Claude Code Expands, Qwen3.5-Omni Ships, and Harness Engineering Takes Center Stage

AI High Signal Digest

2026-03-31

Claude Code Expands, Qwen3.5-Omni Ships, and Harness Engineering Takes Center Stage

By AI High Signal Digest • March 31, 2026

The biggest developments were a more capable Claude Code, Alibaba's Qwen3.5-Omni release, and a growing body of evidence that harness design is becoming a core performance lever. This brief also covers measurable enterprise ROI, faster local AI stacks, new research papers, funding and strategy moves, and governance-related updates.

Top Stories

Why it matters: This cycle's biggest signals were about agent execution: models are getting better at acting on software, multimodal systems are widening the interface, and performance is increasingly coming from the harness around the model as much as the model itself.

Claude Code moved closer to a full software-testing loop

Anthropic added **Computer use** to Claude Code, letting Claude open apps, click through interfaces, and test what it built directly from the CLI; the feature is in research preview on Pro and Max plans [1]. At the same time, Claude Code and Code Review added **GitHub Enterprise Server** support for async workflows on self-hosted repos [2, 3]. Anthropic staff also said they open sourced a plugin so Claude Code users can call **Codex** from a ChatGPT subscription for reviews, adversarial reviews, and rescue flows [4, 5].

Impact: this is a step from code generation toward a tighter **write-build-run-verify** loop, and it makes Claude Code easier to use inside enterprise GitHub setups [6].

Qwen3.5-Omni pushed multimodal interaction further into the product layer

Alibaba released **Qwen3.5-Omni**, a model for text, image, audio, and video understanding with real-time interaction features including semantic interruption, built-in web search, and complex function calling [7]. Alibaba highlighted **script-level captioning**, support for up to **10 hours of audio** or **400 seconds of 720p video**, **113** speech-recognition languages, and **36** output languages, plus an "Audio-Visual Vibe Coding" workflow that turns camera-described ideas into a website or game [7]. The company also said the model is open access via Hugging Face, with the caveat that "omni" here refers to **interpreting** image and voice, not generating them [8].

Impact: Alibaba is packaging multimodal reasoning, voice interaction, and tool use into a surface that looks closer to a general-purpose AI application platform.

Harness engineering is turning into a primary performance lever

Several results this cycle pointed in the same direction: the system around the model matters more than many teams assumed. **Meta-Harness** said prompt/tool/retry/context choices alone can create a **6x** performance gap on the same model, and that harness deltas are now wider than frontier-model deltas [9]. In Matt Maher's **100-feature PRD** benchmark, a post said **Cursor** improved model performance by **11%** on average, including **Opus** from **77%** to **93%** [10, 11]. CMU's **CAID** paper reported **+26.7 points** on PaperBench and **+14.3 points** on Commit0 over single-agent baselines by coordinating isolated git worktrees and explicit integration via git [12].

> "The delta between harness implementations on the same model is not. That's where the leverage is." [9]

Impact: performance gains are increasingly coming from **coordination, evaluation loops, and tool design**, not only from bigger base models.

Enterprise deployments are producing measurable ROI

Two deployment examples stood out for hard numbers. **Novo Nordisk** is using AI agents built on Anthropic and OpenAI models to detect trial risks, automate site selection, and flag process redundancies, shaving **weeks to months** off clinical trials and potentially accelerating time-to-market by **hundreds of millions of dollars** [13]. Separately, a Shopify case study said the company cut annual AI deployment costs from **$5.5M to $73K** by decomposing business logic, modeling intent with **DSPy**, and optimizing a smaller model while maintaining performance; the cited scale-up estimate cut **150,000-shop** coverage from **$41M** to **$73K** [14, 15].

> "The juice is clearly worth the squeeze." [13]

Impact: the strongest enterprise signal in the notes was not hype but **faster trials, lower operating cost, and maintained performance**.

Local AI stacks got faster and more usable

Ollama said it now runs fastest on Apple silicon through **MLX**, Apple's machine-learning framework [16, 17]. Its preview release also added **NVFP4** support, cache reuse across conversations, intelligent checkpoints, and smarter eviction, with a Mac-oriented acceleration path for **Qwen3.5-35B-A3B** on systems with more than **32GB** of unified memory [18, 19, 20]. In parallel, **llama.cpp** reached **100k GitHub stars**, and its creator said local agentic workflows are now practical because tool calling and local models have improved enough to support tasks like search, email, summarization, and home automation [21].

Impact: the local AI stack is getting closer to real everyday agent use on consumer hardware, especially on Macs.

Research & Innovation

Why it matters: Research this cycle focused less on raw scale and more on leverage: better long-context handling, stronger multimodal designs, cheaper training, and harder benchmarks.

