Tokenmaxxing: How Amazon's AI Leaderboard Became an Expensive Joke

Date: May 30, 2026 | Model: anthropic-batch:claude-opus-4-7

Source: Screenshot (OCR via AI)

Contents

Note: the original article is provided as a separate file (attached to the email or downloadable from the website).

1. Reading Passage

In late 2025, Amazon shut down an internal leaderboard called Kirorank that had been quietly tracking how much each of its employees used the company's in-house AI coding platform, Kiro. The dashboard was supposed to celebrate enthusiastic adopters of artificial intelligence. Instead, it became an expensive cautionary tale.

Amazon, a roughly $2.9 trillion company, has pushed its engineers hard toward AI. Internal targets require more than 80% of developers to use AI tools each week, and the company is expected to spend around $200bn on capital projects this year, the vast majority of it on AI infrastructure and data centres. In that environment, Kirorank looked like a natural extension of the strategy: rank engineers by their AI activity, post the scoreboard, and watch adoption climb.

It did climb — just not in a useful way. According to people familiar with the matter, workers began pointing autonomous AI 'agents' (bots that can take multi-step actions on a user's behalf) at unnecessary tasks, running repeated, low-value calls simply to inflate their consumption of 'tokens,' the units of text that AI models process and that cloud providers bill for. Inside Amazon, the practice picked up a name: tokenmaxxing. The result was a leaderboard full of activity that looked productive but wasn't, and a noticeably larger compute bill for Amazon itself. Dave Treadwell, an Amazon senior vice-president, told staff the leaderboard had been built with 'good intentions' but was being deprecated, and asked employees, plainly, not to use AI just for the sake of using AI.

Here's the catch that makes this more than an in-house embarrassment. The economics of corporate AI have shifted. Major model providers — Anthropic, whose systems Amazon uses heavily, among them — have moved away from flat monthly subscriptions toward consumption-based pricing, in which every token processed has a price. Under the old flat-fee world, wasted tokens were essentially free. Under the new one, every gamed leaderboard score is a small but real transfer of money from Amazon to its AI suppliers. Multiply that across a workforce of hundreds of thousands of engineers and the numbers stop being small.

Amazon is not alone. The Financial Times has reported that Meta employees have engaged in similar gaming on their own internal tables, suggesting tokenmaxxing is less an Amazon-specific glitch than a Big Tech pattern. The underlying dynamic is familiar to anyone who has studied incentives: when a measure becomes a target, people optimise for the target, not for the thing the measure was meant to track. A score meant to capture 'how seriously do you take AI?' instead captured 'how willing are you to waste compute?' — and the two answers turned out to be different.

The industry's response is already taking shape. Amazon has begun emphasising a metric it calls 'normalised deployments,' which tries to measure whether AI-assisted code actually shipped and produced value, rather than how much was generated. Expect more of that kind of pivot from rivals as investors start asking harder questions about whether the hundreds of billions flowing into AI data centres are producing real productivity gains — or just very expensive activity charts.

2. Explanation

> ***Amazon built a scoreboard to celebrate employees who used its AI tools the most. Workers promptly figured out that the fastest way to win was to waste the company's money.***

What's Going On?

Amazon has quietly killed an internal leaderboard called Kirorank that ranked employees by how much they used the company's in-house AI coding tool. The tracker lived inside Amazon's Kiro developer platform, and its scores were supposed to celebrate engineers who were eagerly adopting AI in their workflows.

Instead, workers started gaming the system. Some pointed autonomous AI 'agents' at pointless busywork just to rack up activity, a practice Amazon insiders nicknamed 'tokenmaxxing.' Senior vice-president Dave Treadwell told staff the leaderboard had been built with 'good intentions' but was driving up the company's compute bill, so it was being deprecated. Meta engineers, the Financial Times reported, have been playing the same game on their own internal dashboards.

How To Think About It

This is a textbook case of Goodhart's Law: the moment you turn a measurement into a target, people stop optimising for the underlying thing you actually wanted and start optimising for the number itself. The leaderboard didn't measure good engineering — it measured tokens consumed.

- Imagine a teacher who grades essays by word count. Students don't write better essays; they write longer, padded ones — and the teacher still has to read them all.
- Or think of Soviet nail factories famously judged on tonnage: they made giant, useless nails. Judged on quantity, they made millions of tiny ones nobody could hammer. The metric, not the customer, ran the factory.

Key Things To Know

- Amazon is a roughly $2.9 trillion company and plans to spend about $200bn in capital expenditure this year, the vast majority of it on AI and data centres.
- A 'token' is a chunk of text an AI model processes — every token costs real money in GPU time, so inflated usage hits the bottom line directly.
- Amazon has pushed targets requiring more than 80% of its developers to use AI tools each week, which created the pressure to perform usage in the first place.
- AI providers like Anthropic (whose models Amazon uses heavily) have shifted from flat monthly fees to consumption-based pricing, meaning every wasted token is now billed.
- The popular misread: this isn't lazy employees sabotaging Amazon. It's rational employees responding exactly the way the incentive system told them to.

