

Stop Counting Your Tokens

Why "more AI usage" is the wrong scoreboard, and what to measure instead

Contact

**Fraunhofer Institute for Software
and Systems Engineering**

Speicherstraße 6
44147 Dortmund
Germany

info@isst.fraunhofer.de
www.isst.fraunhofer.de/en

Authors

Markus Spiekermann
Head of Division "Manufacturing Ecosystems"
markus.spiekermann@isst.fraunhofer.de

Prof. Dr.-Ing. Boris Otto
Executive Director
boris.otto@isst.fraunhofer.de

Introduction

In April, Meta used its internal leaderboard to rank about 85,000 employees by the number of AI tokens they consumed. Over 30 days, the dashboard recorded more than 60 trillion tokens, and the highest-ranked employee averaged 281 billion. The leaderboard was taken down two days after the media had reported on it (Munis J., 2026). By the end of May, Fortune had already declared 'tokenmaxxing' dead because it hadn't delivered the return companies expected (Kahn, J., 2026).

The same pattern is observed at the market level. According to Gartner's most recent forecast, global AI spending will reach \$2.7 trillion in 2026, representing a year-over-year increase of 49.5% (Gartner, 2026c). A large share of this spending is being directed to infrastructure purchased by vendors and hyperscalers, rather than by enterprises. In an earlier Gartner release, it was stated that "enterprises have yet to really flex their spending potential" and that 2026 is the inflection year when this will change (Gartner, 2026b). Google claims that it now processes more than 3.2 quadrillion tokens per month across its own products, a figure up sevenfold from the previous year (Pichai, S., 2026).

Gartner's view of the position that this leaves enterprises in is direct: AI will still be in the "Trough of Disillusionment" in 2026, and "before enterprises can genuinely scale up AI, the improved predictability of ROI must take place" (Gartner, 2026a). The decision to adopt it was the simple part; what comes after that is having to demonstrate that the money spent was worthwhile, a criterion which is different from the one that most companies have been monitoring.

The Call-Volume Problem

Imagine a sales team whose target is calls per week. Everyone can meet it, and it says nothing about revenue. Token counts work the same way: they're easy to measure and fully automated, which is exactly why they became the default proxy for whether a company is "*doing AI*."

The Financial Times reported that some Amazon employees created agents to perform meaningless tasks to keep their usage figures high. A 2026 study by OpenAI's own research team, which linked ChatGPT usage data with the financial results of more than 1,500 publicly traded companies, found that revenue per employee had only a weak, imprecisely estimated positive relationship with an employee's token usage. Research on coding agents tells the same story from a different perspective: when researchers tested eight frontier models on SWE-bench Verified, identical tasks produced up to 30-fold differences in token usage across runs, and accuracy generally peaked at a moderate cost level before leveling off (Chatterji, A. et al., 2026). Beyond that point, additional tokens brought no benefit.

Why the Bill Keeps Climbing

One reason is the system's architecture. Unlike chat conversation, agents make multiple calls by resending the context they have accumulated at each stage. When Anthropic analyzed its research agent system, it found that agents use about four times as many tokens as a single chat exchange, and multi-agent setups use about fifteen times as many (Anthropic, 2025). A 2026 study on coding agents found an even larger gap: agentic coding tasks used about 1,000 times more tokens than an equivalent code chat, with most of the cost coming from input tokens rather than the model's output (Bai, L., 2026).

Price declines haven't offset this, and Gartner doesn't expect them to. While provider costs per token are dropping rapidly, Gartner predicts that by 2030 running inference on a trillion-parameter model will cost providers more than 90% less than it did in 2025. The same report makes it clear that this won't result in lower bills. "Since token consumption is rising faster than token costs are falling, overall inference costs are expected to go up," says Will Sommer, Senior Director Analyst at Gartner (Gartner, 2026d). Goldman Sachs models the same trend on a larger scale, forecasting that token consumption by consumers and enterprises will increase 24 times between 2026 and 2030, reaching 120 quadrillion tokens per month (Goldman Sachs, 2026). Some companies are already feeling the effects: Uber's leadership has said the company has used up its entire annual AI budget in about four months (Angelo, J., 2026).

