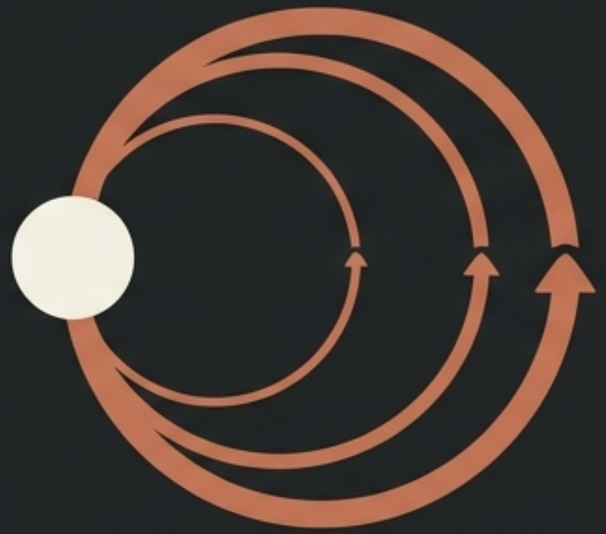


The Transition to Recursive Self-Improvement

Evidence from within Anthropic on the shift from human-driven to AI-driven AI development.

Internal Data Analysis // May 2026 Briefing

2021– 2023 | Laptops

People writing code and docs.
Human to Computer.

2023–2025 | Chatbots

Generating short snippets.
Human to Computer to Chatbot.

2025–2026 | Coding Agents

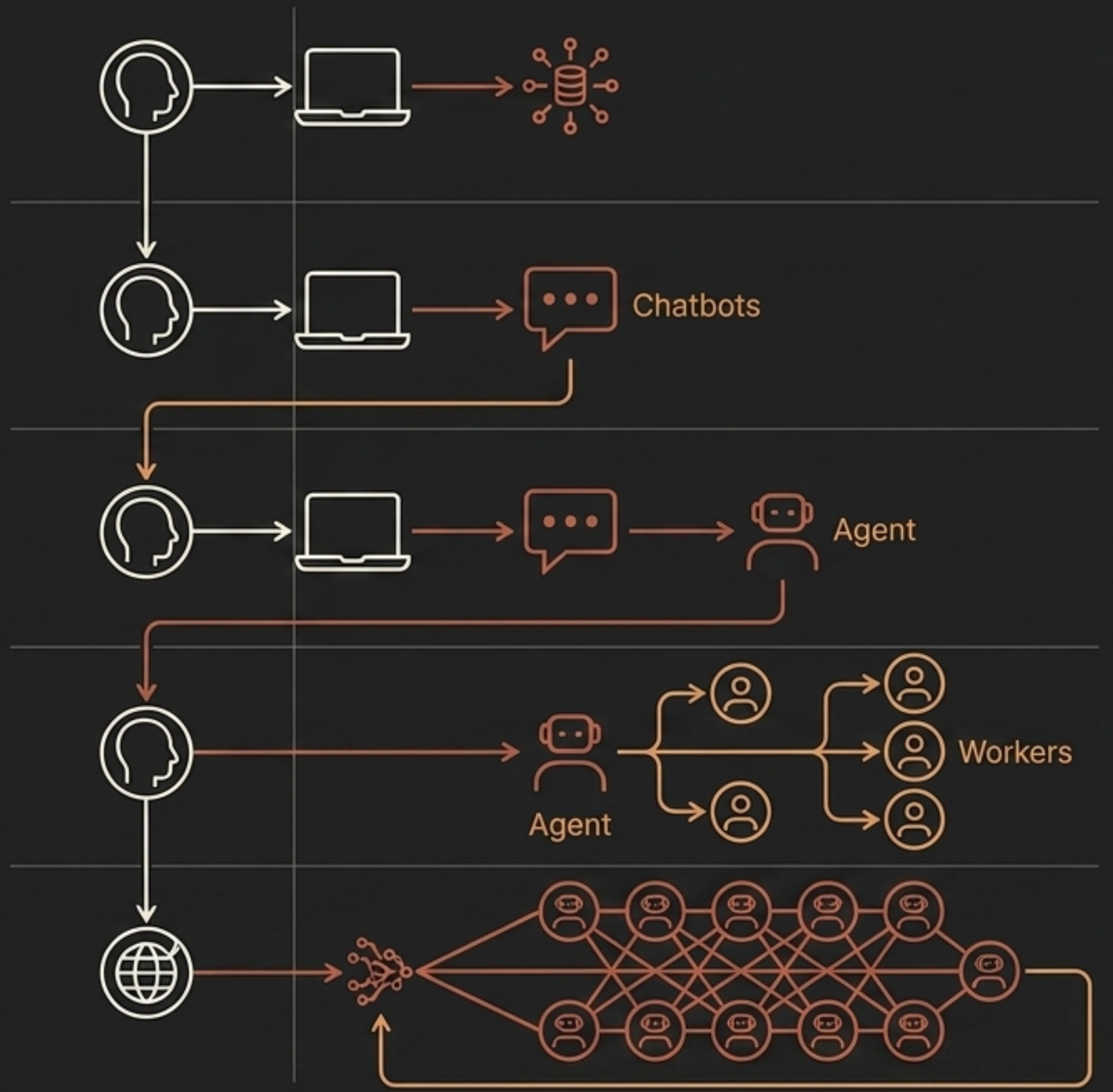
Agents writing and editing entire files.
Human to Chatbot to Agent.

