NIVAAN LTD

What AI Is Costing You

A short guide for financial services firms that have started using AI and do not have a FinOps function

Ten places the spend hides. Four reasons your cloud playbook will not work on it. A twelve point self assessment. And the first thirty days of control, none of which requires buying anything.

Version 1.0 · August 2026 · Nivaan Ltd · cg@nivaan.co.uk

1. The position you are probably in

Someone in your organisation is using AI in production. Possibly several teams. There is an invoice, or part of one buried inside a cloud bill, and it is larger this quarter than last.

If your board asked next week what AI costs the firm and what it is producing, the honest answer would take some finding. Not because anyone has been careless, but because the spend arrived faster than the process for governing it, and because it does not arrive in one place.

This is close to universal right now. The State of FinOps 2026 survey found that the share of FinOps functions managing AI spend went from 31% two years ago to 98%. That shift did not happen because AI became a strategic priority. It happened because the invoices arrived and nobody was ready for them.

There is a second, more encouraging fact. On cloud cost, a large bank with a dedicated team is years ahead of a mid sized firm. On AI cost, almost nobody is ahead. The playbooks are being written now. A firm that gets its arms around this in the next six months is not catching up. It is early.

> **Who this is for**
>
> A financial services firm with cloud infrastructure, no dedicated FinOps function, AI in use somewhere in the business, and a CTO or finance director who owns cost alongside a great deal else. If that is roughly you, everything here applies.

2. Why your cloud playbook will not work on this

Cloud cost management has a settled shape. Provision infrastructure, monitor utilisation, right size what is oversized, forecast against a curve that behaves. It is not easy, but the feedback loop is understood.

AI cost breaks four assumptions in that loop.

One user action is not one billable event

A cloud instance runs for an hour and bills for an hour. An AI feature responds to one click by making a chain of model calls, invoking tools, retrying on failure and replaying earlier context. A single user action can produce twenty billable events. This is the single largest reason AI forecasts miss, and it gets worse as features become more autonomous rather than better.

The billing units are not comparable

Tokens, pages, audio minutes, seconds of generated video, provisioned throughput units. These cannot be added together or trended on one chart without converting to money first. Most internal dashboards do not, which is why very few firms have a single total.

Falling prices hide rising volume

Between early 2025 and early 2026, blended token prices fell by roughly two thirds. Bills went up anyway, because usage per unit of work grew faster than price fell. The consequence is counter intuitive but important: unit price is not a control lever. The number to watch is how many units your work consumes, not what each one costs.

Committed capacity does not correct itself

Variable spend at least signals when something is wrong. A spike is visible. Capacity you have reserved and are not using signals nothing at all. See section four, because this is where the largest avoidable cost usually sits.

3. Ten places AI spend hides

Most firms can point to one invoice. It is usually under half the total. The exercise below takes an afternoon and is worth doing before anything else.

Class	How it is billed	Where it usually sits
Generative text inference	Input, output and reasoning tokens	Direct provider accounts, or inside the cloud bill via Bedrock, Azure OpenAI or Vertex
Agentic and orchestrated systems	Tokens, multiplied by calls per action	Same invoice as the above and indistinguishable from it without tagging
Embeddings and retrieval	Tokens to create, storage and query time to hold	Split between a model provider and a separate vector database vendor
Generative media	Per image, per second of video, per character of speech	Marketing or product budgets, frequently a departmental card
Speech and language services	Per audio minute, per character	Contact centre and customer operations
Vision and document intelligence	Per page, per document	Back office operations, bought as a business tool rather than as AI
Predictive and classical machine learning	Compute hours, per prediction	Data team budget. Rarely counted as AI at all
Custom model work	Training hours, provisioned throughput	Cloud bill, and it charges whether used or not
Supporting infrastructure	Mixed	Data preparation, feature stores, labelling, guardrails, observability
AI inside software subscriptions	Per seat, per month	The software bill. Almost never counted as AI spend

The three that surprise people

- Generative media. Video is the most expensive unit in AI by a wide margin. A marketing team producing content can outspend engineering without appearing on any technical radar.
- Document intelligence. Priced per page. A back office processing a million pages a year has a large, entirely predictable AI bill that nobody labels as AI, and per page rates are negotiable at volume but rarely renegotiated.
- AI inside software. Copilot seats, AI features in your CRM, your service desk, your meeting tools. In a firm that is not technology heavy this is often the single largest AI line, and it arrives on the software bill where nobody looks for it.

> **Practical first step**
>
> Search twelve months of card and expense statements for provider names before you interview anyone. People report what they remember buying. Statements report what is renewing.

4. Sort by cost profile, not by technology

Once you have the list, resist the urge to organise it by vendor or by technology. Organise it by how it is charged, because that determines both what the waste looks like and what you can do about it.

Profile	What it is	What waste looks like	The lever
Usage variable	Scales with volume, costs nothing when idle	Doing the work in a more expensive way than necessary	Fewer units per outcome
Provisioned and committed	Billed on capacity, not use	Idle. Paying in full for capacity nothing is calling	Resize or remove. The saving is immediate and total
Seat and licence	Billed per person per month	Seats bought and never signed into	Reclaim at renewal
Human in the loop	Labelling, annotation, review queues	Reviewing work that did not need reviewing	Sampling against a confidence threshold

The flat line

Of the four, the second deserves particular attention, because it produces the most expensive and least visible failure in AI cost.

A dedicated endpoint, a reserved GPU or a block of provisioned throughput bills at its full rate whether or not anything calls it. On a cost chart it draws a perfectly straight horizontal line. Every instinct a finance or engineering team has says that a flat line is healthy. Spikes get investigated. Flat lines get ignored.

