BURAK GALİBA
thoughts on AI

ESSAY · ENTERPRISE AI

Rationing Tokens Is Not an AI Cost Strategy

Uber burned its 2026 AI budget in four months, then capped spend per engineer. The rationing cycle is predictable, and a uniform cap taxes exactly the people who found the value.

ROI AI STRATEGY LEADERSHIP

Burak Galiba · August 6, 2026 · 4 min read
https://burakgaliba.com/blog/rationing-tokens-is-not-a-cost-strategy

SCAN TO READ ONLINE

Uber rolled Claude Code out to its engineers in late 2025, told staff to use AI as much as possible, and ranked usage on internal leaderboards. By April its CTO was confirming to *The Information* that the entire annual AI budget was gone, four months in. In June, Bloomberg reported the response: a cap of $1,500 per employee per month on agentic coding tools, trackable on a dashboard, exceedable with permission.

Encouragement, then a meter, six months apart, from the same organisation — and neither decision was wrong on its own terms.

It isn't an Uber story. It's the default shape of what happens when an enterprise buys AI broadly without ever deciding what the capacity is for.

The cycle runs the same way every time

An organisation buys licences and hands them out widely — developers, analysts, employees, contractors, anyone who might benefit. Distribution is the whole plan. Almost nobody is taught how to get value out of the thing; people are pointed at the magic and left to find their own way in. Stack Overflow's 2025 survey of roughly 49,000 developers found the top frustration was output that is "almost right, but not quite", at 66%. Spotting almost-right is a learned skill, and nobody taught it.

Consumption spikes past anything anyone modelled, because consumption was never modelled. Finance reacts the only way it can to an unbudgeted line: it rations. Caps per person, or everyone downgraded to the cheap model.

The people who had learned to work with the tools are the ones who hit the ceiling, because they were the ones using them. Usage falls. Leadership reads falling usage as evidence the tools were oversold — and there is no budget line to defend, because AI never got one. The spend arrived as an overrun, so it gets governed like an overrun instead of managed like an investment. The organisation went looking for efficiency and found a cost centre.

The spike is not a forecasting failure, either. Falling unit prices don't lower AI bills — they raise them, by making workloads viable that were previously too expensive to attempt. Apollo's chief economist Torsten Slok named the mechanism in *Fortune* in June: token prices down more than 90% since 2023, aggregate AI spending doubled since late 2025 — Jevons paradox, the nineteenth-century observation that efficient coal engines increased coal consumption. The whole market was having this surprise in public. The unusual failure isn't the spike; it's having no plan for afterwards.

A uniform cap is inverted against value

Take a hypothetical, with round invented numbers to make the arithmetic visible. Five hundred people hold licences. Four hundred and fifty tried the tool, got a mediocre answer, and drifted back to what they knew; their consumption is near zero. Fifty restructured how they work and consume ten to twenty times the median.

Now set a cap at twice the median. It takes nothing from the 450 and removes most of the capability of the 50. On a spreadsheet the cap is admirably even-handed. In practice it is a tax levied exclusively on the people who found the value, at the moment they were the entire evidence base for the investment.

The reporting inverts too. Usage falls, which is what the cap was for — but the fall gets read as a verdict on the tool, and the people best placed to argue otherwise are the ones who were just throttled.

Two decisions, before the next renewal

Give AI a budget line before it needs one. Not a cap — a line, with an owner, forecast against a consumption model rather than a seat count. Seat pricing is already coming apart: Gartner's July forecast puts $234bn of enterprise application spend at risk of repricing toward consumption and outcome models by 2030. Overruns get cut. Investments get argued about on merit, and the second conversation is the one you want.

Allocate unevenly, on purpose. Uniform caps are the cheapest thing to administer and the most expensive thing to be wrong about. Fund the people producing demonstrable output at the level their work requires, stop paying for licences nobody opens, and instrument consumption against outcomes rather than headcount — the same baseline discipline that decides whether an AI investment survives year-end, applied to the cost side.

Both of those are cost management, and cost management is the smaller half of this. A budget line tells you what the capacity costs. It doesn't tell you what the capacity was for — and an organisation that can't answer the second question will keep having the first argument every year. That question has exactly two honest answers, and most organisations pick neither.

If you want a read on whether your organisation has the machinery to answer it, my free AI Readiness Score takes about ten minutes — 20 questions across pilots, data, talent and governance, including the ones this essay turns on: whether AI work has a named owner and a real budget, and whether anyone was trained to use what you bought.

Read the essay online at https://burakgaliba.com/blog/rationing-tokens-is-not-a-cost-strategy

I built a free AI Readiness Score — 20 questions across pilots, data, talent and governance, about ten minutes, no email required: https://burakgaliba.com/readiness