes prompt or code changes, tests variants, and keeps the best-performing version [3]. Today it focuses on optimizing **SKILL.md**, with a roadmap to extend the same loop to tool descriptions, system prompts, and underlying code [3]. The repo is API-only for mutation and evaluation, includes test gates, file-size and scope limits, and still requires human-reviewed PRs before merge [3]. Repo: hermes-agent-self-evolution [3].

Related activity around Hermes points the same direction: one contributor said Hermes was evolved into an autonomous ML/RL research lab with **94 Orchestra primitives** for zero-shot GRPO and distillation [4], while Teknium said more is coming [5]. Separately, a quality-of-life update made Hermes about **20% more likely** to load the right skill for a task, and Teknium said the agent performed the prompt improvement and benchmarking itself [6].

**Impact:** Open agents are being pushed beyond static tool wrappers toward systems that can revise parts of their own operating instructions under guardrails [3].

### 3) A Nature paper showed a strong real-world medical use case for LLMs

University of Pennsylvania researchers analyzed more than five years of Reddit discussion—about **400,000 posts** from **70,000 users**—about GLP-1 drugs including Ozempic and Mounjaro, using LLMs to map patient language to medical terminology [7]. The study surfaced side effects not fully reflected in current labels, including **menstrual irregularities, hot flashes, chills, and fatigue** [7]. The authors frame this as computational social listening and argue that mainstream drugs may need real-time social-media monitoring as an early warning system beyond voluntary FDA-style reporting [7]. Paper: Nature [8].

**Impact:** This is one of the clearest examples in this cycle of LLMs being used to expand surveillance capacity in a high-stakes domain rather than just improve a chatbot [7].

### 4) AI infrastructure risk is becoming geopolitical, not just technical

One analysis tied Middle East tensions to four AI supply-chain bottlenecks: helium, energy, shipping, and advanced chip concentration [9]. It said the strike on Qatar's Ras Laffan facility removed about a third of global helium supply, that helium is essential for sub-3nm fabrication, and that South Korea and Taiwan are highly exposed through suppliers such as SK Hynix and TSMC [9]. The same thread argued that Strait closures could disrupt **25% of world energy supply**, threaten Taiwan's power availability, and block about **30% of global container shipping** for semiconductor equipment, chemicals, and

finished chips [9]. It tied those risks directly to the roughly **$500B** AI datacenter buildout planned for 2026 [9].

**Impact:** This frames compute risk as a question of energy, logistics, and industrial gases—not only model quality and GPU counts [9].

## Research & Innovation

*Why it matters:* The research pipeline this cycle focused on new runtimes, cheaper arithmetic, and better long-context mechanics—not just bigger models.

- **Neural Computers:** Meta AI and KAUST proposed **Neural Computers**, a learned system where computation, memory, and I/O move into the model's runtime state [10]. Early prototypes use video-model-style prediction to simulate terminal and GUI behavior directly from instructions, pixels, and user actions rather than executing code in a conventional OS [10]. The authors and commentators are explicit that long-horizon reasoning, stable symbolic computation, and durable reuse are not solved yet, framing this as a first step toward a **Completely Neural Computer** [10]. Paper: arXiv 2604.06425 [11].
- **Huawei HiFloat4 FP4:** Huawei proposed **HiFloat4 (HiF4)** for Ascend NPUs, reporting BF16-comparable performance with large compute-efficiency gains [12]. A follow-on note summarized the key claim more precisely: about **90%** of training computation can run in FP4 while keeping the loss gap within roughly **1.5%** of a full-precision baseline [13]. The same note said chips designed around this are still more than **1.5 years away** [13]. Paper: arXiv 2604.08826 [14].
- **cuLA for linear attention: cuLA** packages handwritten CUDA kernels for linear attention variants including GLA, KDA, GDN, and Lightning, built with CuTe DSL and CUTLASS for Hopper and Blackwell GPUs [15]. The pitch is straightforward: linear attention gives **O(N)** scaling instead of **O(N²)**, enabling million-token settings, while cuLA reports speedups from **1.32x**–**1.45x** for KDA on Blackwell to **2.19x** for Lightning Attention against an FLA Triton baseline [15]. The migration path is also small—a one-line import swap from `fla.ops.kda` to `cula.kda` [15].
- **Retrieval models are back under debate:** A fresh thread argued that **ColBERTv2**—a **100M** late-interaction retriever—still beats **Qwen3-Embed-8B** on BrowseComp+ and LIMIT, despite the dense retriever being about **80x larger** [16, 17]. Supporters took that as evidence that single-vector dense models generalize poorly out of domain [17, 18]. Critics replied that LIMIT is artificial, that synthetic datasets can mislead, and that quantization often trades only **1–5%** recall for roughly **100x** memory reduction and major speed gains [19, 20, 21]. A separate observation from the same discussion is that retrieval papers are increasingly