Why It Matters

If you're heading into a STEM or business career, you'll spend a lot of time being measured — GPA, KPIs, OKRs, GitHub commits, LinkedIn posts. The Kirorank fiasco is a live demo of why smart people game

dumb metrics, and why your future managers will obsess over 'productivity' numbers that may say almost nothing about whether real work is getting done. It also tells you something important about the AI hype cycle: companies are under so much pressure to look AI-native that they sometimes confuse using AI with creating value.

The Bigger Picture

Watch what happens to AI economics next. Now that providers charge by the token and hyperscalers are sinking hundreds of billions into data centres, every 'tokenmaxxing' spree shows up as wasted capex. Expect a swing in the opposite direction — leaderboards replaced by 'normalised deployments' and quality metrics, plus tighter scrutiny of whether AI adoption is actually producing useful code or just expensive theatre. The second-order effect: investors will start demanding evidence that AI spending translates to revenue, not just usage charts.

3. Key Terms Glossary

Token

The basic unit of text an AI language model reads or writes — roughly a short word or piece of a word. Cloud AI services bill customers per token processed.

Tokenmaxxing

Insider slang for artificially inflating your AI-token usage to look productive on an internal dashboard, even when the underlying work is pointless.

AI agent

An autonomous AI program that can take multi-step actions on a user's behalf — sending emails, running code, calling other tools — rather than just answering one prompt.

Deprecated

Tech-industry term for officially retiring a tool or feature. It still might exist briefly, but it's no longer supported or recommended.

Capital expenditure (capex)

Money a company spends on long-lived physical assets like buildings, servers, or data centres — as opposed to day-to-day operating costs.

Consumption-based pricing

A billing model where you pay per unit used (per token, per API call) rather than a flat monthly fee. It rewards efficiency and punishes waste.

Goodhart's Law

The principle that 'when a measure becomes a target, it ceases to be a good measure' — because people start gaming the metric instead of pursuing what it was meant to track.

Normalised deployment

Amazon's preferred metric: a measure of AI tool use weighted by whether the code actually shipped and produced value, rather than just counting raw activity.

4. Reading Comprehension Quiz

Circle the best answer for each question.

Q1. The passage most directly argues that Amazon's Kirorank leaderboard failed because:

- ◯ **A)** Amazon's AI tools were technically inferior to those of Microsoft and Google.
- ◯ **B)** Employees responded to the incentive by inflating usage rather than improving work.
- ◯ **C)** Senior management never communicated the leaderboard's purpose to staff.
- ◯ **D)** Anthropic's pricing model made the underlying AI tools too expensive.

Q2. Which choice best states the central idea of the passage?

- ◯ **A)** Internal metrics meant to encourage AI adoption can backfire by rewarding wasteful behaviour.
- ◯ **B)** Amazon's $200bn capital spending plan is unsustainable in the current AI environment.
- ◯ **C)** Autonomous AI agents are unreliable and should not be deployed inside large companies.
- ◯ **D)** Meta has overtaken Amazon as the leading employer of AI engineering talent.

Q3. According to the passage, Amazon's compute costs rose because:

- ◯ **A)** Anthropic raised its per-token prices without warning Amazon's procurement team.
- ◯ **B)** Employees deliberately ran unneeded AI tasks to climb the Kirorank rankings.
- ◯ **C)** Amazon's data-centre construction fell behind its rising demand for AI capacity.
- ◯ **D)** Engineers refused to adopt AI tools despite leadership's 80% usage mandate.

Q4. As used in the passage, the word 'inflating' most nearly means:

- ◯ **A)** filling with gas
- ◯ **B)** exaggerating in size
- ◯ **C)** artificially increasing
- ◯ **D)** raising prices on

Q5. As used in the passage, the word 'deprecated' most nearly means:

- ◯ **A)** publicly criticised
- ◯ **B)** formally retired
- ◯ **C)** morally disapproved
- ◯ **D)** discounted in price

Q6. Which statement about consumption-based pricing can most reasonably be inferred from the passage?

- ◯ **A)** It eliminates the financial risk of internal AI experiments.
- ◯ **B)** It makes tokenmaxxing more costly than it would be under flat-fee pricing.
- ◯ **C)** It is unique to Anthropic among major AI providers.
- ◯ **D)** It will be banned by regulators in the coming year.

Q7. The passage suggests that the behaviour of Meta's employees:

- ◯ **A)** is fundamentally different from what occurred at Amazon.
- ◯ **B)** has been more financially damaging than Amazon's tokenmaxxing.
- ◯ **C)** follows the same pattern of gaming internal usage metrics.
- ◯ **D)** was caused by direct pressure from Anthropic's sales team.

Q8. The author's tone in describing Kirorank is best characterised as:

- ◯ **A)** wry and analytical
- ◯ **B)** outraged and alarmed
- ◯ **C)** celebratory and admiring
- ◯ **D)** neutral and statistical

Q9. Which statement about AI adoption inside large tech companies can most reasonably be inferred from the passage?