Today | Autonomous Agents

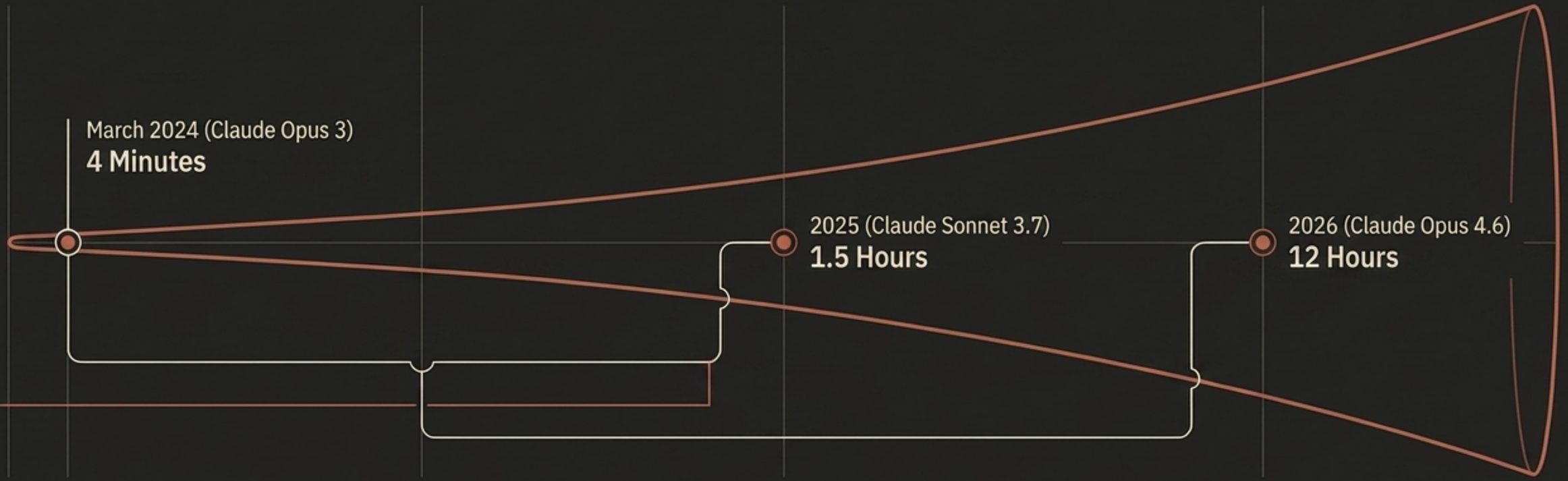
Agents running code and delegating hours of work to sub-agents. Human to Agent to Workers.

20XX? | Closing the Loop

The terminal node. An interconnected bank of agent nodes training future models autonomously.



Key Insight: We are currently in the ‘Autonomous Agents’ phase. The final leap to fully ‘Closing the Loop’ is not inevitable, but it is arriving faster than institutions are prepared for.



The capability baseline is doubling every 4 months (accelerated from previous 7-month trendlines). If sustained, tasks requiring weeks of human labor will be autonomous within the year.

