The Quiet Singularity

Recursive Self-Improvement, Collapsing Costs, and the Feedback Loop No One Is Governing

Aaryush Gupta

February 2026

An observational analysis of three concurrent developments in artificial intelligence — recursive model self-improvement, the commoditization of frontier-class inference, and the compounding automation of cognitive labor — and their implications for institutions designed around assumptions of scarcity.

I. The Shift

We are entering a different phase of artificial intelligence. Not better assistants. Not faster code generation. For the first time, we are observing systems that meaningfully participate in building the next generation of themselves — not in theory, not in research papers, but in production environments at the companies creating them.[1][2]

The evidence for this is no longer speculative. It is publicly stated, on the record, by the leadership of the two most prominent frontier AI laboratories in the world. And it is occurring simultaneously with a collapse in the cost of intelligence that most observers have not yet internalized.

This document examines three developments that, taken together, suggest we have entered a compounding feedback loop — one that is accelerating faster than the institutional, policy, and cultural frameworks designed to govern it.

II. Recursive Self-Improvement

A. OpenAI: GPT-5.3-Codex

On February 5, 2026, OpenAI released GPT-5.3-Codex and made an extraordinary public claim: it was the first model in the company's history that was, in their words, "instrumental in creating itself."[1] The Codex engineering team used earlier versions of the model to debug its own training runs, manage its own deployment pipelines, and diagnose its own test results and evaluations.[2]

OpenAI CEO Sam Altman described this publicly as "the beginning of the intelligence explosion."[3] The model's technical report confirmed it served as the primary engineer for its own final optimization phase and deployment pipeline.[4]

This is not fully autonomous recursive self-improvement in the theoretical sense. But it is the first commercial proof point that AI-assisted AI development is no longer theoretical. If a coding model can meaningfully accelerate its own development, the pace of future improvements compounds.[2]

II. Recursive Self-Improvement (continued)

B. Anthropic: Claude Writing Claude

On February 3, 2026, Anthropic Chief Product Officer Mike Krieger stated at the Cisco AI Summit: "Right now for most products at Anthropic it's effectively 100% just Claude writing, and then what we've done is created all the right scaffolds around it to let us trust it."[5] He described engineers routinely shipping pull requests of 2,000 to 3,000 lines generated entirely by Claude.[6]

Boris Cherny, head of Anthropic's Claude Code division, confirmed he has not written a single line of code by hand in over two months. In a public post, he stated: "I shipped 22 PRs yesterday and 27 the day before, each one 100% written by Claude."[7] An Anthropic spokesperson confirmed the company-wide figure is between 70% and 90%.[7]

Perhaps most remarkably, approximately 90% of Claude Code's own codebase is now written by Claude Code itself.[7][8] Anthropic's Cowork product — launched January 12, 2026 — was built in ten days by four engineers, with most of the code written by Claude Code.[9][10]

These are not startups exaggerating capability. These are the two leading frontier AI laboratories describing what is already happening inside their own walls.

III. The Collapse of Cost

Recursive improvement alone would be significant. But it is occurring alongside a dramatic and accelerating collapse in the cost of frontier-class inference.

MiniMax M2: Near-Frontier Performance at Utility Pricing

MiniMax, a Shanghai-based AI company, released M2 — a 230-billion parameter mixture-of-experts model that activates only 10 billion parameters per token. It ranks in the top five globally on the Artificial Analysis Intelligence Index, scoring 61 — ahead of DeepSeek-V3.2 (57) and trailing Claude Sonnet 4.5 (63) and GPT-5 (69).[11] On agentic benchmarks including Terminal-Bench and SWE-bench Verified, it is competitive with frontier proprietary models.[11][12]

Its API pricing: **$0.30 per million input tokens** and **$1.20 per million output tokens**.[13][14] This represents approximately **8% of the cost** of Claude Sonnet 4.5 ($3.00 / $15.00 per million) with nearly double the inference speed.[15] The model is open-source under an MIT license, with weights freely available.

MiniMax M2.5, released days ago, is the first open-weights model to match Claude Sonnet on independent coding benchmarks.[16]

Model	Input / 1M tokens	Output / 1M tokens	Intelligence Index
Claude Sonnet 4.5	$3.00	$15.00	63
GPT-5 (thinking)	~$3.00	~$15.00	69
MiniMax M2	$0.30	$1.20	61

Table 1. Cost comparison across frontier-class models. Sources: Artificial Analysis[11], MiniMax official pricing[13], Perficient analysis[15].

IV. The Implication

Intelligence is no longer scarce. It is approaching the price profile of a utility — available to anyone, at any scale, for almost nothing.

I am experiencing this directly. I have not written a single line of code by hand in 2026. My speed and quality of execution are higher than at any point in my career. I am building more, shipping faster, and solving harder problems — not because I became a better engineer, but because the nature of what "engineering" means changed underneath me while I was doing it.

That last sentence should sit with you for a moment.

