

RECURSIVE SELF-IMPROVEMENT (RSI)

The recursion premium: self-improving AI is repricing the startup world

AI now writes most of the code at the lab that builds it, a dedicated recursive self-improvement lab has opened in Tokyo, and Nvidia is paying billions for equity in labs whose only product is faster research. Recursive self-improvement has moved from thought experiment to line item, and investors are pricing pre-revenue superintelligence bets above \$25 billion on the strength of it. Over the next two to five years that premium will mint a new class of infrastructure startups while compressing the moats of almost everyone else. Whether the loop actually closes is now the most expensive open question in venture capital.

AUTHOR



Alicante Tech Vibes *Research*
Labs

~9 MIN READ
DATED

• ALL FIGURES SOURCED &

80%+ OF CODE MERGED AT ANTHROPIC NOW AUTHORED BY CLAUDE (MAY 2026)	320 min CLAUDE OPUS 4.5 AUTONOMOUS TASK HORIZON, METR TH1.1 (JAN 2026)	88.6 days TIME-HORIZON DOUBLING RATE FOR MODELS SINCE 2024 (METR)
\$242B VENTURE CAPITAL INTO AI IN Q1 2026 ALONE – 80% OF ALL VC (CRUNCHBASE)	\$32B SAFE SUPERINTELLIGENCE VALUATION WITH ZERO PRODUCT (JUL 2026)	+30 pts SWE-BENCH GAIN SAKANA'S DARWIN GÖDEL MACHINE EARNED BY REWRITING ITSELF

I. The machine starts feeding itself

Recursive self-improvement — RSI, the idea that an AI system can improve the process that builds the next AI system — spent two decades as a philosophical argument. In 2026 it acquired dated, auditable evidence. On June 5, Anthropic published a paper, *When AI builds itself*, disclosing that more than 80% of code merged into its production codebase is now authored by Claude, up from low single digits when Claude Code launched in February 2025; typical engineers merged eight times more code per day in Q2 2026 than in 2024. The company that most visibly benefits from the loop paired the disclosure with a call for a verifiable global brake on frontier development.

The loop is not only running inside the big labs. Tokyo-based Sakana AI formally opened a dedicated RSI Lab in mid-2026, built on results like the Darwin Gödel Machine — an agent that repeatedly rewrote its own codebase and lifted its SWE-bench software-engineering score from 20% to 50% without human intervention — and an AI Scientist system whose paper was published in Nature on March 26, 2026. And the capability floor keeps rising on schedule: METR's Time Horizon 1.1 update (January 29, 2026) measured Claude Opus 4.5 completing tasks that take human experts roughly 320 minutes, with the horizon doubling every 88.6 days for models released since 2024. **The signal is not any single result; it is that self-improvement stopped being a metaphor and became a measured input to production.**

FIG. 1: HOW LONG A TASK CAN FRONTIER MODELS FINISH ALONE? (HUMAN-EXPERT MINUTES, 50% SUCCESS – METR TH1.1)





BARS SCALED TO OPUS 4.5 = 320 MIN. CONFIDENCE INTERVALS ARE WIDE (170-729 MIN FOR THE TOP BAR). SOURCE: METR, TIME HORIZON 1.1, JAN 29, 2026.

II. Compute becomes the casting vote

If models can convert compute into research progress with less human labor in the loop, then whoever allocates compute picks the winners. Nvidia grasped this before most investors: on July 27, 2026 it invested \$5 billion in Ilya Sutskever's Safe Superintelligence — a company with no product and a \$32 billion valuation — granting it access to the Vera Rubin platform and, per the company, an order of magnitude more compute. Nvidia had already put roughly \$800 million into Reflection AI, the open-weight lab founded by ex-DeepMind researchers, which by March 2026 was raising \$2.5 billion at a \$25 billion pre-money valuation — up from about \$545 million a year earlier, a ~45x markup on essentially the same team.

This equity-for-compute pattern redraws the map. The old landscape question was open versus closed models; the new one is who can afford to run automated researchers at scale. Sakana is making the contrarian bet, arguing its sample-efficient methods can deliver RSI to institutions "beyond the world's two largest compute clusters." If that holds, RSI diffuses and prices fall; if it does not, the frontier consolidates around three or four compute-rich labs and their chip supplier — which is also, increasingly, their shareholder.