How Many Tokens Is "Normal"?

No authoritative benchmark specifies token usage per employee by company, and no analyst report breaks down enterprise usage this way. The closest answer comes from the same OpenAI paper cited above, which found something more interesting than a simple look-up table: usage does not increase with the number of employees; instead, it decreases (Chatterji, A. et al., 2026). Among companies that do adopt, output tokens per employee drop as employee count rises. The authors describe this as "a mechanical or organizational scaling effect: larger firms are more likely to adopt, but conditional on having adopted, the measured use per employee is lower." Scaling enables you to roll out AI sooner than you increase AI use per employee.

Different companies estimate the amount involved, and the figures vary considerably. G2, a software-review company with about 300 employees, spent \$1.27 million on AI tokens in 2026 (about 970 billion tokens), with the share of its spending on Claude Code and Cowork rising from 17% in January to 79% by July (Keary, T., 2026). At Zapier, non-technical staff spend a few hundred dollars a month on AI; most developers spend thousands; the heaviest users can

Token

The basic unit of text a language model processes (roughly three quarters of an English word). AI usage and pricing are measured in tokens.

Tokenmaxxing

Treating maximum token consumption as proof of AI adoption, for example by ranking employees on internal usage leaderboards.

reach \$30,000 a month, mainly on greenfield projects with clean feedback loops (ibid.). As for the Meta leaderboard mentioned at the start of this post, it is not a measure of typical usage; it shows what the top end of the distribution looks like when usage becomes a matter of status.

The only rule that can be regarded as genuine rather than merely anecdotal is Gartner's: agentic workloads need between 5 and 30 times as many tokens per task as a typical chatbot interaction, no matter what the size of the company (Gartner, 2026d).

Token Minimizing: The Same Trap

In response to the visible figure, teams capped older models, those with smaller context windows, and those using shorter prompts. The initial round of cuts, which removed the oversized tool catalogs and the out-of-date context that was no longer needed, represented real achievements. Subsequent cuts eliminated the task descriptions, the domain constraints, and the architectural context; as a result, the agents that lost these features produced work that broke the wider system.

The situation is imbalanced: saving 500 input tokens while spending 5,000 on rework amounts to almost nothing in raw dollars; what actually counts as a cost is the human time spent debugging. The total cost hasn't changed; it's just been shifted to a place the dashboard doesn't show.

Blunt caps also fail to account for real factors. Angel Lange, CFO of the contract-management platform Agiloft, found that 74% of users were not reaching their usage limits, showing that broad restrictions did not address the factors that drive spending (Keary, T., 2026). Three things matter more than setting a ceiling: preventing paying twice for the same context, matching the model's capabilities to the task's difficulty, and asking whether a given step actually needs a model.

The first issue concerns the problem of re-reading. At each agent step, all the accumulated context, the system prompt, the conversation history, and the reference documents are resent. By default, the model reprocesses all of this from the beginning, even if it is identical to the previous step. Prompt caching allows the provider to avoid this reprocessing, and the savings are considerable: both Anthropic¹ and OpenAI² charge only about one tenth of the normal input rate for a cache hit, which represents a 90% reduction for the reused material. This situation is not particular to either company; any workflow that repeats context across calls incurs a re-read charge that it does not need to pay, and the way to avoid it is to construct the prompts so that the unchanging part comes first and the changing part comes last.

The second issue concerns matching capabilities. It's not necessary to use the most capable and most expensive model for every task. Just as you would assign simple tasks with clear specifications to smaller models and reserve the more advanced ones for truly difficult reasoning, you should also select the appropriate instance size when using cloud computing: the task determines the capacity needed, not the other way around. The prices published by OpenAI show the range of options available, with a 25-fold difference between its cheapest and most capable text model, according to the same measure. Although that particular ratio applies to a single provider at a specific time, the idea of making model selection depends on each individual task goes well beyond that situation, summarized under the term "model routing" (Chen, L. et al., 2023; Ding, D. et al., 2024).

