Recursive Self-Improvement Takes Center Stage as Fable 5.2 Pulls Ahead and OpenAI's Next Model Stalls

AI High Signal Digest

2026-11-15

Recursive Self-Improvement Takes Center Stage as Fable 5.2 Pulls Ahead and OpenAI's Next Model Stalls

By AI High Signal Digest • November 15, 2026

Dwarkesh Patel's essay argues AI research is the next domain for an "avalanche" like math's, and researchers are split on how close that is. Also: Fable 5.2's jump on SimpleQA, OpenAI's longest release gap in more than a year, new Reuters reporting on Jensen Huang, and claims about inference speed and cost.

The RSI argument goes mainstream

Dwarkesh Patel argues that AI research has verifiable proxies, as math does, so even today's jagged models can match or beat humans at it. In his view, narrow AI researchers could squeeze a GPT-3-to-GPT-6-sized jump into a single year, "a couple years away (and possibly only a couple months away)" [1]. His case for why narrow systems would have broad effects is that better learning algorithms improve everything [1].

Researchers who run these experiments are divided: - **Elie Bakouch** largely agrees but names a source of slowdown: second-order interactions across architecture, data and training stages. One example is that labs keep reversing their attention-design choices. He says compute remains the main bottleneck and the gap is large between closed US labs and everyone else [2]. He also warns that reported pretraining-efficiency gains often mix data improvements with algorithmic ones, and that data tuned to the held-out evaluation set can produce "INSANE efficiency gain" [2]. - **Rohan Anil (@_arohan_)** has yet to see AI find a better optimizer for a frontier model, and says Prime Intellect's results were stuck in a local minimum [3]. Bakouch replies that in his NanoGPT

experiment Opus 5.5 beat the human record, though he puts most of the gains down to hyperparameter tuning that the environment rewards [4]. He guesses that Anthropic's internal models are already at or above human researchers on such tasks [4].

On the more alarmed side, Redwood's Buck Shlegeris puts the chance of superintelligence within a year at about one in three. He cites METR research he says found current AIs raise Anthropic's rate of AI progress by about 70% [5]. He gives roughly 20% odds that misalignment upends ordinary life within a year, and says he wrote the post quickly [5]. Neel Nanda called the conclusion "very concerning" [6].

The Anthropic roadshow chart covered in the last brief is now being debated more calmly. Bakouch says Anthropic shouldn't be blamed for a graph marked illustrative, and that large gains are plausible against a GPT-2 baseline once both data and algorithms are counted [7]. Omar Khattab (@lateinteraction) says he believes the gains are substantial but wants labs to share only "non-vacuous information" [8].

Frontier race: Fable surges, OpenAI is quiet

Andrew Curran notes that Fable 5.2 "jaggedly jumped ahead of the entire pack" on Epoch's SimpleQA Verified [9]. Teortaxes thinks Anthropic built a new RL routine that draws out knowledge already in the pretrained model, and expects it to be "quickly commoditized" [10]. Curran expects two signals next week, though neither is confirmed. One is an Anthropic internal progress report with updated RSI numbers. The other is GPT-6.2 Astra, possibly released alongside another major math result [9].

OpenAI has gone 72 days without a frontier release, its longest gap since GPT-5 to 5.1 [11]. Chubby (@kimmonismus) says several people told him none of the commonly floated explanations account for the GPT-6.2 delay. His guess, which he admits is speculation, is that OpenAI may skip to GPT-6.5 [12].

Contributor toyama nao's tests on Zhihu found Google's Gemini 4 Hydrogen beat Argon on coding at similar cost. That comparison also included an Open-Code upgrade [13]. On reasoning, Hydrogen was weaker, with a median 61.66 against Argon's 69.13 [13], and it scored lower on hallucination tasks, where retries rarely helped [13].

Nvidia after Huang

Reuters reports, via a relayed post, that Jensen Huang told Nvidia's general counsel and CMO in October that his relationship with Alice Huang was standard customer business. He also wanted legal action looked into against the NYT, though no suit was ever filed [14]. Elon Musk warns that many of Huang's supply-chain deals were handshakes, not written contracts, and says some "will

be lost" [15]. SemiAnalysis points out that six of Nvidia's seven biggest customers are building their own chips; SpaceX is the exception [16]. It says its benchmarks show the TPU is "just as good as NVIDIA," and that Huang's skill at getting suppliers to invest is the edge Nvidia has [16].

