Empty desks dissolve into autonomous agents, but one seat is re-forming. What comes back is the judgment the agent could not hold.

The AI Cost Bet That Is Quietly Hiring People Back

Syed Umair Shoaiby
Technical Program Manager at Cirrus Logic | Leading Semiconductor Silicon NPI from Tape-out to Production | AI...

June 3, 2026

Klarna replaced 700 people with AI. This year it started hiring them back.

The reason is not the one you would guess. The agents did not simply underperform. They turned out to be expensive in a way the business case never modeled, and the math is now reversing across the industry.

Start with the part most companies got backwards. AI looks cheaper every quarter. The blended cost of AI fell from $18.40 to $6.07 per million tokens in a single year, and Gartner expects inference on a large model to cost over 90% less by 2030. So leaders assumed agents would be cheap and cut staff against that assumption.

But the unit you buy is not the unit you spend. A chatbot answers once. An agent loops. It plans, calls a tool, reads the result, checks itself, retries, and then calls another tool. Gartner puts that at 5 to 30 times more tokens per task, and total consumption is forecast to grow 24 times by 2030. Per token, AI gets cheaper. Per task, it gets more expensive. Inference is now about 85% of the enterprise AI budget. The cost did not disappear when the headcount did. It moved into an operating line that scales with every loop.

Then there is the capability the demo oversold. Agents are genuinely good at the routine, high-volume layer. They fall apart on judgment, escalation, institutional knowledge, and the messy cross-functional calls that hold a

program together. Klarna's CEO admitted the company overestimated AI and underappreciated the human part of the work.

Klarna is not the outlier. It is the preview. 55% of companies that replaced workers with AI now regret it. Nearly a third have already reopened the exact roles they cut, and several found the cost of rehiring was higher than what they saved. Analysts expect half of the companies that blamed cuts on AI to be rehiring for the same functions by 2027.

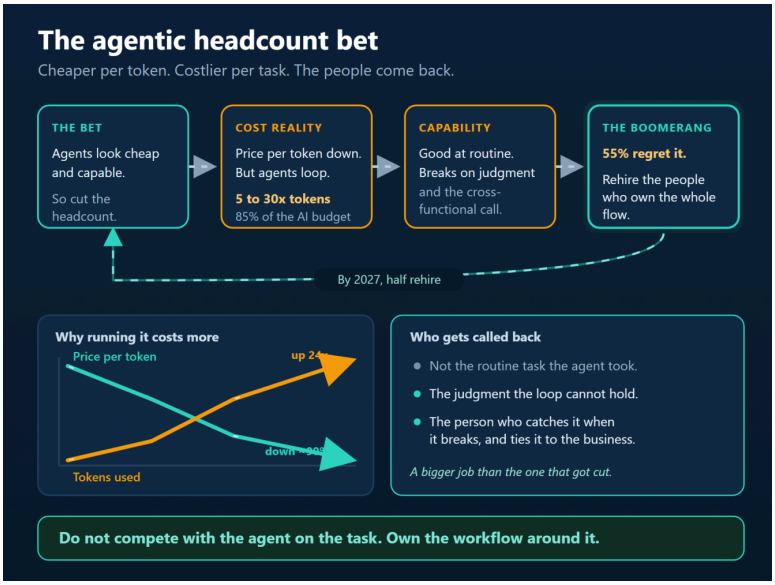
So the real question was never how many people to remove. It is which people to keep and what they need to become.

The people who get called back are not the ones doing the routine task the agent now handles. They are the ones who understand the entire flow. They can see where the agent quietly breaks, design it to take fewer steps instead of more expensive ones, own the judgment the loop cannot, and tie the whole thing back to what the business actually needs. That is a bigger job than the one that got cut, not a smaller one.

The lesson for anyone in this field is the same. Do not compete with the agent on the task. Learn the workflow around it well enough to own it. The cost of agentic work is rising, and its reliability is still catching up, and both of those facts pull in the same direction. They make the person who understands the system more valuable, not less.

If your team runs on agents now, do you actually know who understands the whole flow well enough to catch it when it breaks? And are you growing those people, or did you let them go?

#AI #AgenticAI #ProgramManagement #FutureOfWork #Leadership



The agentic headcount bet. Cheaper per token but costlier per task, and fragile on judgment, so the people who own the whole flow get called back.