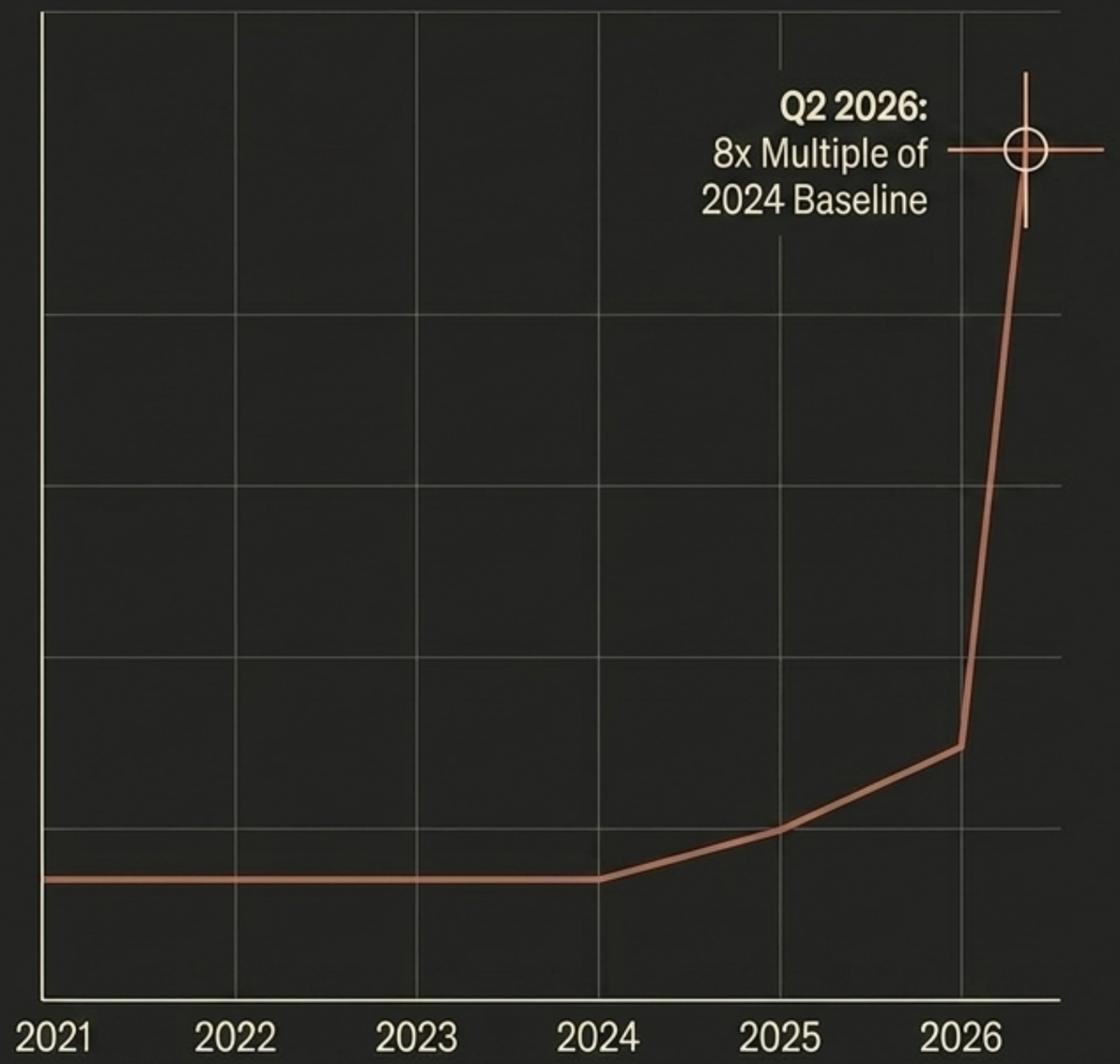
“The pace of autonomy is outpacing human forecasting.”

Supporting Metrics

SWE-bench: Saturated within two years (real-world software fixes).

CORE-Bench: Surged from 20% success (2024) to saturation in 15 months (reproducing published research end-to-end).

