ggml joins Hugging Face as NVIDIA open-sources DreamDojo, India demand spikes, and agent stacks mature

AI News Digest

2026-02-21

ggml joins Hugging Face as NVIDIA open-sources DreamDojo, India demand spikes, and agent stacks mature

By AI News Digest • February 21, 2026

Key moves today span open-source local AI (ggml/llama.cpp joining Hugging Face), a major open robotics world-model release (NVIDIA's DreamDojo), and fresh signals from India on both usage growth and "sovereign" model building. Also: new agent-layer tooling, updated task-horizon evals for Claude Opus 4.6, and escalating policy/IP tensions around military use, safety reporting, and generative video.

ggml (llama.cpp) joins Hugging Face, doubling down on local AI

ggml.ai + Hugging Face: make local inference easier to use

The ggml.ai team (behind **llama.cpp**) is joining **Hugging Face**, with a stated plan to keep building ggml, make llama.cpp "more accessible," and "empower the open-source community" around **local AI on personal hardware** [1][2].

> "Our joint mission is to make local AI easy and efficient to use by everyone on their own hardware." [3]

More detail was shared via Hugging Face's post: https://huggingface.co/blog/ggml-joins-hf [4].

[1] post by @ggerganov

[2] post by @Thom_Wolf

[3] post by @ggerganov

[4] post by @Thom_Wolf

Robotics: NVIDIA open-sources an interactive world model (DreamDojo)

DreamDojo: "Simulation 2.0" from videos → robot-readable actions

NVIDIA's Jim Fan announced **DreamDojo**, an **open-source, interactive robot world model** that takes robot motor controls and generates future frames "in pixels," explicitly positioning it as a shift away from traditional simulators with engines/meshes/hand-authored dynamics [5].

The system is pretrained on **44K hours of human egocentric video** and uses **latent actions** inferred from video to make the data "robot-readable" across different hardware, with a claim of **zero-shot generalization** to unseen objects/environments [6][7]. It also includes **post-training** to adapt to a specific robot's actuation mechanics via gradient descent [8].

Real-time interaction + reported task gains

DreamDojo includes a real-time distilled version reported to run at **10 FPS**, enabling live teleoperation, policy evaluation, and model-based planning "inside a dream" [9][10]. In one example, model-based planning is described as improving real-world success by **+17%** out of the box on a fruit-packing task [11].

Links: project https://dreamdojo-world.github.io/ [12] • paper https://arxiv.org/abs/2602.06949 [13] • code https://github.com/NVIDIA/DreamDojo [14].

India signals: demand + "sovereign" model building

OpenAI: Codex usage in India spikes

Sam Altman said he met with India's Prime Minister Narendra Modi "to talk about the incredible energy around AI in India" [15]. He also said India is OpenAI's **fastest-growing market for Codex**, with weekly users **up 4× in the past two weeks** [16].

Altman separately warned against slow enterprise AI timelines, describing a company planning to spend 2026–2028 "strategizing" and saying that approach "will be a catastrophic mistake" for AI adoption [17].

[5] post by @DrJimFan
[6] post by @DrJimFan
[7] post by @DrJimFan
[8] post by @DrJimFan
[9] post by @DrJimFan
[10] post by @DrJimFan
[11] post by @DrJimFan
[12] post by @DrJimFan
[13] post by @DrJimFan
[14] post by @DrJimFan
[15] post by @sama
[16] post by @sama
[17] Rapid Fire With Sam Altman: His Take on AGI, Musk & the Future of AI | Express Adda

Rapid Fire With Sam Altman: His Take on AGI, Musk & the Future of AI | Express Adda (22:09)

Sarvam AI: training 30B + 105B "from scratch in India" with a small team

A Sarvam AI post described training **30B** and **105B** models "from scratch in India" with a team of **15**, with benchmarks and Hugging Face links said to be forthcoming [18]. Vinod Khosla also highlighted SarvamAI as "outdistancing" others in an "India sovereign AI model race" based on presence at an India AI summit event [19].

Yann LeCun: coalition-built open frontier models as a sovereignty path

In a Delhi interview, Yann LeCun argued that countries could collaborate to build **open-source frontier models** using regional data "not accessible" to proprietary models, and that this could allow open models to "become better" than closed proprietary ones [20]. He also argued for more investment in research and PhD training in India to build local expertise [21].

