Autonomous Research Advances as Anthropic Pushes into Biotech and Rethinks Agent Access

AI High Signal Digest

2026-04-06

Autonomous Research Advances as Anthropic Pushes into Biotech and Rethinks Agent Access

By AI High Signal Digest • April 6, 2026

Autonomous research systems, Anthropic's biotech acquisition, and tighter controls on agent compute dominated this cycle. The brief also covers Gemma 4's spread into local developer workflows, new long-context research, evolving agent infrastructure, and policy moves in China and Maine.

Top Stories

Why it matters: This cycle's biggest signals were about autonomous research, vertical expansion into biotech, the economics of agent usage, wider local-model distribution, and early labor-market measurements.

ASI-Evolve claims end-to-end autonomous AI research

Shanghai Jiao Tong University researchers released **ASI-Evolve**, an open-sourced system described as running the full AI research loop itself: reading papers, forming hypotheses, designing and running experiments, analyzing results, and iterating without human intervention [1]. In neural architecture search, it ran **1,773 rounds**, generated **1,350** candidates, and produced **105** models that beat the best human-designed baseline; the top model exceeded DeltaNet by **+0.97 points** [1]. The same framework reportedly improved data curation by **+3.96** average benchmark points and **+18** on MMLU, and produced RL algorithms that beat **GRPO** by up to **+12.5** on competition math [1].

> "This is the first system to demonstrate AI-driven discovery across all three foundational components of AI development in a single framework." [1]

A biomedicine test also showed **+6.94** points in drug-target prediction on unseen drugs [1]. One critic argued the work is not the first effort of its kind and said frontier labs still rely on data intuitions that may not be offloaded to a scaffold [2].

Impact: The paper presents this as a single framework improving **architecture, data, and algorithms** rather than optimizing only one part of the stack [1].

Anthropic acquires Coefficient Bio for biotech workflows

Anthropic acquired **Coefficient Bio** for about **$400M**. The sub-10-person startup builds AI to plan drug R&D, manage clinical regulatory strategy, and identify new drug opportunities [3]. The team joins Anthropic's healthcare and life sciences group, which already works with Sanofi, Novo Nordisk, AbbVie, and others [3].

> "I'm talking about using AI to perform, direct, and improve upon nearly everything biologists do." [3]

Posts around the deal frame it as execution on Dario Amodei's **"virtual biologist"** idea, with Coefficient Bio covering drug discovery, clinical trials, and regulatory submissions end to end [3].

Impact: The deal pushes Anthropic further from being only a general-model vendor and deeper into healthcare-specific workflows [3].

Claude access rules now reflect harness economics

A notice said **Claude subscriptions** will no longer cover usage on third-party tools like **OpenClaw**, though users can still buy extra usage bundles or use a Claude API key [4]. A later analysis argued some harnesses send repeated low-value requests with long contexts—often over **100K tokens**—making costs **tens of times** higher than a subscription price [5]. Another post framed Anthropic's position as allowing products that complement **Claude Code** but not direct competitors [6], a characterization one critic rejected [7].

Impact: The dispute is no longer just about model quality. It is about who gets subsidized compute, who has to pay API rates, and how efficiently agent frameworks use context and caching [5].

Gemma 4 is turning into a distribution story

Gemma 4 is now integrated into **Android Studio** Agent mode for local feature development, refactoring, and bug fixing [8, 9]. Separate posts highlighted **1,500 free daily requests** to **Gemma 4 31B** in Google AI Studio [10], and one user described Gemma running locally on a Pixel phone with no connectivity [11]. Gemma 4 was also cited as the **#1 trending model on Hugging Face** [12, 13].

Impact: Gemma 4 is showing up across IDEs, hosted inference, and offline edge use, which is a stronger adoption signal than benchmarks alone [8, 11, 10].

Goldman Sachs sees a net labor-market drag from AI substitution

Goldman Sachs estimated that, over the past year, AI substitution reduced monthly payroll growth by roughly **25,000** and raised unemployment by **0.16 percentage points**, while augmentation added about **9,000** jobs and lowered unemployment by **0.06 points** [14]. Netting the two implies a **16,000** monthly drag on payroll growth and a **0.1 point** boost to unemployment, with the negative effects concentrated among less experienced workers [14].

Impact: The note argues that today's net labor effect is already negative and is falling disproportionately on entry-level workers [14].

Research & Innovation

Why it matters: The strongest technical work this cycle focused on cheaper long-context inference, better credit assignment for reasoning, and more formal ways to generate theory with LLMs.

