Karpathy's Capability Funnel Lands as Mathematicians Reckon with OpenAI's 700+ Manuscripts

AI High Signal Digest

2026-10-12

Karpathy's Capability Funnel Lands as Mathematicians Reckon with OpenAI's 700+ Manuscripts

By AI High Signal Digest • October 12, 2026

Karpathy describes a widening gap between casual LLM users and the few people with frontier access, while mathematicians respond to OpenAI's AI-generated manuscripts. Also covered: OpenAI extends Sol's context window, more open-weight inference undercuts frontier pricing, and new multi-agent research.

Karpathy: the capability gap is now a funnel

Andrej Karpathy revisited his April post and argued that the gap in understanding of LLM capability is widening into "a sharp funnel." His rough numbers ("napkin math") break down as follows: - About 6B people have barely touched LLMs. - 1–2B are casual free-tier users. - About 20M professionals in math and code see agents finish multi-week projects from a prompt. - About 5,000 people have internal access to frontier systems [1].

Karpathy describes what that last group sees as swarms of thousands of agents working for weeks on software, cyber, science and math. He says "human review and comprehension are starting to fall behind," and points to the 722 frontiermath manuscripts as an example [1]. He gives three causes: how ambitious the task is and how long it runs; capability peaking in domains that are digital, verifiable and economically valuable, because of RL on verifiable rewards; and access tier [1]. This framework helps explain why reactions to AI progress this week varied so widely.

Mathematics absorbs the OpenAI release

Le Monde, via a widely shared post, reports that Fields medallist Hugo Duminil-Copin was "deeply shaken" by OpenAI's release of more than 700 AI-generated manuscripts, which include work on 16 questions directly relevant to his research. He is quoted saying, "They have annihilated my field of research" [2]. Fellow Fields medallist Jacob Tsimerman thinks mathematics will soon rest on proofs no one understands. He says that is already true of many of his own theorems, which build on results whose proofs he doesn't know [3].

Hugging Face turned OpenAI's problems into open-source RL environments. Clément Delangue calls them "early and rough": the verifier accepts only the exact original formalization, so an equivalent proof can score zero [4]. A slide from Terence Tao's "Math 2.0" lecture at Caltech set off a separate argument about whether a cure needs to be understood if it works [5]. François Fleuret, for example, noted that many drugs have unknown mechanisms of action [6].

Agent-assisted open research is also moving fast. An integer-multiplication effort crossed $= 10^{3}$, with much of the day's progress coming from optimizing and canceling existing parts of the algorithm [7]. Of 151 PRs opened in 24 hours, 89 disclosed using Codex and 48 Claude. Codex produced the most advances, but the top teams used both models [7].

Frontier model updates

- **GPT-6.1 Sol context:** the default context in Codex and Work went from 272k to 500k tokens, about an 80% increase. The goal is fewer compactions. Users can set it as high as 1M, and OpenAI says it will go higher [8].
- **Long-horizon behavior:** one researcher reports that GPT-6 Astra followed requested work durations precisely, but only for tasks over 2 hours. Asked to work 72 hours, Astra stopped at exactly 72h while Fable 5.1 gave up after 3h. The researcher asks whether stopping to check in might actually be better [9].
- **Anthropic rumors (unconfirmed):** users say requests appear to be going to a new Opus 5.5 or 5.6 checkpoint that "looks really, really good" [10]. Others say Fable 5 is routing to a possible Fable 5.5. One commentator speculates it will be released ahead of Anthropic's IPO [11]. Anthropic has not confirmed any of this.

Open weights: cheaper inference, specialist fine-tunes, and a margin debate

Two new inference providers launched on DeepSeek V4.1 Flash: - **River AI** prices DeepSeek v4.1 Flash and GLM-5.3 Flash 20% below DeepSeek's and ZAI's own offerings. It credits a training and inference stack it wrote from scratch in Rust [12]. - **Dolphin** claims "Opus 5-class" coding at 97% lower cost

than frontier models [13].