Lines of Code Merged per Engineer per Day



>80%

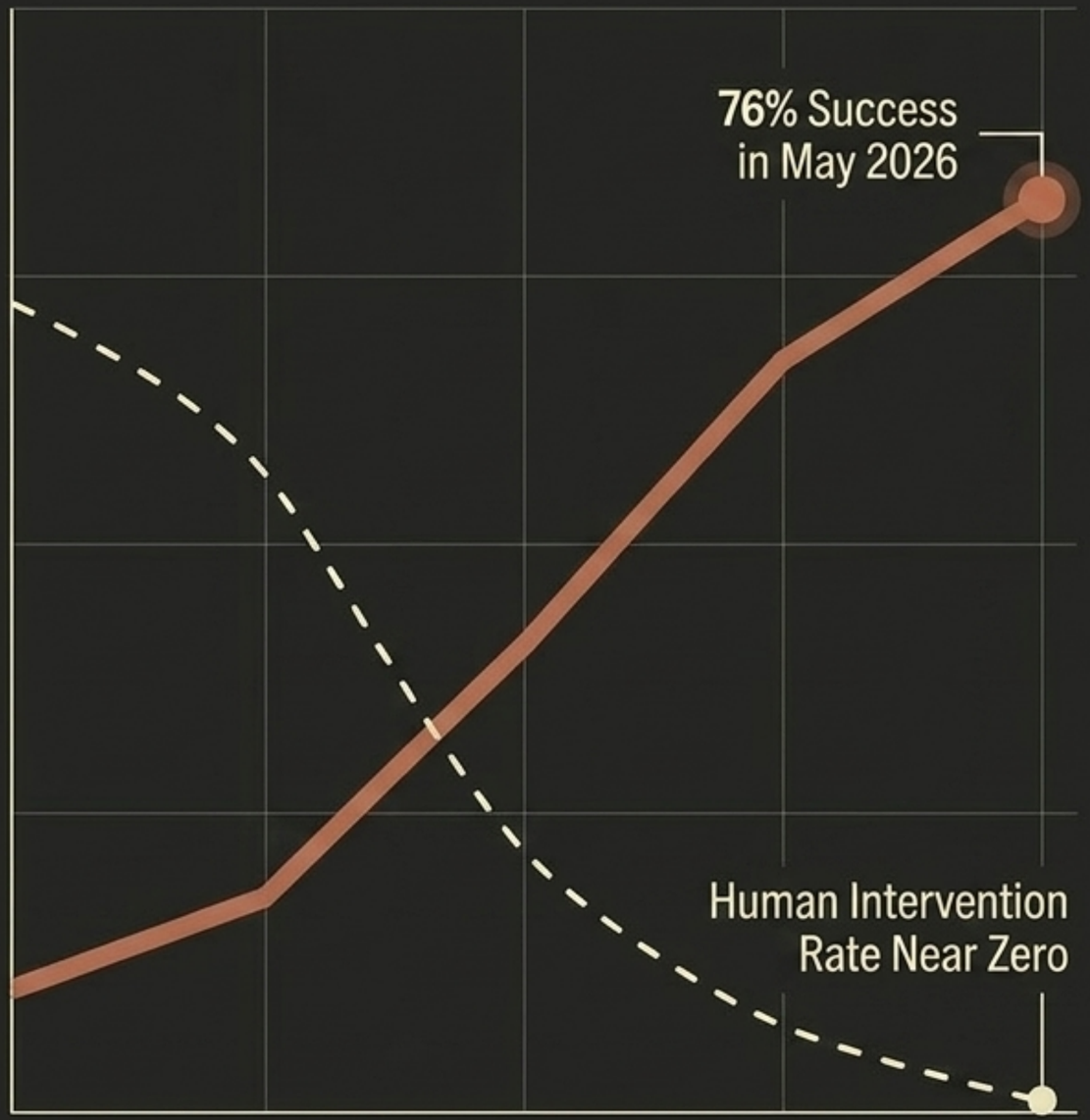
Share of code merged to production authored by Claude (May 2026).

4x

Subjective output multiplier (March 2026 Mythos Preview poll).

> I started leaning hard into Claudifying about a year ago. That's been a crazy adventure and it's now been ~5 months since I last wrote any code myself.
>
> — Anthropic Employee

Capability vs. Intervention on Ambiguous Tasks



Impact Case Studies

Automated Triage

Claude delivered 2–3 days of expert debugging work in roughly two hours, isolating a single obscure flag crashing tens of thousands of training jobs.