built on decoder-only LLMs rather than encoder-only models like BERT [22].

## Products & Launches

*Why it matters:* New releases were less about generic chat and more about workflow control, evaluation, search, and document manipulation.

- **OpenClaw v2026.4.11** added ChatGPT import, a new **Memory Palace** for exploring chats as structured memory, guided plugin setup, richer chat rendering, better video generation handling, and stronger integrations with Teams, Feishu, WhatsApp, and Telegram [23]. Changelog: openclaw v2026.4.11 [23].
- **Opik** was highlighted as an open-source tool for debugging, evaluating, and monitoring LLM apps, RAG systems, and agent workflows with tracing, automated evals, and dashboards [24]. Repo: comet-ml/opik [24].
- **Qdrant Query Language (QQL)** introduced a SQL-like interface for vector search, combining hybrid retrieval, filtering, and semantics in a more structured query flow [25]. Article: QQL overview [25].
- **Verso AI** launched with an engine that converts PowerPoint's OOXML format into a representation that is easier for LLMs to edit; the team said its system beats competing tools on its internal benchmark [26].
- **Hermes Telegram Dashboard** was announced for Hermes users, bringing terminal access, system-resource monitoring, and cron management into Telegram [27].

## Industry Moves

*Why it matters:* Distribution, platform reach, and infrastructure economics continue to matter as much as model releases.

- **MiniMax M2.7 widened its distribution footprint.** Together AI said the model is live on its platform, with day-0 availability on both serverless and dedicated infrastructure [28, 29]. Fireworks separately announced day-0 hosted availability, describing the model as production-ready with **200K+ context**, strong reasoning, and native multi-agent support [30]. At the same time, separate discussion around the open-weight release focused on the **non-commercial license**, with one user saying the restriction prevents at-home use [31, 32].
- **Perplexity launched a company-building competition around Perplexity Computer.** The **Billion Dollar Build** is an eight-week program for teams building a company with a path to a **$1B** valuation, with finalists eligible for up to **$1M** from the Perplexity Fund and up to **$1M** in Computer credits [33].

- **Infrastructure remains a constrained market.** Notes from a HumanX fireside chat said the AI infra market is still early, that AI-native firms are using post-training to reach frontier quality more cheaply and quickly, and that inference remains constrained by both supply and talent [34].
- **Claude's consumer reach kept rising.** Similarweb reported **341.26% YoY** growth in Claude website traffic in Q1 [35]. One commentator attributed some of the appeal to Claude's less chatty, more concise responses [36].

## Policy & Regulation

*Why it matters:* This cycle had fewer formal government actions, but access control, compliance, and labor-policy debates continued to shape how AI systems are deployed.

- **Capability control is tightening around agent ecosystems.** A Zhihu Frontier roundup said Anthropic removed **OpenClaw** from the Claude whitelist, describing it as a sign of tighter control over agent ecosystems [37]. The same roundup said **Anthropic Mythos** was released only to select partners, framing that as a capability-control issue rather than a broad public rollout [37].
- **Compliance remains a live cross-border constraint.** The same roundup said Slack's shutdown in Greater China was presented as a compliance matter rather than an abrupt exit [37].
- **AI-risk activism remained under scrutiny after the attack on Sam Altman's home.** PauseAI said it unequivocally condemned the attack, said the suspect had only minimal participation in its public Discord, and stated he had no role in the group's campaigns or events [38]. The organization also argued that concern about advanced AI risk is shared by major researchers, legislators, and institutions, and positioned its own work around protests, petitions, and policy advocacy as a peaceful outlet [38].
- **The labor-policy conversation is widening.** Sam Altman was cited as proposing a **4-day, 32-hour workweek** and a *new social contract* in response to AI and robotics [39]. At the same time, Alex Karp argued AI will destroy humanities jobs and increase the importance of vocational training and hands-on skills [40], while Aaron Levie argued the opposite dynamic may hold in law, predicting more lawyers because AI will generate more legal questions and new areas of compliance work [41].