Brian Armstrong says Coinbase fine-tuned Qwen3.5-9B on internal data for fraud prevention, and that it beat frontier models at lower cost. Coinbase has added a fine-tuning skill to its internal harness, and Armstrong expects some backend services to be replaced this way [14]. The company's write-up projects fraud-prevention savings of roughly 3× the inference cost [15].

Jaya Gupta argues that inference is a "permanent reverse auction." In her view it is weakly defensible because switching providers is "one line of code" and neolabs keep adding spare capacity [16]. Sarah Guo replies that inference is "a scale game" and that leading providers rarely lose customers [17]. In a SemiAnalysis clip, the speaker argues that even the labs underestimated their own demand and signed GPU contracts at high premiums. They call it "a revenue and profitability fast takeoff" rather than an AI one [18].

Small "decision models" kept arriving after last period's launches from Microsoft and OpenAI: - vllm-sr released Decision 3.0, open weights from 0.6B to 27B, and claims #1 on both Jev Decision Index boards in its own evaluation [19]. - Liquid AI released open-weight d1-3B and d1-omni-600M, which make decisions from images. Liquid reports 35 ms per image on Jetson AGX Thor and is targeting on-device checks for robots [20].

Multi-agent research

Meta Superintelligence Labs' **GitSwarm** gives hundreds of asynchronous agents one shared Git repository, with no central planner. With GPT-5.5 Codex workers it solved all 30 IMOProofBench-Advanced problems in one run. It scored 79.4% on ProgramBench, against 65.1% for a single agent at a similar budget, and the score rose as workers were added. 70.1% of failed research commits were cited by later accepted work [21].

Sakana released new detail on **MASS**, its self-improvement loop where one model acts as solver, optimizer and evaluator. Training only on task-solving data also improved the optimizer and evaluator roles. Multi-agent traces made more efficient training data than single-agent ones, and optimizer capacity can limit how fast it improves [22]. Sakana warns that when a model is its own evaluator, shared blind spots can let mistakes be "accepted as progress" [22].

Other signals

- **Meta Muse retention:** above ChatGPT at day 7 (>30%) and just below it at day 14 (22%) [23]. Scott Stevenson notes the curve keeps falling instead of flattening, and guesses users switch off its proactive notifications [24].
- **XPeng's Iron humanoid:** a post says mass production is beginning, with a European rollout in H2 2027 [25].

- **Baidu's Unlimited OCR:** a 3B open-source model that parses 100-page PDFs in one pass [26]. Jerry Liu's independent check puts it at about 46.2% on ParseBench (#115), noting it skips visual elements and the most complex tables [27].
- **Packaging:** Cu/diamond composites debuted at Semicon Taiwan at about 800 $\mathrm{W/m \cdot K}$, roughly double pure copper's ~400 $\mathrm{W/m \cdot K}$. The main obstacle is bonding copper to diamond [28].

Sources

1. X post by @karpathy
2. X post by @miniapeur
3. X post by @TOEwithCurt
4. X post by @ClementDelangue
5. X post by @AndrewCurran_
6. X post by @francoisfleuret
7. X post by @RohanArun
8. X post by @thsottiaux
9. X post by @maksym_andr
10. X post by @kimmonismus
11. X post by @kimmonismus
12. X post by @ibab
13. X post by @dphnAI
14. X post by @brian_armstrong
15. X post by @karinanguyen
16. X post by @JayaGup10
17. X post by @saranormous
18. X post by @SemiAnalysis_
19. X post by @XunzhuoLiu
20. X post by @TheTuringPost
21. X post by @omarsar0
22. X post by @SakanaAILabs
23. X post by @deedydas
24. X post by @scottastevenson
25. X post by @LinusEkenstam
26. X post by @BrooksWhaleX
27. X post by @jerryjliu0
28. X post by @SemiAnalysis_