- **Massive-context agents without giant context windows:** one paper places very large text corpora into directory structures and lets off-the-shelf coding agents navigate them with shell commands and Python instead of stuffing everything into the context window. The reported results were **88.5%** on **BrowseComp-Plus** versus **80%** best published, **33.7%** on **Oolong-Real** versus **24.1%**, and operation up to **3 trillion tokens** [22]. Paper: https://arxiv.org/abs/2603.20432 [22].
- **LongCat-Next:** a new multimodal model was presented as "lexicalizing modalities as discrete tokens," with claims that it matches or beats SOTA across multimodal benchmarks, delivers SOTA audio on both recognition and TTS accuracy, and adds vision/audio without hurting core language performance [23]. Resources: paper [24], GitHub [24], Hugging Face [24].
- **daVinci-LLM:** this pretraining paper was summarized as matching larger-model performance with **half the size**, adding **23 points** on **MATH**, and arguing that **data quality** can matter more than dataset scale [25]. Resources: paper [26], repo [26].
- **Reasoning and optimization: ParaGator** trains candidate generation and aggregation end-to-end for parallel reasoning, using **pass@k** for generation and **pass@1** for aggregation, with the stated goal of avoiding mode collapse and improving math/scientific reasoning [27]. On the systems side, **Gram Newton-Schulz** was introduced as a drop-in replacement

for Newton-Schulz in **Muon**, with **up to 2x faster** performance while preserving validation perplexity within **0.01** [28].

- **Benchmarks remain hard: PRBench** introduced **30** expert-curated paper-reproduction tasks across **11** physics subfields, and the cited result was stark: **all agents showed zero end-to-end callback success** [29]. **Tau Bench** added a banking domain with **698 documents** across **21** product categories; best models were cited at **25%** task success and under **10%** on pass@4 [30].

Products & Launches

Why it matters: Product work moved toward usable systems: better voice models, more local tooling, and clearer paths from research models to daily workflows.

- **Voice products improved at both ends of the stack.** OpenAI said **gpt-realtime-1.5** improves instruction following, tool calling, and multilingual accuracy in the Realtime API, while a new OpenAI developer post summarized **Perplexity's** lessons from running voice agents in production around context, audio pipelines, and turn-taking [31, 32]. Separately, **Cohere Transcribe** launched as a **2B-parameter** open-weights speech-to-text model with **4.7% AA-WER**, roughly **60x real-time** transcription, training from scratch on **14 languages**, and availability both through Cohere's API and on Hugging Face under Apache 2.0 [33, 34].
- **Local agent tooling kept expanding. ARC (Agent Remote Control)** introduced a browser-based remote monitor for local agents, with real-time tool-call visibility, approvals, messaging, native **Hermes Agent** integration, open source distribution, and end-to-end encryption [35]. **AutoClaw** launched as a way to run **OpenClaw** locally with no API key, support for any model or **GLM-5-Turbo**, and fully local data handling [36]. **litesearch** packaged a fully local document-ingestion and retrieval stack for agents like Claude Code, using LiteParse, local embeddings, local Qdrant storage, and CLI-native search [37, 38].
- **Security-conscious agent wrappers are becoming their own category. PokeeClaw** positioned itself as an enterprise-secure alternative to OpenClaw, with a secure sandbox architecture, isolated environments, approval workflows, role-based access control, audit trails, and lower token usage [39, 40].
- **Composable agent skills are spreading. Base44** added **130+** built-in "Superagent Skills" across marketing, operations, data analysis, design, content, coding, and research, with custom skills created from natural-language descriptions and reusable across workflows [41, 42, 43].

Industry Moves

Why it matters: Corporate signals this cycle were about who owns the agent operating layer, who controls deployment, and where new capital is going.

- **SycamoreLabs** launched as a "trusted agent OS for the enterprise" with a **$65M seed** led by **Coatue** and **Lightspeed**, alongside AbstractVC, Dell Technologies Capital, 8VC, Fellows Fund, e14 Fund, and angel investors [44].
- **Figure AI** described its breakup with OpenAI in unusually direct terms. CEO Brett Adcock said Figure got "no value" from the relationship beyond early fundraising, said Figure's internal team outperformed OpenAI's daily, and said the real break came when OpenAI planned to restart robotics, which would have turned Figure's work into competitor training [45]. Figure has since built its own **vision-language-action** model, **Helix**, and the cited post said the company is valued at **$39B** [45].
- **Anthropic's growth is creating infrastructure strain.** A cited report described the company's success as sparking a **server crunch** [46].
- **Hugging Face is explicitly pushing a builder strategy.** Clement Delangue said the goal is to help "millions" build AI themselves rather than remain API users, and pointed to **hf-autoresearch** as an example of agent collaboration around checkpoints, datasets, papers, and Hub workflows [47, 48].
- **Internal agent deployments are becoming business functions.** A post about LangChain said its internal **GTM agent** drove **250% more lead conversions**, using Deep Agents for orchestration, multiple data sources for context, and Slack for approvals [49]. A separate build log said a similar agent was rebuilt on **DeeplineCLI + Deep Agents** in under an hour with roughly **200 lines** of config [49].