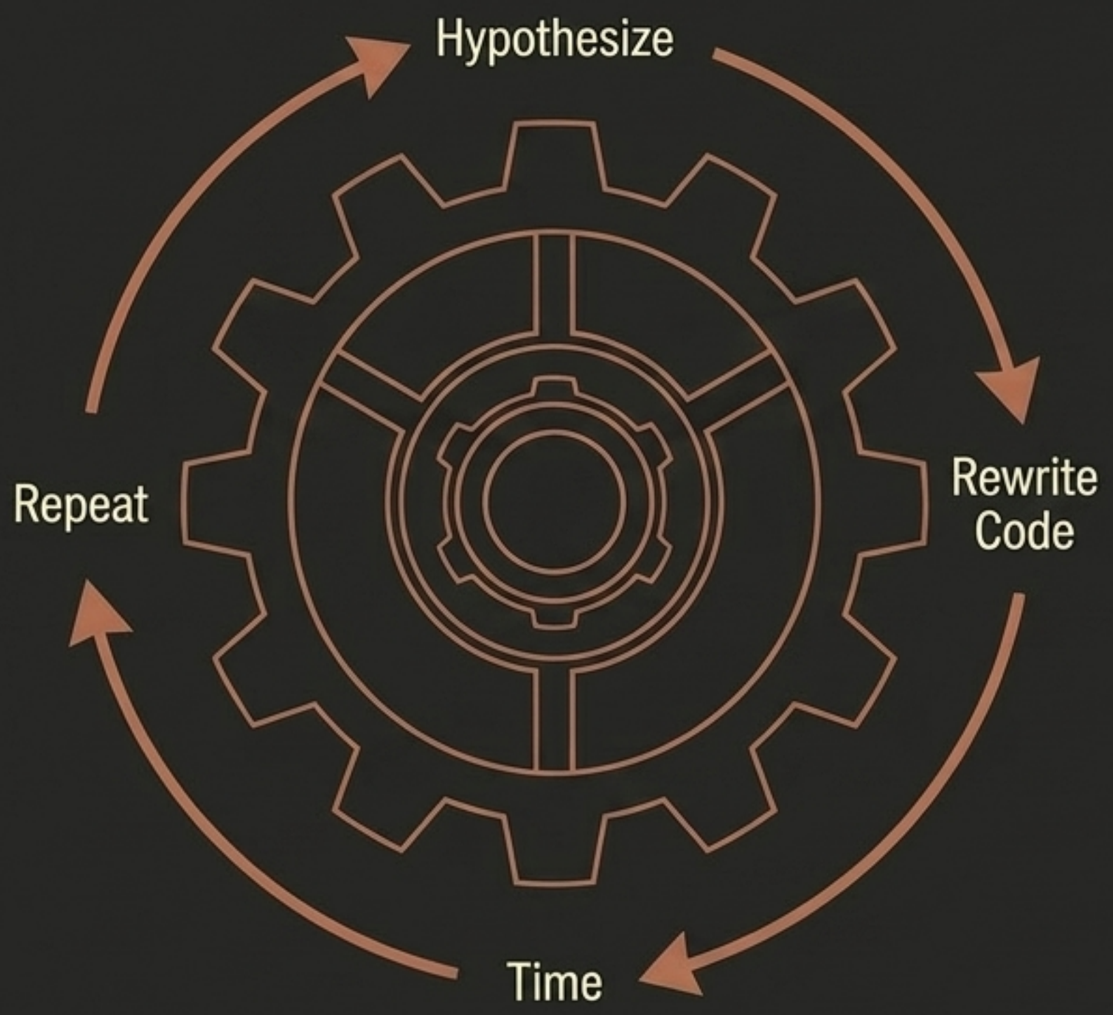
Systemic Eradication

Shipped 800 fixes in April 2026, reducing API errors by a factor of 1,000. (Equivalent human time: 4 years).