- ◯ **A)** Mandating high AI usage rates can produce the appearance of adoption without its substance.
- ◯ **B)** Most engineers at Amazon and Meta secretly oppose using AI in their work.
- ◯ **C)** AI agents are technically incapable of completing meaningful engineering tasks.
- ◯ **D)** Tech companies will soon abandon all internal AI tracking entirely.

Q10. Which detail from the passage best supports the answer to the previous question?

- ◯ **A)** Amazon plans roughly $200bn in capital expenditure this year, mostly on AI and data centres.
- ◯ **B)** Amazon set targets requiring more than 80% of its developers to use AI tools each week.
- ◯ **C)** Senior vice-president Dave Treadwell said the leaderboard had 'good intentions.'
- ◯ **D)** Anthropic recently shifted from flat monthly fees to consumption-based pricing.

My Score: _____ / 10

5. Answer Key with Explanations

Q1. The passage most directly argues that Amazon's Kirorank leaderboard failed because:

Answer: B

The passage's central claim is that workers gamed the metric by 'tokenmaxxing,' raising costs without raising real output. D names a real factor mentioned in the passage but it's a background condition, not the cause of the leaderboard's failure — Trap C (true-ish but not what the passage argues). SAT Tip: 'Most directly argues' questions want the thesis, not a supporting detail; pick the option that captures the cause-and-effect spine of the whole passage.

Q2. Which choice best states the central idea of the passage?

Answer: A

The passage uses Kirorank as evidence for the broader idea that adoption-pressure metrics can produce perverse outcomes. B touches a passage fact but inflates it into a claim the author never makes — Trap C (true-sounding but unsupported). SAT Tip: A 'central idea' answer should fit the entire passage, not just the most memorable paragraph.

Q3. According to the passage, Amazon's compute costs rose because:

Answer: B

The passage explicitly says workers used AI agents for needless tasks to climb the rankings, which drove up token consumption and infrastructure costs. D inverts the actual behaviour described — Trap A (right scope, wrong direction). SAT Tip: For 'according to the passage' questions, your answer must be literally stated, not just plausible — beware options that reverse the cause and effect.

Q4. As used in the passage, the word 'inflating' most nearly means:

Answer: C

In context, employees were artificially boosting their token consumption — pumping the number up without real cause. A is the literal common meaning (Trap B: passage vocab in the wrong sense). SAT Tip: On vocab-in-context, substitute each option back into the sentence — the right answer keeps the original meaning intact; the common-definition trap usually doesn't.

Q5. As used in the passage, the word 'deprecated' most nearly means:

Answer: B

The passage uses 'deprecated' in its software-industry sense — the dashboard has been shut down and is no longer supported. A and C reflect the everyday English meaning of 'deprecate' (to express disapproval) — Trap B (familiar word in an unfamiliar technical sense). SAT Tip: When a word has both a technical and an everyday meaning, the SAT almost always wants the meaning that matches the passage's domain.

Q6. Which statement about consumption-based pricing can most reasonably be inferred from the passage?

Answer: B

If providers like Anthropic charge per token rather than flat monthly fees, every wasted token now generates a real bill — making the gaming behaviour directly expensive. D is dramatic but unsupported anywhere in the passage — Trap C (plausible-sounding but invented). SAT Tip: Inference answers should be a small logical step from the text, not a leap; reject options that introduce brand-new claims.

Q7. The passage suggests that the behaviour of Meta's employees:

Answer: C

The passage groups Meta's engineers with Amazon's as examples of the same gaming pattern across Big Tech. B introduces a comparison the passage never makes — Trap C (extra-textual claim). SAT Tip: 'The passage suggests' is still a textual question — if a comparison or ranking isn't supported, don't supply it from your imagination.

Q8. The author's tone in describing Kirorank is best characterised as:

Answer: A

The author treats the episode as a darkly funny illustration of perverse incentives, mixing analysis with light irony (e.g. 'expensive theatre,' 'gaming dumb metrics'). D ignores the clear authorial attitude that runs through the piece — Trap B (using a passage word, 'data,' to suggest a clinical tone the text doesn't actually have). SAT Tip: For tone questions, scan for the author's adjectives and word choices, not the seriousness of the topic — a serious subject can still be discussed wryly.

Q9. Which statement about AI adoption inside large tech companies can most reasonably be inferred from the passage?

Answer: A

The Kirorank story shows that aggressive usage mandates can create incentives to perform AI use rather than benefit from it. D overshoots into a sweeping prediction the passage doesn't make — Trap C (absolute claim unsupported by text). SAT Tip: Among inference options, distrust absolutes like 'all,' 'never,' or 'entirely' — the correct answer is usually the more measured one.

Q10. Which detail from the passage best supports the answer to the previous question?

Answer: B

*The 80% mandate is the clearest example of an adoption target that pressures employees into showing usage regardless of value — exactly the dynamic Q9 identifies. A is a real passage fact but speaks to spending scale, not to the adoption-vs-substance gap — Trap C (true detail, wrong job). SAT Tip: On evidence-pairing questions, lock in your Q9 answer first, then hunt for the line that proves *that specific claim* — don't pick a quote just because it sounds important.*