A pilot finishes, the team moves on, and the endpoint it stood up keeps charging. Twelve months later it is still there. In many firms this single item is larger than everything variable put together, and it is almost never volunteered, because nobody thinks of it as a cost. They think of it as something they set up once.

> **Ask this question this week**
>
> What dedicated endpoints, reserved capacity or provisioned throughput do we have, who requested each one, and what is its current utilisation? If the answer takes more than a day to produce, that is itself the finding.

5. A twelve point self assessment

Score each capability Crawl, Walk or Run. Crawl scores one, Walk two, Run three. The total runs from twelve to thirty six. Be honest rather than generous, because the value of this is in showing where to start.

#	Capability	Crawl	Run
1	Inventory	One or two invoices are known	All ten classes inventoried and reviewed quarterly
2	Attribution	One key serves everything, no breakdown possible	Per team keys or a gateway tagging every request
3	Committed capacity	Reserved capacity exists, nobody tracks utilisation	Every line has an owner, a utilisation floor and a review date
4	Unit economics	Cost is a monthly total	Cost per claim, per document, per ticket, reported monthly
5	Forecasting	Last month plus a growth guess	Driver based, with an autonomy expansion scenario
6	Budgets and caps	Nothing. The invoice is the first signal	Hard caps and anomaly alerts on every provider
7	Ownership	Nobody, or technology generally	Every line has a named owner who sees their number
8	Model selection	One model for everything regardless of difficulty	Routing by task difficulty, reviewed as prices move
9	Efficiency practice	No caching, no output limits, untuned retrieval	Applied as engineering standards across teams
10	Shadow AI	Personal cards, unknown scale	Consolidated, with a sanctioned route that is easier
11	Seat discipline	Bought, never reviewed	Licences reconciled against actual sign in activity
12	Value measurement	Cost and value discussed in separate meetings	Each workload carries a cost and a measured benefit

Reading your score

Score	Stage	Where to start
12 to 18	Exposed	Spend is arriving and nobody owns it. A single runaway workload could become material before anyone notices. Start with inventory, caps and committed capacity.
19 to 27	Aware	The total is known, the breakdown is not. Start with attribution and unit economics, because until you have them you cannot decide whether to optimise or invest.
28 to 36	In control	Spend is attributed, owned and tied to outcomes. Move from control to efficiency and value.

Most firms beginning AI adoption score between fourteen and twenty, including firms with strong cloud cost discipline. The assumptions do not carry across. A low score here is a starting position, not a failure.

6. The first thirty days

None of the following requires buying a product, deploying a tool or starting a programme. All of it can be done by people you already employ.

Week one

- Search twelve months of card and expense statements for AI provider names. Do this before asking anyone anything.
- Filter your cloud bills for AI, machine learning and GPU service codes, including provisioned throughput and reservations.
- List every dedicated endpoint and reserved capacity block, with who requested it and its current utilisation.

Week two

- Put a hard spend cap and an anomaly alert on every provider account, routed to a named person rather than a shared mailbox.
- Shut down or resize anything from the endpoint list that is not being used. This is usually the largest single saving available and it needs no engineering.
- Cross reference AI seat counts in your software subscriptions against actual sign in activity.

Weeks three and four

- Name an owner for every cost line. Not a committee. One person per line who sees their own number each month.
- Bring shadow AI onto company accounts, and make the sanctioned route easier to use than the unofficial one. If it is harder, people will keep using cards.
- Pick your single largest workload and work out what one unit of business work costs. One claim, one document, one ticket. That number will start more useful conversations than any dashboard.

> **If you do only one thing**
>
> Do the endpoint and reserved capacity review in week one. It takes a day, requires no tooling, and in most firms it is worth more than everything else on this list combined.

7. When it is worth getting help

Much of the above is genuinely a do it yourself exercise, and this guide is written so that you can. Three situations tend to be the exception.

- Attribution is impossible internally, because everything runs through one key and reconstructing it requires sampling work nobody has time for.
- The people who could do it are the same people delivering the roadmap, so it slips quarter after quarter.
- The number has to go to a board or a regulator, and an internal estimate of your own waste is a harder document to defend than an independent one.

How Nivaan works

Stated plainly, because the objection behind most polite refusals is not price. It is that nobody has capacity for another initiative.

What we need	What we do not need
Read access to billing and usage exports	Access to any production system
One hour with your executive sponsor	Customer, member or personal data
One hour with finance	Software to buy, install or integrate
One hour with your platform or cloud lead	An agent or collector deployed in your estate
One hour with an engineer close to the workloads	A procurement exercise
Two weeks of application or usage logs	An internal project team or steering group

> **Six hours**
>
> That is the total time your organisation spends across a three week review. Four conversations and one data request. Everything else happens off to the side and comes back as a finished report: a figure in pounds with the evidence behind it, a breakdown by team and unit of work, savings ranked by value against effort, and a ninety day plan your own team can execute.

Independence

Nivaan sells no cloud, no tooling, no licences and no managed services, and takes no commission from any vendor. There is no product behind the advice. A recommendation to keep doing exactly what you are doing is an outcome we are paid the same for.

About

Nivaan Ltd is an independent cost advisory practice for UK financial services, drawing on twenty years of programme and portfolio delivery at HSBC, Barclays, Lloyds Banking Group, RBS, Refinitiv, Capgemini and Sopra Steria.

cg@nivaan.co.uk · nivaan.co.uk

Figures cited from the State of FinOps 2026 survey, FinOps Foundation published research, and 2026 analysis of enterprise AI billing. Nivaan Ltd is registered in England and Wales. This guide is general information, not advice on your specific circumstances.