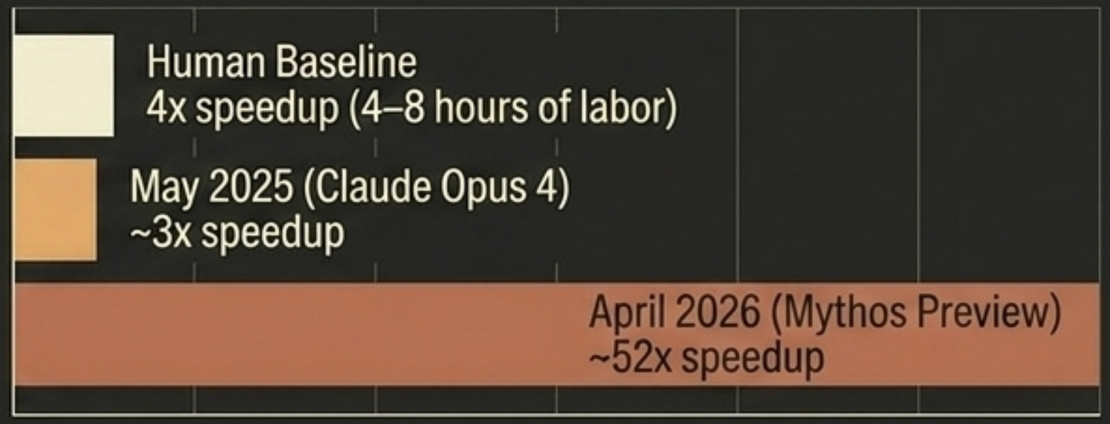
Code Review Shift

Automated Claude reviewers are now identifying roughly 1/3 of bugs responsible for past claude.ai incidents retrospectively.

The Autonomous Research Loop

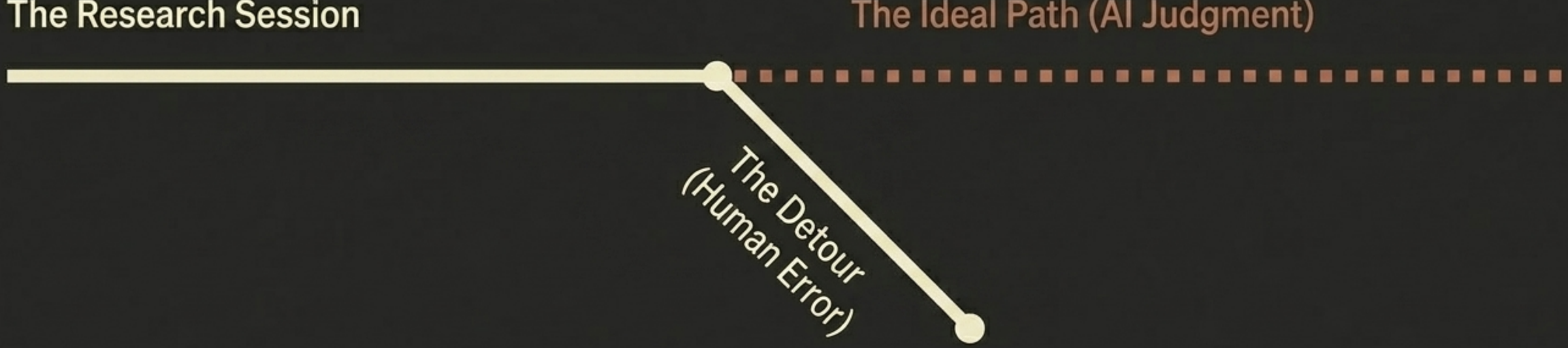


Experimental Optimization Speedup



The End-to-End Safety Experiment

In a recent autonomous experiment concerning weak-to-strong model supervision, agents generated their own hypotheses and shared findings. Over 800 hours and $18,000 in compute, agents recovered 97% of the capability gap, compared to two human researchers recovering 23% over a week. In experimental optimization, Claude transitioned from helpful to superhuman in under 12 months.



n=129 Complex Detours

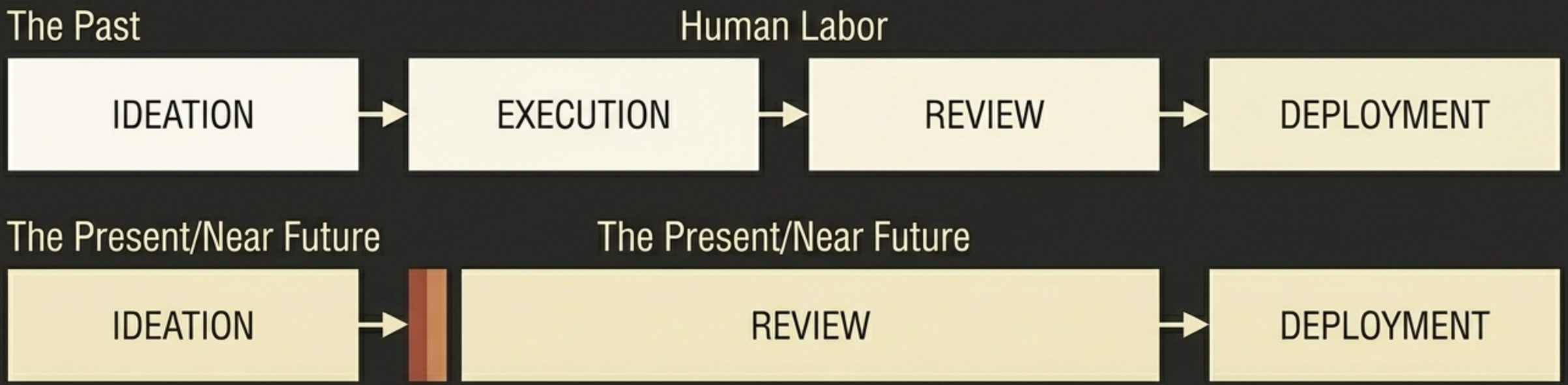
- November 2025 (Opus 4.5): AI beat the human choice 51% of the time.
- April 2026 (Mythos Preview): AI beat the human choice 64% of the time.