Inference economics

Zyphra is serving GLM-5.3-Flash in FP4 on AMD MI300X at 25% below its FP8 price [17]. It tested on SemiAnalysis's AgentX, a replay of multi-turn coding-agent sessions, at 96 concurrent sessions. Per-user p90 generation was 44.6 tokens/s, against 4.7 for stock SGLang [18]. Zyphra says each GPU brings in $5.00/hour against rental prices of $1.85–2.99 [19].

A third party re-tested the Baseten DeepSeek V4.1-Flash engine covered last time. Using the same 4×B300 setup and fixed acceptance, tuned vLLM reached 843 TPS against about 1,010 for the proposed deployment. That is only 1.2×, falling to 1.13× at 700K context [20].

Research worth noting

- **Weight forensics, extended:** Alapan Chaudhuri says that given a model's starting weights, you can recover which tokens were seen, the LR schedule and the step count from its released weights [21]. Examples include 20,341 OLMo 2 embedding rows that never moved, and Pythia-1B's true peak LR of 2.5e-4 rather than the 3.0e-4 on its model card [21]. Lucas Beyer notes this works for fine-tuning; for pretraining, the traces likely wash out [22]. Separately, Bojie Li scanned 265 open LLMs, inferred the GPU family for 185, and says he independently confirms Will Depue's Blackwell findings [23].
- **Agent memory:** MemCredit picks memories by measuring how much each one changes the agent's decision, not by similarity. On ALFWorld it reached 86.57% success against 65.67% for BM25, using 74.5% fewer memory tokens [24].
- **Prompt injection:** MetaPI turns AppWorld, VitaBench and Terminal-Bench into injection tests. Even the most robust closed models, gpt-5.6-luna and claude-sonnet-5, were attacked successfully more than 20% of the time [25].
- **EmbeddingGemma 2** was Hugging Face's #1 trending model for over a week [26]. It is distilled from Gemini Embedding 2 and supports Matryoshka sizes from 768 down to 128 [27].
- **e-means** from Meta and ENPC is a drop-in replacement for k-means. The authors report it beats k-means accuracy in a tenth of the epochs, with 10–45× GPU speedups [28].

Policy and practice

New York's RAISE Act is in effect. Frontier developers with more than $500M in annual revenue must register by January 1, 2027 and report critical incidents within 72 hours [29].

Perplexity's Aravind Srinivas reports that agent-hours from Teammate are now 11× September levels, up from 7× a week earlier. He estimates $80M–180M a year in labor value against about $10M in tokens [30].

Steve Yegge argues the other way: pace AI output to human speed. He says his agents swamped his players, his accountant and his merge queue, and launched broken OSS without permission [31]. He has burned down five software factories because rewriting them became cheaper than fixing them [31]. Anthropic also published a guide to always-on agents, covering event inboxes, timers and memory [32].

Sources

1. X post by @dwarkesh_sp
2. X post by @eliebakouch
3. X post by @_arohan_
4. X post by @eliebakouch
5. X post by @bshlgrs
6. X post by @NeelNanda5
7. X post by @eliebakouch
8. X post by @lateinteraction
9. X post by @AndrewCurran_
10. X post by @teortaxesTex
11. X post by @synthwavedd
12. X post by @kimmonismus
13. X post by @ZhihuFrontier
14. X post by @kimmonismus
15. X post by @elonmusk
16. X post by @SemiAnalysis_
17. X post by @ZyphraAI
18. X post by @ZyphraAI
19. X post by @ZyphraAI
20. X post by @zakariaornot
21. X post by @alapanchaudhuri
22. X post by @giffmana
23. X post by @bojie_li
24. X post by @omarsar0
25. X post by @dair_ai
26. X post by @gabriberton
27. X post by @gabriberton

28. X post by @TheTuringPost
29. X post by @dl_weekly
30. X post by @AravSrinivas
31. X post by @Steve_Yegge
32. X post by @omarsar0