Long-context methods keep chipping away at attention cost

- **HISA** replaces a flat sparse-attention token scan with a two-stage **block-then-token** pipeline, eliminating the indexing bottleneck at **64K context** without extra training [15, 16].
- **Screening Is Enough / Multiscreen** replaces softmax-style global competition with threshold-based screening, matching Transformer-like validation loss with **40% fewer parameters** and reducing inference latency by up to **3.2x** at **100K** context [17, 18].
- Commentary around this work framed sparse attention as a form of **maximum inner product search**, while noting that approaches with better theoretical complexity still have to work on GPUs and at datacenter scale to matter in practice [19, 20].

New training methods target deeper reasoning and smaller working contexts

- **FIPO** uses discounted future-KL signals in policy updates, pushing average chain-of-thought length past **10,000 tokens** and reaching **56.0% AIME 2024 Pass@1** on **Qwen2.5-32B** [21].
- **SKILL0** tries to internalize agentic skills into model weights instead of retrieving them at runtime, reporting gains of over **9%** on **ALFWorld** and **6%** on **Search-QA** while cutting context usage to under **0.5K tokens per step** [22].
- **Principia** introduces benchmarks and training recipes for deriving mathematical objects, with gains from on-policy judge training and verifiers that

also transfer to standard numerical and multiple-choice math benchmarks [23].

LLMs are starting to participate in theoretical science workflows

steepest-descent-lean formalizes convergence bounds and hyperparameter scaling laws in Lean using Codex [24, 25]. The work reproduces prior-style results under weaker assumptions, including support for **Nesterov momentum** and **decoupled weight decay** [26, 27, 28], and recovers a fixed-token-budget scaling law of **BS** $\mathbf{T^{2/3}}$ [29]. Its stated workflow is simple: formalize a peer-reviewed proof, ask an LLM to weaken assumptions and re-derive theorems, then keep only the changes that preserve or better match empirical results [30, 25]. The repo is here: steepest-descent-lean [24].

Products & Launches

Why it matters: Useful product progress this cycle was less about one big model launch and more about better infrastructure around coding agents, memory, routing, and interface layers.

GitNexus adds a code graph for agent workflows

GitNexus indexes a codebase into a graph using Tree-sitter, mapping calls, imports, inheritance, execution flows, and blast radius before code changes [31]. The pitch is that agents get the repo's dependency structure precomputed at index time, so smaller models can answer architecture questions without repeated exploration [31]. Setup is a single command: `npx gitnexus analyze` [31]. The project was cited as already reaching **9.4K GitHub stars** and **1.2K forks** [31].

New building blocks are landing for agent memory and control planes

- **Memvid** offers a single-file memory layer for agents with instant retrieval and portable, versioned long-term memory without a database [32].
- **Plano** is an open-source AI-native proxy and data plane for agentic apps, with built-in orchestration, safety, observability, and smart LLM routing [33].

Hermes Agent expands its interfaces

Hermes Agent added support for **OAuth-authenticated MCP servers** [34], can expose an **OpenAI-compatible endpoint** for use with **OpenWebUI** as a chat interface [35], and now ships a **Manim** skill for generating programmatic math and technical animations via `/manim-video <prompt>` [36, 37]. One demo combined the Manim skill with **Math Code** to produce an explanatory video for **Jordan's Lemma** [38].

DESIGN.md turns visual style into plain text for coding agents

The **awesome-design-md** repo packages design-system descriptions from **31 real websites** into markdown files that agents can read directly, covering colors, typography, spacing, buttons, shadows, and responsive rules [39]. The project was presented as a way to avoid repetitive default AI UI aesthetics, and it has now been integrated with Hermes Agent [39, 40].

Industry Moves

Why it matters: The business story this cycle was about sustainable economics, talent concentration, hardware experimentation, and the changing labor and data supply chains behind AI systems.

New pricing models are emerging for agent-heavy workloads

Alongside Anthropic's tighter subscription rules, **MiMo** launched a **Token Plan** that supports third-party harnesses through token quotas and frames the model as long-term, stable delivery rather than open-ended subscription usage [5]. The surrounding commentary argued the market is moving toward a combination of more token-efficient agent harnesses and more efficient models, not simply cheaper tokens [5, 41].

PrimeIntellect added open-source training talent

Open-source researcher **Elie Bakouch** said he is joining **PrimeIntellect** to work on **pre/mid training**, citing the team's open-frontier mission and the leverage of a small focused group [42]. Peers called the hire an **"unbelievable get"** and a **"phenomenal choice"** [43, 44].

Neuromorphic computing patent activity is accelerating

A PatSnap-cited note said neuromorphic computing has moved from academic prototype to commercial product, with **596 patents** filed through early 2026 and a **401%** surge in activity during 2025 [45, 46].

A once-important data-labeling channel in China is weakening

Data-labelling workshops in rural **Guizhou** that were once part of China's poverty alleviation effort and helped build AI systems are now struggling as state support and industry demand have fallen [47, 48]. One commentator suggested similar programs could still be repurposed toward graduate unemployment, but that was presented as an open question rather than a current policy [48].