REFERENCE GUIDE

"The blended cost of AI dropped from $18.40 to $6.07 per million tokens in a year"

AI Inference Cost Crisis 2026 (Oplexa). A 67% year-over-year drop in blended per-token cost, Q1 2025 to Q1 2026. Same source: inference is now about 85% of the enterprise AI budget.

https://oplexa.com/ai-inference-cost-crisis-2026/

"Inference on a large model will cost over 90% less by 2030, yet overall cost rises"

Gartner, press release, March 25, 2026. Per-token inference cost on a trillion-parameter model falls over 90% by 2030, but Gartner states plainly that lower unit cost drives disproportionately higher demand, so overall inference cost is expected to rise.

https://www.gartner.com/en/newsroom/press-releases/2026-03-25-gartner-predicts-that-by-2030-performing-inference-on-an-llm-with-1-trillion-parameters-will-cost-genai-providers-over-90-percent-less-than-in-2025

"Agents use 5 to 30 times more tokens per task"

Gartner, same release. Agentic models require 5 to 30x more tokens per task than a standard chatbot, because one task expands into a loop of reasoning steps, tool calls, and retries.

https://www.gartner.com/en/newsroom/press-releases/2026-03-25-gartner-predicts-that-by-2030-performing-inference-on-an-llm-with-1-trillion-parameters-will-cost-genai-providers-over-90-percent-less-than-in-2025

"Token consumption growing 24 times by 2030"

Goldman Sachs. Token consumption forecast to grow about 24x, to roughly 120 quadrillion tokens per month, between 2026 and 2030, driven by enterprise and consumer agent adoption.

https://www.goldmansachs.com/insights/articles/ai-agents-forecast-to-boost-tech-cash-flow-as-usage-soars

"Klarna said AI did the work of 700 agents, then started rehiring"

Reworked. Klarna publicly claimed AI replaced about 700 customer service agents, then reversed course and began rehiring as service quality on complex cases declined. CEO Sebastian Siemiatkowski admitted the company overestimated AI and underappreciated the human part of the work.

https://www.reworked.co/employee-experience/klarna-claimed-ai-was-doing-the-work-of-700-people-now-its-rehiring/

"55% of companies regret replacing workers with AI"

Orgvue and Forrester survey data, reported in the "AI boomerang" coverage. 55% of companies that replaced human workers with AI now regret it; hybrid human-AI models consistently outperform full automation.

https://hrexecutive.com/as-ai-layoff-regret-surges-will-boomerang-employees-make-a-comeback/

"Nearly a third reopened the exact roles, and rehiring often cost more than was saved"

The Interview Guys, summarizing Robert Half and Careerminds data. About 29% of companies that cut staff for AI have already reopened those exact positions; of 600 HR leaders, roughly two-thirds had already rehired some laid-off staff, and nearly a third found rehiring cost more than the original cuts saved.

https://blog.theinterviewguys.com/why-55-of-companies-regret-cutting-jobs-for-ai/

"Analysts expect half to rehire for the same functions by 2027"

AI boomerang analysis. By 2027, about 50% of companies that attributed headcount reductions to AI are expected to rehire staff for essentially the same functions, often under different titles.

https://blog.theinterviewguys.com/why-55-of-companies-regret-cutting-jobs-for-ai/

"AI-native gross margins near 52 percent against 75 to 85 percent for mature software"

AI Agent Economics, citing ICONIQ Capital 2026 data (TechTimes). A 23 to 33 point gross-margin gap driven almost entirely by inference spend. Supporting context for the rising operating cost of agentic work.

https://www.techtimes.com/articles/317542/20260601/ai-agent-economics-token-tax-locks-gross-margins-30-points-below-saas-baseline.htm

---

Disclaimer: The views shared here are my own and do not represent the positions of my employer or any organization I am affiliated with. Company examples, including Klarna, are drawn from public reporting and are used for illustration, not as commentary on any specific business. All figures come from the sources linked with this article and reflect the data available at the time of writing.

No comments, yet.

Be the first to comment.

Start the conversation

Syed Umair Shoaiby

Technical Program Manager at Cirrus Logic | Leading Semiconductor Silicon NPI from Tape-out to Production | AI Enabled Program Execution, Workflow Automation & Operational Strategy | Scaling Cross Functional Programs