III. What the loop can and cannot do yet

The honest technical read is narrower than the headlines. Today's "self-improvement" is mostly AI accelerating the *engineering* around AI — writing code, running ablations, tuning scaffolds — not autonomously producing the conceptual breakthroughs that define new model generations. Sam Altman's own framing concedes the gap: OpenAI targets an "intern-level" research assistant by September 2026 and a "legitimate AI researcher" only by 2028.

The skeptic case has data behind it. METR's own randomized study of experienced open-source developers (July 2025) found they were 19% *slower* when using AI tools — while believing they were faster — a warning that self-

reported productivity inside labs deserves discounting. Nathan Lambert's [catalog of brakes on an intelligence explosion](#) notes that research progress depends on messy data intuition and that labs devote surprisingly little compute to research itself. And the most rigorous economic treatment to date, [Whitfill and Wu \(August 2025\)](#), finds the answer hinges on one parameter: if compute and cognitive labor are substitutes, a software-only explosion is plausible; their frontier-experiments specification instead estimates them as near-perfect complements — meaning armies of automated researchers still queue for the same GPUs. METR's headline numbers also carry [wide confidence intervals](#), and only 5 of its 31 longest tasks have real human baselines. The loop is real; its slope is genuinely disputed.

“The world needs a verifiable global mechanism to slow or temporarily pause frontier AI development.”

ANTHROPIC, "WHEN AI BUILDS ITSELF," JUNE 5, 2026 — PUBLISHED ALONGSIDE ITS DISCLOSURE THAT CLAUDE AUTHORS 80%+ OF ITS OWN PRODUCTION CODE. THE LAB RUNNING THE LOOP FASTEST ASKED FOR A BRAKE.

IV. The recursion premium

Capital markets are already paying for the loop. Global venture funding hit [\\$300 billion in Q1 2026](#), an all-time quarterly record, with [\\$242 billion — 80% — going to AI](#), up from a 55% share a year earlier. Four rounds (OpenAI's \$122B, Anthropic's \$30B, xAI's \$20B, Waymo's \$16B) absorbed \$188 billion of it, and late-stage funding grew 205% year over year while seed grew just 31%. Call the gap between these valuations and any defensible revenue multiple the **recursion premium**: the price investors pay for the possibility that a lab's next model builds the one after that.

The premium is easiest to see where there is no revenue at all. SSI at \$32 billion with zero product is not a bet on a product roadmap; it is a bet that research velocity itself compounds. Reflection's 45x markup in twelve months prices the same thesis. Even the application layer trades on it: Cursor-maker Anysphere was in talks in April 2026 to raise at a [\\$50 billion valuation](#), on the logic that AI-written software is the first mass market the loop touches. If the frontier-experiments economics of Whitfill and Wu is right, much of this premium is

misallocated — cognitive labor without matching compute buys little — and the correction lands hardest on pre-revenue labs.

V. Startups in the shadow of the loop

For founders and investors outside the frontier labs, RSI cuts both ways. The threat first: if lab models improve at METR-measured rates — task horizons doubling every three to seven months — then any startup whose moat is "a wrapper of engineering effort around a model" is building on melting ice. Feature roadmaps that took a Series A to execute in 2024 are a fine-tune away for a frontier lab in 2027. Expect smaller rounds to build the same product, more seed-strapped teams, and diligence that asks "what happens to you when the model gets 10x better?" as the first question rather than the last.

The opportunity is the picks-and-shovels layer that the loop itself demands. Every automated-researcher pipeline needs evaluation and monitoring infrastructure (METR-style task suites are becoming an industry input, not an academic curiosity), reinforcement-learning environments and verified reward signals, agent orchestration and audit tooling, and safety verification that enterprises and regulators will require before letting self-modifying systems near production. These markets share a useful property: they get *bigger* the faster the frontier moves, whichever lab wins. The precedent worth holding in mind is the late-1990s fiber buildout — capital raced far ahead of demand, most operators died, and the infrastructure repriced to become the substrate of the next two decades. A recursion-premium correction would likely rhyme: the compute, the tooling, and the trained models survive their cap tables.

VI. How it plays out: three scenarios

Horizon 2026–2030. Probabilities are analytical judgment, not measurement.

1 · Compounding, not exploding ~55%

2026–2030 · STEADY

AI R&D automation keeps accelerating engineering but stays complementary to compute and human judgment; no discontinuity. The recursion premium partially deflates — one or two \$20B+ pre-revenue labs take down rounds — while eval, RL-environment, and agent-infrastructure startups quietly become the decade's best seed vintage. Watch: METR doubling times holding near 7 months rather than shortening; OpenAI's September 2026 "intern" landing as a useful tool, not a researcher; AI's share of VC settling between 60–80%.

