NeuraSec

AI FinOps Wake-Up Call

Helping CIOs and CFOs be ready

The $750 Wake-Up Call: Why You Need an AI FinOps Strategy Right Now

By Gary Willock · NeuraSec · June 2026

The bill that landed this morning

At midnight last night, GitHub flipped a switch.

GitHub Copilot - the AI coding assistant used by an estimated 20 million developers worldwide - moved from a flat $10/$19/$39 monthly subscription to **token-based billing**. Same tool. Same vendor. Same login. Completely different bill at the end of the month.

The transition was announced on 27 April by GitHub's Chief Product Officer. The mechanics are simple: instead of a fixed fee, you now get a monthly allotment of "GitHub AI Credits" (1 credit = $0.01) which are consumed by every input token, output token, and cached token you push through the service. When the credits run out, you either top up or you stop.

By breakfast time on day one, developers were posting screenshots online of projected bills moving from **$29/month to $750/month**. One team estimated their seat cost would jump from **$50 to over $3,000**. For users of the more advanced "agentic" features - the ones where Copilot writes whole files, fixes whole pull requests, runs code review autonomously - increases of **10x to 50x** are now routine.

You can call it bait-and-switch. You can call it a market correcting itself after eighteen months of vendors burning capital to win mindshare. Both are partly true. What is undeniably true is this:

The era of flat-rate AI is ending. And most organisations are nowhere near ready for what comes next.

This is not a Copilot story. It's a category story.

If GitHub were an outlier, you could shrug, switch tool, and move on. They are not. This is the curve every major AI product is moving along:

- **Cursor**, the developer-darling IDE, already runs a credit-pack model with rapidly diminishing "fast request" allowances
- **Anthropic's Claude Code** charges per-token from the first keystroke
- **OpenAI's ChatGPT Enterprise** is layering usage caps and "deep research" credit limits on top of seat licences
- **Microsoft 365 Copilot** has quietly added "metered consumption" alongside the per-user fee for agentic and Copilot Studio scenarios
- **Google's Gemini for Workspace**, **AWS Q**, **Salesforce Einstein** - all moving in the same direction

The economics make this inevitable. Frontier models are expensive to run. A "request" is no longer a request - a single agentic task can consume millions of tokens of context, retry on failure, call sub-agents, and read entire codebases. Vendors who priced on requests in 2024 were quietly subsidising heavy users with light ones. That subsidy is gone.

What that means for you, as a buyer:

In 2025, your AI cost was a line on the IT budget. In 2026, it is a line in cost-of-goods-sold, and nobody in your finance team knows how to forecast it.

That is the problem.

Cloud déjà vu - but worse

We have done this before. Badly.

When the major enterprises moved to public cloud between 2013 and 2020, the pitch was straightforward: pay only for what you use, scale elastically, retire your datacentre. The reality was that finance teams kept treating Azure and AWS like the old infrastructure budget - a fixed annual capex figure, signed off once, reviewed at year-end.

The result was predictable. Run rates ballooned. Test environments left on overnight. Storage tiers untouched. Reserved Instances unbought. Tags missing. By 2019 the average enterprise was wasting between **30% and 35%** of its public cloud spend, according to every reputable benchmark - Flexera, Gartner, FinOps Foundation, take your pick.

The discipline that emerged to fix this - **FinOps** - is now well established. There are certifications, frameworks, dedicated tooling (Apptio Cloudability, IBM Turbonomic, Microsoft Cost Management, native CSP advisors), and a body of practice around showback, chargeback, unit economics, and engineering accountability.

Here's the uncomfortable truth: **most UK organisations still haven't actually implemented it.** Cloud FinOps remains, for many councils, NHS trusts, housing associations, and mid-market enterprises, an aspiration rather than a discipline. The CFO still sees one large bill. The architects still spin up what they need. Nobody owns the unit cost of a transaction.

Now imagine layering on top of that:

- **Five to fifteen different AI vendors**, each with their own metering model
- **Token-based billing** that varies by model (a query to GPT-5 costs 8x what the same query costs against a smaller model)
- **Cached vs uncached pricing** that rewards certain prompt patterns and punishes others
- **Agentic workflows** that consume tokens for tasks the user never sees
- **Per-user, per-team, per-department, per-customer** consumption patterns that change weekly
- **A regulatory environment** (EU AI Act, ICO guidance, sector-specific rules) that may require you to **justify** not just what AI you use, but how much, and for what

This is cloud 2.0. On steroids. And the organisations that failed to grip cloud FinOps in the 2010s are about to do the same thing with AI - at a faster cadence, with less visibility, and with every department procuring independently.

What changes when AI is metered

Three things break the moment your AI moves from flat-rate to consumption-based.

1. Your budget model

A predictable £39/user/month line item becomes a variable cost that can swing by an order of magnitude based on which features people use and how they use them. The same engineer doing the same job can cost £30 one month and £600 the next, with no change in headcount or contract. Annual budgets become meaningless without forecasting telemetry.