Three Cost Levers

- 1 Reuse context
- 2 Match model to task
- 3 Skip models if possible

¹ <https://platform.claude.com/docs/en/about-claude/pricing>, accessed 09/28/2026

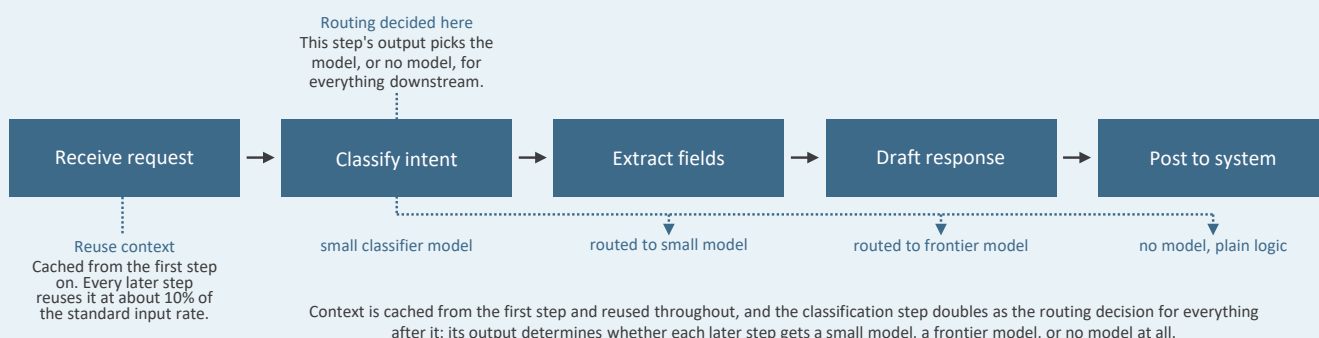
² <https://developers.openai.com/api/docs/pricing>, accessed 09/28/2026

The third case is even more straightforward: does this stage actually need a model? Many workflow steps involve lookups, formatting, or other fixed actions and are passed to a model simply for convenience rather than because they are essential. In its analysis of 24 workflows, Zapier made a distinction between these cases and discovered that by keeping AI to the steps which truly require reasoning and handling all the rest using ordinary logic, it was able to reduce model-based workflow costs by 71% when compared to sending all the steps through a model (Zapier AI Workflow Index, Q2 2026). Zapier notes that this cost model is based on assumed pricing, not actual customer spending. Nevertheless, the pattern identifies that most workflows are mostly deterministic and that AI only carries out a smaller, focused portion of the work.

Figure 1: The combination of caching, model routing, and selective use of AI throughout a workflow.

Where the Three Cost Levers Apply in a Real Workflow

Example: an automated support-ticket workflow



Value Maximization: Measure Outputs

The solution is to reformulate the question, not just provide a different answer. Rather than asking how many tokens a team used, we should ask what resulted from that usage: how many deployments were released, how much developer time was saved, how much rework was avoided, and how many vulnerabilities were resolved? All these things are entirely ordinary; they are the same kinds of delivery metrics that engineering teams have already been tracking, but now they are being applied to AI spending instead of being focused on it.

That is a completely different question from "are we using enough AI?" A team that used more tokens this month is no clearer indication of better work; it is a team that shipped more, reworked less, and closed more vulnerabilities. When you look at outcomes like these, continuous token spending ceases to be something to minimize and instead becomes a stake to be priced.

Several companies have already put this into practice. G2 has an internal 'AI Efficiency Index' which rates its employees according to the number of tokens obtained per dollar spent rather than simply by the number of tokens spent, with the aim of rewarding efficient use rather than extensive use (Forbes, 2026). Angel Lange sums up the change in a single statement: "The question isn't how many tokens did we use? It's what did those tokens enable?" The company claims a 67% increase in developer productivity from using AI and monitors this alongside its token expenditure rather than in place of it (ibid.).