Policy & Regulation

Why it matters: The policy signal this cycle was less about sweeping AI laws and more about concrete requirements on how companies govern AI internally

and how jurisdictions handle AI infrastructure growth.

Beijing now requires AI ethics committees

New Beijing rules require all Chinese companies engaging in AI activities to establish internal **AI ethics committees**, effective immediately [49]. The final version removed earlier wording that made such committees conditional on circumstances [49], and the move follows a 2023 ethics review system that had been criticized as too narrow and too perfunctory for AI-specific issues [49]. One commentator said the plain reading could be especially hard on smaller startups and questioned how it will be enforced [50].

Maine is moving to pause large data-center projects

Maine is on track to become the first U.S. state to pause construction of large data centers—projects over **20 megawatts**—until **November 2027** while it studies environmental and energy impacts [51]. Commentary around the move acknowledged concerns about rising electricity costs while arguing that infrastructure limits should not become a blanket brake on AI development [51]. The linked report is from the *Wall Street Journal*: Maine data center ban [52].

Quick Takes

Why it matters: These smaller items are useful signals for where local deployment, inference optimization, tooling behavior, and public understanding are moving next.*

- A local **grounded reasoning** demo paired **Gemma 4** with **Falcon Perception**, using Gemma to decide what to inspect and Falcon to return pixel-accurate coordinates; one example checked whether a soccer player was offside, fully local on **M3** hardware [53].
- **TorchSpec** said its **kimi-k2.5-eagle3** draft model hit **40K downloads** on Hugging Face in two weeks, and **vLLM** said it adopted the open-source **EAGLE3** draft model for low-latency inference on **Kimi 2.5** [54, 55].
- A weekend project showed a fully local coding agent using **Qwen3.5 30B A3B**, **llama.cpp/lemonade**, **ngrok**, and **OpenHands**; the builder said performance was better than expected [56].
- **Claude Code** now reportedly throws an error when asked to analyze its own source code [57]. Separately, one user said the tool now refuses some system-fixing workflows they previously relied on, while **Codex** still accepts them [58, 59, 60].
- François Chollet highlighted a tutorial for **fine-tuning Gemma on TPU v5** using **Kinetic + Keras + JAX**, with a quick-start repo here: kinetic-finetuning-on-cloud-tpu [61, 62].
- Ryan Greenblatt argued that statements like **"AGI is here"** or **"we're far from AGI"** are not meaningful unless the speaker defines the term being used [63].

- The documentary **The AI Doc** is now showing in hundreds of theaters, and one commentator said non-technical viewers valued its plain explanation of how LLMs work [64, 65].

Sources

1. X post by @heynavtoor
2. X post by @teortaxesTex
3. X post by @kimmonismus
4. X post by @bcherny
5. X post by @_LuoFuli
6. X post by @paradite_
7. X post by @Teknium
8. X post by @osanseviero
9. X post by @osanseviero
10. X post by @WinterArc2125
11. X post by @matvelloso
12. X post by @_philschmid
13. X post by @ClementDelangue
14. X post by @BrianSozzi
15. X post by @HuggingPapers
16. X post by @TheAITimeline
17. X post by @TheAITimeline
18. X post by @mithernet
19. X post by @part_harry_
20. X post by @teortaxesTex
21. X post by @TheAITimeline
22. X post by @TheAITimeline
23. X post by @TheAITimeline
24. X post by @leloykun
25. X post by @leloykun
26. X post by @leloykun
27. X post by @leloykun
28. X post by @leloykun
29. X post by @CevherLIONS
30. X post by @leloykun
31. X post by @heygurisingh
32. X post by @NerdyRodent
33. X post by @dl_weekly
34. X post by @Teknium
35. X post by @Teknium
36. X post by @NousResearch
37. X post by @Teknium
38. X post by @prompterminal

39. X post by @ihtesham2005
40. X post by @Teknium
41. X post by @teortaxesTex
42. X post by @eliebakouch
43. X post by @code_star
44. X post by @TheZachMueller
45. X post by @kimmonismus
46. X post by @kimmonismus
47. X post by @vince_chow1
48. X post by @teortaxesTex
49. X post by @vince_chow1
50. X post by @teortaxesTex
51. X post by @kimmonismus
52. X post by @kimmonismus
53. X post by @dahou_yasser
54. X post by @lightseekorg
55. X post by @vllm_project
56. X post by @gneubig
57. X post by @theo
58. X post by @theo
59. X post by @theo
60. X post by @theo
61. X post by @fchollet
62. X post by @jigyasa_grover
63. X post by @RyanPGreenblatt
64. X post by @slashdot
65. X post by @dbreunig