2. Your vendor strategy

When the cost of intelligence is variable, the question is no longer "which AI vendor do we standardise on?" - it's "which AI do we use **for which workload**, at what cost, with what fallback?" Some workloads will stay on flat-rate licences where they exist. Some will run on metered cloud APIs. Some will move to **local inference** on hardware you own - modern NVIDIA accelerators, the

new generation of inference-optimised silicon from AMD, Groq, Cerebras, and Apple - running open-weight models like Qwen, Llama, DeepSeek, or Mistral. The economics of "should this run on someone else's GPU or mine?" are about to become a routine architectural decision.

3. Your governance model

When every team can spin up an AI agent that bills by the token, **shadow AI becomes shadow spend.** Marketing's experiment with Claude. Finance's bot on ChatGPT. The product team's vibe-coded Cursor project. The HR team's CV-screener on Gemini. Each one a credit card. Each one untracked. Each one a potential data leak, compliance breach, and budget surprise - usually all three at once.

What "good" looks like

The organisations that will navigate this well are the ones that, in the next six to twelve months, do four specific things.

1. Build a single AI inventory. What models, from what vendors, are being used where, by whom, against what data? Most CIOs cannot answer this today.

2. Establish a tiering strategy. Not every problem needs the most expensive model. A well-designed AI estate routes routine work to cheap or local models, escalates to mid-tier for nuance, and only reaches for the frontier when the value justifies the cost. This is the same logic as compute right-sizing in the cloud - and it is, if anything, more impactful.

3. Put real telemetry in place. Token-level visibility per user, per team, per use case. Cost per task. Cost per outcome. Cost per customer served. Without this, every conversation about value is opinion-based.

4. Bring the conversation back to value. AI spend without a clear link to a business outcome is the most dangerous kind of cost: invisible enough to grow, plausible enough to justify, and large enough to hurt when somebody finally looks. FinOps for AI is not about minimising spend - it is about **making spend defensible.**

These are not hypothetical concerns. They are the same conversations Gary's clients are having today as they look at their first agentic deployments, their first Copilot Studio rollouts, their first attempts to wire ChatGPT into their service desks.

The window is open. Briefly.

Here is the strategic observation that should land hardest.

In 2014, a relatively small number of organisations took cloud FinOps seriously **before** their bills got out of control. They are the ones who, ten years later, have predictable cloud unit economics, healthy engineering accountability, and CFOs who trust the IT budget.

The vast majority did not and have spent the decade since playing catch-up. Some still are.

In 2026, the same window is open for AI. It will close faster - probably within twelve to eighteen months - because adoption is accelerating and the costs are now visible. The organisations that get ahead of this now will spend the late 2020s with a defensible, optimised, well-governed AI estate. The ones that do not will spend the same period explaining to their boards why their AI line item tripled, twice, in eighteen months, with no measurable change in output.

This is not a tooling problem. There will be plenty of tools. It is a **strategy, architecture, governance and finance** problem - exactly the kind of problem that requires a clear-eyed, vendor-neutral perspective from someone who has done it before, in cloud, and is now doing it again in AI.

How NeuraSec helps

This is precisely why we have built **AI FinOps Readiness** as a new service line at NeuraSec.

It is not a tool. It is not a dashboard. It is a structured engagement that delivers what most organisations are missing right now:

- **An AI inventory and current-state assessment** - what you're using, where, by whom, at what cost, with what risk
- **A tiering and routing strategy** - flat-rate, metered cloud, and local inference workloads mapped to the right model for the right job
- **A target operating model** - who owns AI spend, how it's forecast, how it's charged back, who signs off the next agentic deployment
- **An EU AI Act and governance overlay** - because cost discipline without compliance is just a different kind of risk
- **A 12-month roadmap** - sequenced, prioritised, and costed, with the quick wins called out first

If today's Copilot news made you wonder what's hiding in your own AI estate, that's a good instinct. It is almost certainly hiding more than you think.

We'd like to help you find out before the next billing cycle does.

👉 **Learn more about AI FinOps Readiness:** https://www.neurasec.co.uk/ai-finops.html

--

Gary Willock is an independent Enterprise Architect and the founder of NeuraSec, an independent AI consultancy helping UK public sector and enterprise clients adopt AI securely, defensibly, and economically. With 20+ years in enterprise architecture, cyber, and cloud — and as the founder of both Magma Cloud and Digital Scaffold — Gary brings the same FinOps discipline that reshaped cloud to the AI era.

📩 **Get in touch:** https://www.neurasec.co.uk · contact@neurasec.co.uk

---

Sources: GitHub Blog (27 April 2026 announcement), TechCrunch, Techtimes, How2Shout, Enterprise DNA, GitHub Community Discussion #192948, FinOps Foundation State of FinOps 2025, Flexera State of the Cloud 2025.