Gartner also makes the same point at the market level. In its July 2026 report on the market for AI models and platforms, which forecasts sales at \$64 billion in 2026, which is a CAGR of 63%, it states that enterprise budgets are coming under "increased focus on usage efficiency, cost control and measurable outcomes", with money being directed towards providers who can show value in terms of cost, latency, performance, and reliability (Gartner, 2026e). Both individual teams and the market as a whole are reaching the same conclusion: the figure worth watching is not how much AI is being used, but what it produces.

When Models Become Infrastructure, the System Becomes the Product

When outcomes matter more than the number of tokens, systems matter more than tokens. Once a capable model is available to all, access to it ceases to be an advantage. The kind of system a team builds around the model, such as context management, workflow orchestration, governance, and optimization, distinguishes two teams using the same model. This is also why model orchestration is becoming more important than model selection: IDC predicts that by 2028, 70% of leading AI-driven enterprises will operate advanced multi-tool architectures that dynamically route tasks among multiple models rather than sticking with a single model (Ward-Dutton, N., 2025). Choosing "the best model" is a one-off decision; however, building the system that routes, caches, and monitors around it is a continuous effort, and that's where real differentiation now exists.

Developers already treat cloud spend, database load, and application performance as things to actively manage. Token use deserves the same discipline and the FinOps community, which built that discipline for cloud in the first place, now says so explicitly. J.R. Storment, Executive Director at FinOps Foundation (2026) states "Tokenomics is best understood as FinOps applied to AI," treating cost-per-token as a unit metric the same way cloud teams already track vCPU-hours (FinOps Foundation, 2026). The Linux Foundation made it official in June 2026, launching a dedicated Tokenomics Foundation to standardize AI cost practices across enterprises and model providers (Tokenomics Foundation, 2026).

Rigor starts with visibility, since management can't happen without collected figures. According to G2, 80% of software buyers now allocate a specific token or LLM usage budget to their technical teams, treating AI costs like any other expense (Forbes, 2026). Jensen Huang, CEO of Nvidia has said he can see every engineer at the company eventually working within an annual token budget of their own (Huang, J. 2026). While this approach is not opposed to adoption of AI, a budget that does not have an outcome metric linked to it still results in the same "tokenmaxxing" problem, only in this case the ceiling is fixed rather than open.

In simple terms, if a company wants to move from mere adoption to real value, it must monitor token usage by team, workflow, and outcome, rather than relying on its monthly invoice as its only source of visibility. Since adoption didn't require much instrumentation, the only question was whether people were using AI at all. The question is now different and answering it demands data that most companies still fail to collect. Only by tracking can this question be answered, and right now, tracking is the step most companies still skip.

FinOps

The discipline of actively managing cloud spending, now increasingly applied to AI as well.

Tokenomics

FinOps applied to AI. Cost per token is tracked as a unit metric, just as cloud teams track vCPU-hours.