2 · The loop tightens ~25%

2027–2029 · FAST

Automated researchers start contributing genuine algorithmic advances; horizon doubling drops below three months and lab model-release cadence visibly shortens. Capital hyper-concentrates: frontier labs plus Nvidia capture most AI value, app-layer startups consolidate or die, and governments treat RSI capability as strategic infrastructure. Watch: labs attributing named architectural improvements to AI systems; a second Anthropic-style disclosure putting AI-authored research (not just code) above 50%; emergency-tempo policy responses echoing Anthropic's pause proposal.

3 · The premium pops ~20%

2026–2028 · CORRECTION

Compute complementarity binds (the Whitfill–Wu frontier-experiments case), benchmark gains stop translating into revenue, and the 19%-slower result generalizes: measured productivity disappoints. Late-stage AI funding contracts sharply from its 205% growth rate; SSI-style bets mark down; surviving startups are the ones selling measurable cost reduction today. Watch: METR TH1.2 showing lengthening doubling times; Reflection's \$25B round failing to close; Q1 2027 AI funding falling below 50% of global VC.

WHAT WOULD FALSIFY THIS THESIS

The thesis: measurable AI self-improvement is real enough to restructure startup economics within five years. It fails if the loop's output is illusory — if AI-authored code share rises while lab research output per dollar stays flat, or if METR's next update shows doubling times lengthening rather than holding. The strongest counter-case is already published: Whitfill & Wu's complements estimate implies cognitive labor without compute buys almost nothing, and METR's developer study showed AI tools making experts 19% slower while feeling faster. If enterprise revenue attributable to agentic coding stalls through 2027 while valuations keep climbing, this brief described a bubble, not a regime change.

WHAT TO WATCH, IN ORDER

1. **OpenAI's September 2026 deadline.** Altman promised an "intern-level" automated research assistant by next month — checkable within weeks via OpenAI announcements and independent researcher reports.
2. **METR Time Horizon 1.2** (metr.org/time-horizons, expected H2 2026): does the ~88-day post-2024 doubling hold, shorten, or lengthen? Single most load-bearing public number in the RSI debate.
3. **Reflection AI's \$2.5B round:** whether it closes at or above \$25B (press and filings, Q3–Q4 2026) is the cleanest live read on the recursion premium's health.
4. **Sakana RSI Lab's first publications** (sakana.ai, ongoing): do Darwin-Gödel-style self-modification gains transfer beyond SWE-bench to open-ended research tasks — and at low compute?
5. **Crunchbase Q3 2026 data** (October 2026): AI's share of global venture above or below 80%, and whether seed-stage growth keeps lagging late-stage by 6x.

SOURCES

1. Anthropic Institute, "When AI builds itself," June 5, 2026, via The Next Web · thenextweb.com
2. METR, "Time Horizon 1.1," January 29, 2026 · metr.org
3. Sakana AI, "Introducing Sakana AI's Recursive Self-Improvement (RSI) Lab," June 2026 · sakana.ai

4. Zhang, Hu, Lu et al., "Darwin Gödel Machine," arXiv:2505.22954, 2025 · arxiv.org
5. TechCrunch, "Ilya Sutskever's Safe Superintelligence partners with Nvidia," July 27, 2026 · techcrunch.com
6. Crunchbase News, "Q1 2026 Shatters Venture Funding Records," April 2026 · news.crunchbase.com
7. The AI World, "Reflection AI Eyes \$2.5B Round at \$25B Valuation," March 30, 2026 · theaiworld.org
8. TechCrunch, "Cursor in talks to raise \$2B+ at \$50B valuation," April 17, 2026 · techcrunch.com
9. TechCrunch, "Sam Altman says OpenAI will have a 'legitimate AI researcher' by 2028," October 28, 2025 · techcrunch.com
10. Nathan Lambert, "State of play of AI progress (and related brakes on an intelligence explosion)," Interconnects, April 30, 2025 · interconnects.ai
11. Whitfill & Wu, "Will Compute Bottlenecks Prevent an Intelligence Explosion?," arXiv: 2507.23181, August 2025 · arxiv.org
12. METR, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," July 10, 2025 · metr.org

METHOD NOTE: FIGURES ARE ATTRIBUTED AND DATED INLINE; WHERE INDEPENDENT SOURCES DISAGREE, THE DISAGREEMENT IS REPORTED RATHER THAN AVERAGED. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT.

Built by the [Alicante Tech Vibes](#) community, a tope.

2026-08-12