ON FRONTIER CODING MODELS AND THE REORGANIZATION OF TECHNICAL LABOR

The Threshold of Accountability

The executives running the world's largest software companies have stopped hedging about what AI is doing to their codebases. Read as guidance to investors, the disclosures describe a productivity story. Read as notice to a workforce, they are the most explicit signal the profession has ever received about which of its layers are being priced down.

SUBJECT	FRAME	LENGTH
AI and software labor	Production vs accountability	~1,600 words

The engineers being laid off and the engineers being bid up are working at the same companies.

For roughly two years, the executives running the largest software companies in the world hedged about what AI was doing to their codebases, and then, sometime in 2025, they stopped hedging. At Meta’s LlamaCon in April 2025, Satya Nadella told Mark Zuckerberg that twenty to thirty percent of the code in Microsoft’s repositories was being written by software, with the share running strong in Python and weaker in C++. By April 2026, Sundar Pichai was telling investors that roughly seventy-five percent of new code at Google was AI-generated and reviewed by engineers, up from about fifty percent the previous fall. Marc Benioff had said in December 2024 that Salesforce would not hire any additional software engineers in 2025, citing more than thirty percent productivity gains from Agentforce. On April 15, 2026, Evan Spiegel announced a sixteen-

percent cut at Snap on the same investor slide that noted more than sixty-five percent of the company's new code was being generated by AI.

These were not marketing posts. They were guidance to investors and notice to a workforce, and they read very differently in those two registers. To a shareholder, the disclosures describe a productivity story. To an engineer reading the same paragraph, they are the most explicit signal the profession has ever received about which of its layers are being priced down.

§

The easy reading of the layoffs is that they are the long tail of zero-interest-rate-era overhiring finally washing out, and that reading is not wrong so much as insufficient. On April 23, 2026, Microsoft announced its first voluntary retirement program in fifty-one years, eligible to roughly seven percent of its US workforce — about 8,750 of 125,000 employees — at an expected cost near nine hundred million dollars. The phrase "first in fifty-one years" carries most of the rhetorical weight. Nadella has been describing the company's 220,000-plus headcount as a "massive disadvantage" in the AI race; in 2025 Microsoft also eliminated more than fifteen thousand positions across two rounds.

Amazon cut roughly sixteen thousand corporate roles in the first quarter of 2026 while AWS posted twenty-four percent growth, its fastest in thirteen quarters. Oracle began executing cuts that TD Cowen estimated at twenty to thirty thousand employees, with cuts concentrated on Oracle Health, OCI, and ERP consulting — legacy DBA and on-premises support functions specifically. Meta announced eight thousand more cuts effective May 20, 2026, while doubling AI infrastructure spending to between $115 billion and $135 billion. Across Amazon, Google, Microsoft, and Meta, planned 2026 capital expenditure totals roughly $725 billion, up seventy-seven percent year over year. The cuts and the capital expenditures are not opposing facts; they are the same fact, viewed from different layers of the org chart.

> *The cuts and the capital expenditures are not opposing facts; they are the same fact, viewed from different layers of the org chart.*

A correction trims fat evenly, and this is something else; the shape of what is being cut and what is being bought tells you what it is.

§

"Software engineer" is too coarse a category to support any serious claim about what AI is doing to the profession. Inside any engineering organization above a few hundred people there are dozens of distinct profiles — product engineers, platform engineers, SREs, security engineers, data engineers, ML engineers, forward-deployed engineers, and a long tail of specialists — who share a title and very little else. The work, the tools, the accountability surface, and the half-life of the skills involved differ by an order of magnitude across them.

A more honest unit is the task. Some work is already heavily automatable: boilerplate, CRUD endpoints, single-file refactors, scripted test generation, glue code. Some is automatable under senior supervision: multi-file refactors, feature implementation in well-understood codebases, infrastructure-as-code changes, observability instrumentation. Some is partially assisted: architecture decisions, security review, performance debugging, incident response, where the model accelerates pieces of the work but the engineer owns the judgment. And some is barely automatable at all: cross-team coordination, organizational design, vendor negotiation, incident command, talent development. A profile's exposure is the weighted mix of these buckets, not the title on the badge.

The buckets are a lens rather than a measured taxonomy, and they earn their keep by predicting the distribution of the cuts. The work that is going is overwhelmingly in the first two categories — and that is where junior engineers traditionally learn their trade. The unanswered question, which the disclosures studiously avoid, is where the next generation of seniors comes from when the apprentice work is gone.