References

- Angelo, J. (2026, May 26). *Uber burned through its entire 2026 AI budget in four months. Now its COO is questioning whether it's worth it.* Fortune. <https://fortune.com/2026/05/26/uber-coo-ai-spending-tokens-claude-code/>
- Anthropic. (2025, June 13). *How we built our multi-agent research system.* <https://www.anthropic.com/engineering/multi-agent-research-system>
- Bai, L., Huang, Z., Wang, X., Sun, J., Mihalcea, R., Brynjolfsson, E., Pentland, A., & Pei, J. (2026, April 24). *How do AI agents spend your money? Analyzing and predicting token consumption in agentic coding tasks.* arXiv. <https://arxiv.org/abs/2604.22750>
- Chatterji, A., Holtz, D., Rakholia, N., Tambe, P., & Weeratunga, G. (2026, August 12). *How organizations use AI: Evidence from ChatGPT.* arXiv. <https://arxiv.org/abs/2608.12236>
- Chen, L., Zaharia, M., & Zou, J. (2023, May 9). *FrugalGPT: How to use large language models while reducing cost and improving performance.* arXiv. <https://arxiv.org/abs/2305.05176>
- Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Ruhle, V., Lakshmanan, L. V. S., & Awadallah, A. H. (2024, April 22). *Hybrid LLM: Cost-efficient and quality-aware query routing.* arXiv. <https://arxiv.org/abs/2404.14618>
- Gartner. (2026a, July 20). *Gartner forecasts worldwide AI platforms and models market to grow 63% in 2026* [Press release]. <https://www.gartner.com/en/newsroom/press-releases/2026-07-20-gartner-forecasts-worldwide-ai-platforms-and-models-market-to-grow-63-percent-in-2026>
- Gartner. (2026b, May 19). *Gartner forecasts worldwide AI spending to grow 47% in 2026* [Press release]. <https://www.gartner.com/en/newsroom/press-releases/2026-05-19-gartner-forecasts-worldwide-ai-spending-to-grow-47-percent-in-2026>
- Gartner. (2026c, September 16). *Gartner forecasts worldwide AI spending to grow 49.5% in 2026* [Press release]. <https://www.gartner.com/en/newsroom/press-releases/2026-09-16-gartner-forecasts-worldwide-ai-spending-to-grow-49-point-5-percent-in-2026>
- Gartner. (2026d, March 25). *Gartner predicts that by 2030, performing inference on an LLM with 1 trillion parameters will cost GenAI providers over 90% less than in 2025* [Press release]. <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>
- Gartner. (2026e, January 15). *Gartner says worldwide AI spending will total \$2.5 trillion in 2026* [Press release]. <https://www.gartner.com/en/newsroom/press-releases/2026-1-15-gartner-says-worldwide-ai-spending-will-total-2-point-5-trillion-dollars-in-2026>
- Goldman Sachs. (2026, May 20). *AI agents forecast to boost tech cash flow as usage soars.* <https://www.goldmansachs.com/insights/articles/ai-agents-forecast-to-boost-tech-cash-flow-as-usage-soars>
- Google. (2026, May). *Google I/O 2026: Sundar Pichai's opening keynote.* The Keyword. <https://blog.google/innovation-and-ai/sundar-pichai-io-2026/>
- Kahn, J. (2026, May 28). *Tokenmaxxing is over. That's because it never measured what really counts to see ROI from AI.* Fortune. <https://fortune.com/2026/05/28/tokenmaxxing-is-dead-companies-didnt-get-the-roi-from-ai-they-wanted-to-see/>
- Keary, T. (2026, August 18). *AI token spend is rising, but measuring value remains challenging.* Forbes. <https://www.forbes.com/sites/timkeary/2026/08/18/ai-token-spend-is-rising-but-measuring-value-remains-challenging/>
- Munis, J. (2026, April 9). *A Meta employee created a dashboard so coworkers can compete to be the company's No. 1 AI token user—and Zuckerberg doesn't even rank in the top 250.* Fortune. <https://fortune.com/2026/04/09/meta-killed-employee-ai-token-dashboard/>
- Storment, J. R. (2026, May 10). *Token economics: The atomic unit of AI value.* FinOps Foundation. <https://www.finops.org/insights/token-economics-the-atomic-unit-of-ai-value/>
- Tokenomics Foundation. (2026). *The Tokenomics Foundation: AI value.* Retrieved September 28, 2026, from <https://www.tokeneconomics.com/>
- Ward-Dutton, N. (2025, December 11). *The future of AI is model routing.* IDC Blog. <https://www.idc.com/resource-center/blog/the-future-of-ai-is-model-routing/>
- Zapier. (2026). *How enterprises put AI to work: The AI workflow index (Q2 2026 report).* <https://zapier.com/ai/workflow-index/q2-2026-report>