[18] post by @pratykumar

[19] post by @vkhosla

[20] India should invest more in research and PhDs: AI Pioneer Yann LeCun in Delhi

[21] India should invest more in research and PhDs: AI Pioneer Yann LeCun in Delhi

Agents & developer tooling: orchestration layers proliferate

“Claws” emerge as a new layer on top of LLM agents (with security caveats)

Andrej Karpathy described “**Claws**” as a new layer on top of LLM agents—pushing orchestration, scheduling, context, tool calls, and persistence forward [22]. He expressed concern about running **OpenClaw** given its scale (“400K lines”), reports of exposed instances, RCE vulnerabilities, supply chain poisoning, and malicious/compromised “skills,” calling it a “wild west” and “security nightmare” even while endorsing the concept [23].

He highlighted **NanoClaw** as a smaller core (~4000 lines) with containerized defaults and a “skills”-driven configurability approach (e.g., `/add-telegram` instructing an agent to modify code to integrate Telegram) [24]. Separately, swyx said NanoClaw addresses OpenClaw complaints as a minimal hackable reproduction (~700 LOC) that uses **Apple Containers** for sandboxing/security [25].

Google open-sources an Agent Development Kit (ADK)

A LocalLLM post says Google has officially launched the **Agent Development Kit (ADK)** as **open source** [26][27].

antaris-suite 3.0: in-process agent memory/guard/router/context as Python packages

Antaris Analytics announced **antaris-suite 3.0**, described as six **zero-dependency** Python packages covering an agent turn’s infrastructure: memory, safety/guard, routing, context management, pipeline coordination, and shared contracts [28]. It also describes an OpenClaw integration where memory recall and ingest hook into every agent turn automatically, including “compaction-aware session recovery” for long-running agents [29].

The release also notes it ran a “three-model gauntlet” (Claude, ChatGPT, Gemini) that found issues before shipping, and that those were resolved with **1,465 tests passing** [30]. GitHub: https://github.com/Antaris-Analytics/antaris-suite [31].

[22] post by @karpathy
[23] post by @karpathy
[24] post by @karpathy
[25] post by @swyx
[26] r/LocalLLM post by u/Fun-Necessary1572
[27] r/LocalLLM post by u/Fun-Necessary1572
[28] r/LocalLLM post by u/fourbeersthepirates
[29] r/LocalLLM post by u/fourbeersthepirates
[30] r/LocalLLM post by u/fourbeersthepirates
[31] r/LocalLLM post by u/fourbeersthepirates

Model capability measurement: longer task horizons (with big error bars)

METR: Claude Opus 4.6 time-horizon estimate on software tasks

METR Evals reported an estimate that **Claude Opus 4.6** has a "50%-time-horizon" of ~**14.5 hours** on software tasks (95% CI: **6 to 98 hours**), while noting the measurement is "extremely noisy" because the task suite is "nearly saturated" [32].

A separate Reddit discussion notes a METR update where Claude Opus 4.6 "hits 50%" on **multi-hour expert ML tasks** (example: "fix complex bug in ML research codebase"), with confidence bands described as wide but the trend "clear" [33][34].

Policy, safety, and IP friction

Pentagon vs. Anthropic: "legal use cases" vs. red lines

A segment summarized by Matt Wolfe describes tension where the Pentagon wants to use Anthropic models for "all legal use cases," while Anthropic does not want its models used for **mass surveillance** or **fully autonomous weapons without a human in the loop** [35].

In a separate interview, Anthropic CEO Dario Amodei said Anthropic has deployed models for U.S. national security for "quite a while," but reiterated concern about **fully autonomous weapons** (no human-in-the-loop) and **domestic mass surveillance of Americans**, framing these as important red lines for compatibility with democracy and the company's culture [36].

Report: OpenAI flagged violence-related writings; leaders chose not to alert authorities

A post citing a WSJ Tech link says OpenAI internal systems flagged a "Canadian trans shooter" for writings about real-world violence including gun violence [37]. It also says over a dozen OpenAI employees debated telling law enforcement, but OpenAI leaders decided not to inform authorities about a "potential mass murder" [38]. Elon Musk responded: "Troubling" [39]. WSJ Tech link (as shared): https://x.com/wsjtech/status/2024960405915787751 [40].