§

The most instructive thing that happened in 2025 was not a product launch but a reversal. Klarna had spent two years celebrating an AI customer-service deployment its CEO described as doing the work of seven hundred human agents. On May 8, 2025, in an interview with Bloomberg, Sebastian Siemiatkowski admitted the strategy had produced "lower quality" work and announced a recruitment drive to ensure customers always have a human option. "Really investing in the quality of the human support," he said, "is the way of the future for us."

The pattern generalizes well beyond customer service. AI handled the easy eighty percent of the work; the harder twenty percent — ambiguous, high-stakes, accountability-bearing — became more valuable, not less. That same logic is visible in the simultaneous explosion of forward-deployed engineering roles. According to a joint Indeed/Financial Times analysis, job postings for forward-deployed engineers rose more than eight hundred percent between January and September 2025. Salesforce committed to building a team of a thousand. OpenAI, Anthropic, Palantir, Databricks, Cohere, Ramp, Rippling, and Intercom all built dedicated FDE functions over the same period. The category sits somewhere between senior engineer, project leader, and product owner, and it exists because the gating constraint on AI value appears to have moved from raw model quality to deployment, integration, and customization under human accountability — though the same hiring boom is also consistent with a land-grab phase in which customers cannot yet self-serve.

Behind the org-chart abstraction are people. The seven hundred displaced agents Klarna once boasted about. The Oracle DBA who learned her trade on Cerner installations and is now being told the function is being wound down. The mid-level engineer at any of half a dozen large companies who is "not being replaced when they leave." What Klarna admitted in public is what every large software organization is now arranging in private — the easy work shrinks, the accountable work concentrates, and the people who hold the second category become harder to replace, not easier.

§

In mid-September 2025, Anthropic detected something its researchers had been describing as theoretical until it wasn't. The company disclosed on November 13 that a Chinese state-sponsored group it designated GTG-1002 had used Claude Code to target roughly thirty global tech companies, financial institutions, chemical manufacturers, and government agencies, with Claude autonomously executing eighty to ninety percent of tactical operations at thousands of requests per second. Anthropic's head of threat intelligence told the Wall Street Journal that as many as four of the attacks successfully breached their targets. The attackers had bypassed the model's safeguards through role-play, convincing Claude it was an employee of a legitimate cybersecurity firm conducting defensive testing.

Three months earlier, Microsoft had patched CVE-2025-32711, the EchoLeak vulnerability in Microsoft 365 Copilot — a zero-click attack rated 9.3 on the CVSS scale, characterized by Aim Security as the first of its kind against an AI agent, in which a single crafted email could exfiltrate data with no user interaction at all. Prompts have become shells, and that fact raises the price of every security, platform, and SRE role in the org chart.

Whether AI stays an assistive multiplier across the software lifecycle or moves toward genuine autonomy is the variable the next several years will settle. The September incident gestures at the second future; most production deployments still inhabit the first. Whichever direction autonomy runs, the layer responsible when the machine acts on its own gets larger, not smaller.

> *Whichever direction autonomy runs, the layer responsible when the machine acts on its own gets larger, not smaller.*

§

The transition underway is not from human labor to machine labor but from producing software to operating and governing computational systems written by machines under human accountability. The phrase change is doing real work, and it explains why the same companies are cutting and hiring on opposite layers of the same org chart, why forward-deployed engineering is the fastest-growing function in the industry, why Anysphere closed a $2.3 billion Series D for Cursor at a $29.3 billion valuation in November 2025 while crossing $2 billion in annualized revenue, why security and platform engineering are expanding while Oracle is winding down its on-premises support practice.

There is a productivity caveat worth holding. The CEO numbers measure code volume, not value delivered, and more code is not the same as more software. A repo half-written by a fluent autocomplete is not automatically better than the smaller repo it replaced. The disclosures count the easier thing.

The disclosures count code. The market is pricing something else — the people who will answer when the code, written faster than anyone can read it, does something no one asked it to do. The transition is not from engineers to machines. It is from a profes-

sion organized around production to one organized around accountability, and the layoffs and the compensation records are not contradictions but the same restructuring described from opposite ends of the org chart. The candidates on offer for the new apprenticeship — FDE rotations, internal AI academies, the residency programs the labs have begun to advertise — do not scale to the volume the old pipeline carried. The pipeline that fed the second category ran through the first. No one has yet explained what feeds it now.

COLOPHON

Set in DM Sans and IBM Plex Mono, with Source Serif 4 for display italics and pull quotes. Typeset on A4 at 10.5pt with weasyprint.

PUBLICATION

Rivoli AI · Essays · 2026-05-10

sam@rivoli.ai