Policy & Regulation

Why it matters: The notes were light on formal government action, but governance questions around data consent, auditing, and safety evaluation were prominent.

- **GitHub Copilot training consent:** a widely shared warning said GitHub had opted users into training its models on their code by default, including paying customers, and pointed users to **Settings > Privacy** to disable it [50].
- **Governance proposals are getting more concrete:** Will MacAskill and Fin Moorhouse proposed eight projects aimed at improving the transition to superintelligence, including independent evaluation of AI character traits, benchmarking strategic and philosophical reasoning, auditing mod-

els for sabotage and backdoors, and building monitoring and verification tools for collective coordination [51].

- **Safety debate stayed active:** Boaz Barak published a new post titled *the state of AI safety in four fake graphs*, which Sam Altman publicly endorsed as "a very good post" [52, 53].

Quick Takes

Why it matters: These smaller items help fill in the operating picture around models, agent frameworks, and supporting infrastructure.

- **Qwen 3.6 Plus Preview** went live on **OpenRouter** for a limited free period; Alibaba asked for feedback and noted prompts/completions may be collected during the preview [54, 55].
- **Codex auto compaction** was reported to improve long-session coherence, with one user saying Codex remembers tiny details across multiple rounds of compaction [56, 57].
- **Hermes Agent** added **Multi Agent Profiles**, giving independent bots separate memory, gateway connections, skills, and chat histories [58, 59].
- A new **BOOT.md hook** in Hermes lets agents save state before restarts and resume with what one post described as zero context loss [60].
- **OpenAI's Codex App Server** is fully open source, includes **sign in with ChatGPT**, and powers Codex integrations in products like the Codex app and external tools such as JetBrains and T3 Code [61].
- **PixVerse V6** launched on **fal.ai** with text-to-video, image-to-video, transition, and extend endpoints, while PixVerse separately promoted V6 as offering more control, better performance, and 15-second 1080p audiovisual generation [62, 63, 64].
- **LisanBench** launched a live benchmark site with leaderboard visualizations, and its creator said a meta leaderboard is next [65, 66].
- **Triton-Ascend** is now public, giving Huawei Ascend hardware a Triton kernel programming model that commenters said could help frameworks like **sglang** and **vLLM** run on Ascend without learning AscendC [67, 68].
- **Gemini Live** is now powered by **Gemini 3.1 Flash Live** [69].

Sources

1. X post by @claudeai
2. X post by @_catwu
3. X post by @katchu11
4. X post by @romainhuet
5. X post by @reach_vb
6. X post by @Yuchenj_UW
7. X post by @Alibaba_Qwen
8. X post by @kimmonismus

9. X post by @LiorOnAI
10. X post by @edwinarbus
11. X post by @theo
12. X post by @omarsar0
13. X post by @kimmonismus
14. X post by @kmad
15. X post by @lateinteraction
16. X post by @ollama
17. X post by @awnihannun
18. X post by @ollama
19. X post by @ollama
20. X post by @ollama
21. X post by @ggerganov
22. X post by @dair_ai
23. X post by @arankomatsuzaki
24. X post by @arankomatsuzaki
25. X post by @arankomatsuzaki
26. X post by @arankomatsuzaki
27. X post by @jaseweston
28. X post by @jcz42
29. X post by @arankomatsuzaki
30. X post by @_philschmid
31. X post by @OpenAIDevs
32. X post by @OpenAIDevs
33. X post by @ArtificialAnlys
34. X post by @ArtificialAnlys
35. X post by @winglian
36. X post by @Zai_org
37. X post by @llama_index
38. X post by @jerryjliu0
39. X post by @Pokee_AI
40. X post by @fchollet
41. X post by @kimmonismus
42. X post by @Base44
43. X post by @kimmonismus
44. X post by @SriViswan
45. X post by @mikekalilmfg
46. X post by @steph_palazzolo
47. X post by @ClementDelangue
48. X post by @ClementDelangue
49. X post by @jai__toor
50. X post by @GergelyOrosz
51. X post by @willmacaskill
52. X post by @boazbaraktcs
53. X post by @sama
54. X post by @OpenRouter

55. X post by @Alibaba_Qwen
56. X post by @markchen90
57. X post by @alxfazio
58. X post by @Teknium
59. X post by @NousResearch
60. X post by @louislaurent
61. X post by @dkundel
62. X post by @fal
63. X post by @fal
64. X post by @PixVerse_
65. X post by @scaling01
66. X post by @scaling01
67. X post by @_rozzai
68. X post by @teortaxesTex
69. X post by @GeminiApp