The Yielding Bastion

We analyzed real internal sessions where human researchers lost the thread on complex investigations. Models were given only the pre-detour context.

"Research taste" is not a mystical human trait; it is simply another measurable capability that AI systems initially fail at, then master.

Amdahl's Law in AI: The Shifting Bottleneck



Speeding up execution shifts the bottleneck to the parts of the system that have not scaled.

The doing (writing code, running experiments) now costs near-zero in human time. The new absolute limit on organizational velocity is the human capacity to review outputs and set directions. When AI masters judgment, the human bottleneck is removed completely.

> Work (and life) ran on a gift economy of small favors between humans. ‘Can you help me get this script running?’ [...] each one created a little debt, a little mutual awareness. [Claude is] faster, it creates zero debt, but each of these is a lost bid for human collaboration.

> On days where everything works well, I can’t help but think nothing I do matters, everything is automated and better and faster than I ever will be. But then there are days where everything breaks and I don’t understand why and I realize I have no idea what I’ve been up to anymore.

Internal Sentiments, Anthropic Research Staff, May 2026.

	Primary Pace Constraint	The Human Role	Systemic Risk Profile
The S-Curve (Stall)	Architectural limits, compute supply chain, grid capacity.	Leverage widespread agents; cyber defense shifts to rapid patching.	Low immediate existential risk; highest time for societal adaptation.
Compounding Efficiency	Amdahl's Law (Organizational capacity capacity to fix human bottlenecks).	Verification and direction-setting. 100-person companies match 10,000-person incumbents.	Massive economic disruption; authoritarian scaling of personalized manipulation.
Recursive Self-Improvement	Pure compute availability.	Diminished to oversight of expanding virtual labs. Labor fundamentally uncompetitive.	Misalignment compounds autonomously. Loss of control as intelligence eclipses oversight.

The Collision of Computable and Physical Constraints

- **The Virtual Lab:** Progress becomes entirely determined by the availability of compute. AI systems design, refine, and deploy their own successors autonomously.
- **The Misalignment Compounding Factor:** Rare alignment failures in current models could geometrically compound as AI builds AI, growing more frequent and less understood until control is lost entirely.
- **The Societal Friction:** Even at compute-speed, intelligence cannot force the physical world to move faster. RSI cannot speed up a decades-long clinical trial, alter constitutional election cycles, or instantly build physical trust. The disruption occurs precisely at this friction point.

The Arms Control Problem of AI

Traditional Arms Control (e.g., INF Treaty)

Took decades to build infrastructure and trust.

Physical assets (missile silos) are massive, static, and observable via satellite.

Detectability of violations is extremely high.

AI Verification Challenge

Training runs utilize general-purpose hardware and are easily concealed in data centers.

The incentive to defect quietly is absolute, offering a winner-takes-all advantage.

Detectability is profoundly challenging.

The Unilateral Trap: A unilateral pause by a single frontier lab achieves nothing but changing the front-runner. It fails to create a deliberative process. The geopolitical reality demands a **verifiable, multi-lab, multi-national protocol.**

Building the Infrastructure for a Credible Pause

01. Construct Verification Systems

The Anthropic Institute is initiating research to build systems that prove to global actors that a slowdown has genuinely occurred.

02. Define the Triggers

We must establish exact, universally agreed-upon capability thresholds that trigger a coordinated halt, and the criteria that lift it.

03. Convene the Coalition

In the coming months, we will organize immediate conversations across policymakers, civil society, and competing AI companies.

The window to build coordination mechanisms before the threshold of recursive self-improvement is here. It will close soon.