[32] post by @METR_Evals

[33] r/MachineLearning post by u/thefuturespace

[34] r/MachineLearning post by u/thefuturespace

[35] AI News: 5 New Models Dropped This Week!

[36] Not Trying To Replace Existing IT Industry: Anthropic CEO Dario Amodei | EXCLUSIVE

[37] post by @KatieMiller

[38] post by @KatieMiller

[39] post by @elonmusk

[40] post by @KatieMiller

Hollywood backlash over ByteDance's Seed Dance 2.0; ByteDance says it will add safeguards

Matt Wolfe described statements from SAG-AFTRA, Disney, and the Motion Picture Association condemning ByteDance's AI video model **Seed Dance 2.0** over alleged unauthorized use of voices/likeness and IP, and reported ByteDance saying it will add safeguards and strengthen protections against unauthorized IP/likeness use [41].

Inside companies: Amazon's writing culture vs. AI-written six-pagers

A Big Technology report describes tension inside Amazon as leadership pushes internal AI tools (including "Cedric," described as a ChatGPT-style tool) that promise "six-page narratives in seconds," while employees characterize the tools as "comically inadequate," citing hallucinations and a lack of clear training/measurement [42][43][44][45].

Veterans in the report worry Amazon is losing sight of the idea that "writing is thinking," describing a loop of "chatbots writing six-pagers to be summarized by other chatbots," while newer employees describe pressure to increase output volume in part because AI can summarize longer docs [46][47].

Research papers to skim (synthetic evals + end-to-end summarization)

- **LOLAMEME**: A synthetic evaluation framework comparing GPT-2 (Transformer), Hyena (convolution), and a hybrid architecture (THEX) on logic+memory tasks with features like global variables and mixed-language syntax; the authors report THEX outperforming Hyena and GPT-2 on several tasks and argue attention/convolution are complementary [48][49][50][51]. Paper: https://arxiv.org/abs/2406.02592 [52].
- **JADS**: An end-to-end model unifying multi-document topic discovery and summarization, described as outperforming a two-step pipeline (BERTopic + Longformer) by **8–9 ROUGE points** using self-supervised data cre-

[41] AI News: 5 New Models Dropped This Week!
[42] Writing Crystalized Thinking At Amazon. Is AI Muddying It?
[43] Writing Crystalized Thinking At Amazon. Is AI Muddying It?
[44] Writing Crystalized Thinking At Amazon. Is AI Muddying It?
[45] Writing Crystalized Thinking At Amazon. Is AI Muddying It?
[46] Writing Crystalized Thinking At Amazon. Is AI Muddying It?
[47] Writing Crystalized Thinking At Amazon. Is AI Muddying It?
[48] r/MachineLearning post by u/djaym7
[49] r/MachineLearning post by u/djaym7
[50] r/MachineLearning post by u/djaym7
[51] r/MachineLearning post by u/djaym7
[52] r/MachineLearning post by u/djaym7

ation, with a Longformer encoder-decoder processing up to **16K tokens** [53][54][55][56]. Paper: https://arxiv.org/abs/2405.18642 [57].

Sources

1. post by @ggerganov
2. post by @Thom_Wolf
3. post by @DrJimFan
4. post by @DrJimFan
5. post by @sama
6. Rapid Fire With Sam Altman: His Take on AGI, Musk & the Future of AI | Express Adda
7. post by @pratykumar
8. post by @vkhosla
9. India should invest more in research and PhDs: AI Pioneer Yann LeCun in Delhi
10. post by @karpathy
11. post by @swyx
12. r/LocalLLM post by u/Fun-Necessary1572
13. r/LocalLLM post by u/fourbeersthepirates
14. post by @METR_Evals
15. r/MachineLearning post by u/thefuturespace
16. AI News: 5 New Models Dropped This Week!
17. Not Trying To Replace Existing IT Industry: Anthropic CEO Dario Amodei | EXCLUSIVE
18. post by @KatieMiller
19. post by @elonmusk
20. Writing Crystalized Thinking At Amazon. Is AI Muddying It?
21. r/MachineLearning post by u/djaym7
22. r/MachineLearning post by u/djaym7

[53] r/MachineLearning post by u/djaym7

[54] r/MachineLearning post by u/djaym7

[55] r/MachineLearning post by u/djaym7

[56] r/MachineLearning post by u/djaym7

[57] r/MachineLearning post by u/djaym7