When execution becomes abundant, leverage shifts — toward judgment, taste, intent, and meaning. The things we spent decades treating as soft skills quietly became the only skills that differentiate human contribution from machine output.

V. The Feedback Loop

What is striking is not a single breakthrough. It is the structure of the dynamic itself.

Better tools build better tools. Those tools lower the cost of building. Lower costs expand access. Expanded access accelerates the next cycle. Each iteration makes the subsequent one faster, cheaper, and more capable — and the humans in the loop shift from writing to directing, from executing to deciding.

This is not a metaphor. It is the operational reality at the organizations building the most advanced AI systems on the planet. OpenAI explicitly described GPT-5.3-Codex as "the first self-developing AI coding model."[2] Anthropic's CPO confirmed "Claude is being written by Claude."[6] SemiAnalysis reports that 4% of all GitHub public commits are now authored by Claude Code, projecting 20%+ by end of 2026.[8]

And it is compounding.

VI. The Institutional Gap

We are not prepared for what this implies.

We do not have language for a world where cognitive labor compounds on itself. We do not have policy for an economy where the marginal cost of expertise trends toward zero. We do not have institutions designed for a rate of capability growth that outpaces the legislative, educational, and cultural systems meant to govern it.

We are still running on frameworks built for scarcity — of knowledge, of skill, of access — and those frameworks are dissolving faster than we are building replacements.

This does not feel like an explosion. It feels like a slope that just became steeper than anything underneath us was designed for.

And the unsettling part is not the speed. It is the quiet. It is how few people have noticed that the ground shifted.

We may look back on this period not as the moment AI arrived, but as the moment the assumptions underneath work, expertise, progress, and purpose began to shift — slowly at first, and then all at once.

References

[1] NBC News. "OpenAI says new Codex coding model helped build itself." February 5, 2026.
https://www.nbcnews.com/tech/innovation/openai-says-new-codex-coding-model-he...

[2] The New Stack. "OpenAI's GPT-5.3-Codex helped build itself." February 5, 2026.
https://thenewstack.io/openais-gpt-5-3-codex-helped-build-itself/

[3] Let's Data Science. "GPT-5.3 Codex Explained: OpenAI's Self-Developing Agent." February 2026.
https://www.letsdatascience.com/blog/gpt-5-3-codex-openai-just-released-an-ai...

[4] Creati.ai. "OpenAI Launches GPT-5.3-Codex: Revolutionary AI Model That Helped Build Itself." February 6, 2026.
https://creati.ai/ai-news/2026-02-06/openai-gpt-5-3-codex-launch-self-buildin...

[5] IT Pro. "Anthropic Labs chief Mike Krieger claims Claude is essentially writing itself." February 2026.
https://www.itpro.com/software/development/anthropic-labs-chief-mike-krieger-...

[6] TechStory. "'Claude Writing Claude': The Code of Anthropic is Now Nearly 100% AI-Generated." February 9, 2026.
https://techstory.in/claude-writing-claude-the-code-of-anthropic-is-now-nearl...

[7] Fortune. "Top engineers at Anthropic, OpenAI say AI now writes 100% of their code." January 29, 2026.
https://fortune.com/2026/01/29/100-percent-of-code-at-anthropic-and-openai-is...

[8] SemiAnalysis. "Claude Code is the Inflection Point." February 2026.
https://newsletter.semianalysis.com/p/claude-code-is-the-inflection-point

[9] Axios. "Anthropic's Claude Cowork wrote itself." January 13, 2026.
https://www.axios.com/2026/01/13/anthropic-claude-code-cowork-vibe-coding

[10] The Pragmatic Engineer. "How Claude Code is built." 2025.
https://newsletter.pragmaticengineer.com/p/how-claude-code-is-built

[11] DeepLearning.ai / The Batch. "MiniMax-M2's Lightweight Footprint and Low Costs Belie Its Top Performance." November 6, 2025.
https://www.deeplearning.ai/the-batch/minimax-m2s-lightweight-footprint-and-l...

[12] Artificial Analysis. "MiniMax-M2 — Intelligence, Performance & Price Analysis." 2025.
https://artificialanalysis.ai/models/minimax-m2

[13] MiniMax Official Pricing. "Pricing — MiniMax API Docs."
https://platform.minimax.io/docs/guides/pricing

[14] MiniMax Official. "MiniMax M2 & Agent: Ingenious in Simplicity."
https://www.minimax.io/news/minimax-m2

[15] Perficient. "Minimax M2: Innovative Reasoning Strategy from Open-Source Model." November 19, 2025.
https://blogs.perficient.com/2025/11/19/minimax-m2-open-source-interleaved-re...

[16] OpenHands. "MiniMax M2.5: Open Weights Models Catch Up to Claude Sonnet." February 12, 2026. https://openhands.dev/blog/minimax-m2-5-open-weights-models-catch-up-to-claude

This isn't speculation.

These are the companies' own words.

I don't have answers. But I am increasingly worried that we are not even asking the right questions yet.

Aaryush Gupta

linkedin.com/in/aaryush