## Quick Takes

*Why it matters:* Smaller items still help map where capability, demand, and uncertainty are moving next.*

- **Unitree R1** is now available for global pre-order on AliExpress starting at **$6,806**, with June 30 deliveries for the base **R1 AIR** trim; the EDU version adds custom software support, full SDK access, and optional Jetson Orin [42].
- **Muse Spark** kept attracting favorable external reactions. Meta said it rebuilt its AI stack from scratch and that Muse Spark now powers Meta AI [43]. François Fleuret said the model passed his tests, including image generation [44, 45], and Alexandr Wang said it is particularly good at finding open-source data and analyzing it [46].
- **Claude Opus 4.6 sparked another benchmark dispute.** BridgeBench claimed its hallucination accuracy fell from **83.3%** to **68.3%** and described the model as nerfed [47]. Critics replied that the newer run used **30 tasks** instead of **6**, and said shared-task scores changed only from **87.6%** to **85.4%**, which they argued could be statistical noise [48]. Separate commentary noted that inference at scale can legitimately create performance variability across services [49].
- **LangChain clarified its agent-SDK stack.** Harrison Chase described `create-agent` as minimal, **deepagents** as more batteries-included, and middleware as the advanced customization layer across both [50]. In a related exchange, he pointed to **deepagents** when asked for a strong open-source alternative to the Claude Agent SDK [51, 52].
- **Open-source AI security tooling got a fresh checklist.** A roundup highlighted **NeMo Guardrails, Promptfoo, LLM Guard, garak, DeepTeam, Llama Prompt Guard 2-86M, ShieldGemma 2, OpenGuardrails, Cupcake,** and **CyberSecEval 3 – Visual Prompt Injection** as lightweight alternatives while Mythos remains closed [53].

---

**Sources**


1. X post by @_lewtun
2. X post by @cmpatino_
3. X post by @Jason23818126
4. X post by @ruffy0369
5. X post by @Teknium
6. X post by @Teknium
7. X post by @TheRundownAI
8. X post by @TheRundownAI
9. X post by @kimmonismus

10. X post by @TheTuringPost
11. X post by @TheTuringPost
12. X post by @arankomatsuzaki
13. X post by @teortaxesTex
14. X post by @arankomatsuzaki
15. X post by @ZhihuFrontier
16. X post by @lateinteraction
17. X post by @lateinteraction
18. X post by @lateinteraction
19. X post by @gabriberton
20. X post by @gabriberton
21. X post by @gabriberton
22. X post by @gabriberton
23. X post by @TheTuringPost
24. X post by @dl_weekly
25. X post by @qdrant_engine
26. X post by @AymericRoucher
27. X post by @mr_r0b0t
28. X post by @togethercompute
29. X post by @MiniMax_AI
30. X post by @FireworksAI_HQ
31. X post by @xeophon
32. X post by @TheZachMueller
33. X post by @perplexity_ai
34. X post by @saranormous
35. X post by @Similarweb
36. X post by @kimmonismus
37. X post by @ZhihuFrontier
38. X post by @PauseAI
39. X post by @kimmonismus
40. X post by @kimmonismus
41. X post by @levie
42. X post by @TheHumanoidHub
43. X post by @alexandr_wang
44. X post by @francoisfleuret
45. X post by @francoisfleuret
46. X post by @alexandr_wang
47. X post by @bridgemindai
48. X post by @paul_cal
49. X post by @dbreunig
50. X post by @hwchase17
51. X post by @matt_ambrogi
52. X post by @hwchase17
53. X post by @TheTuringPost