Agents Stretch Into Multi-Day Work as Inference Takes Center Stage

AI News Digest

2026-03-28

Agents Stretch Into Multi-Day Work as Inference Takes Center Stage

By AI News Digest • March 28, 2026

Today's clearest thread is duration: leading researchers say agents are starting to sustain hours- or days-long work, while chips and datacenters are being redesigned around inference-heavy demand. The same cycle is sharpening the labor debate and the fight for default distribution on major devices.

The big shift

Longer-running agents are moving from benchmark wins to extended workflows

Onstage at GTC, Jeff Dean said the clearest recent gains are on verifiable tasks such as math and coding, citing Gemini gold medals in the IMO and ICPC, and added that agent workflows now let models pursue tasks that run for hours or even days with less close supervision [1]. Jack Clark described agents more plainly as language models that use tools over time, and said Anthropic now sees some research projects shrink from two to three weeks to one to two days [2].

A separate controlled experiment shared on /r/MachineLearning pointed in the same direction: a Claude Code agent with access to 2M+ CS papers beat an otherwise identical setup by 3.2% on TinyStories after 100 automated experiments [3]. *Why it matters:* The newest gains are increasingly tied to sustained tool use, retrieval, and iteration, not just better one-turn answers [1, 2, 3].

The infrastructure response

Inference is overtaking training as the main systems problem

> "Inference is the job now." [1]

Bill Dally said more than 90% of datacenter power is already going to inference, with different hardware needs emerging for training, prefill, and decode; he said Nvidia is targeting low-latency architectures that could run relatively large models at 10,000-20,000 tokens per second per user [1]. He also argued that moving data dominates energy cost: reading a value from external memory can be roughly 1,000x costlier than a multiply-add, which is why Nvidia is exploring SRAM-local computation and stacked DRAM [1].

The same session suggested AI is starting to reshape chip design itself. Nvidia said NBCell can port cell libraries overnight with results that match or exceed human designs, Prefix RL produced adders 20-30% better on size, power, and timing metrics, and Google said AlphaChip has already helped with multiple TPU generations [1]. *Why it matters:* The center of gravity is moving from raw training scale toward inference latency, memory movement, and AI-assisted hardware design [1].

Demand is showing up in both revenue and concrete

In a Plain English interview, the host said analysts estimate Anthropic's annual recurring revenue more than doubled from $9B in December 2025 to more than $20B in March 2026, with no known precedent at that scale [2]. Physical capacity is moving in parallel: Microsoft said it is partnering with Crusoe on a 900MW AI factory campus in Abilene, Texas, and Sam Altman said the first steel beams went up this week at the Michigan Stargate site with Oracle and Related Digital [4, 5].

Why it matters: Strong revenue growth is still being translated into very large compute buildouts, which makes AI demand look durable rather than purely speculative [2, 4, 5].

The policy and distribution questions

Labor warnings are getting sharper, but the timeline is contested

Senator Mark Warner said recent college graduate unemployment could rise from about 9% to 30%, pointed to law firms pausing first-year associate hiring, slashing back-office headcount, and cutting internships, and backed bipartisan bills to report AI-driven job losses and generate policy responses [6]. He said government and society are not ready for the next three to five years of disruption [6].

Jack Clark, by contrast, said he does not agree with Dario Amodei's forecast of 20% unemployment and the loss of half of entry-level white-collar roles within about five years; he argued big employment shifts usually take time, policy choices matter, and today's agents multiply productivity more than they fully replace people [2]. *Why it matters:* The debate has shifted from whether AI will affect white-collar work to how fast the shock arrives and what policy response should be ready first [6, 2].

Default distribution is becoming a strategic front

Perplexity said it is deepening its Samsung partnership to power AI in the browser preinstalled on more than 1B Samsung devices with 100M+ active users, extending existing work on Bixby and preload on Galaxy S26 devices alongside Gemini [7]. Separately, Big Technology highlighted a report that Apple will open Siri to AI assistants from rival companies in iOS 27 [6].

Why it matters: The contest is no longer just about model quality; placement inside browsers, assistants, and operating systems could decide who actually reaches users at scale [7, 6].

Sources

1. Jeff Dean and Bill Dally, Advancing to AI's next Frontier, Nvidia GTC 2026
2. Anthropic Thinks AI Might Destroy the Economy. It's Building It Anyway. | Plain English
3. r/MachineLearning post by u/kalpitdixit
4. X post by @mustafasuleyman
5. X post by @sama
6. Senator Mark Warner on AI's Risks: "I Want To Be More Optimistic, But I Am Terrified."
7. X post by @AravSrinivas