

The hidden cost of AI: why most businesses are paying more and getting less

AI spend is fragmented, surprising and rarely tied to outcomes. This paper shows where the money goes, why the bill keeps surprising finance teams, and the discipline that turns AI into a cost the board can see and a return it can measure.

Executive summary

In one paragraph

Most UK businesses are now paying for AI in at least four places at once: per-seat licences, consumption-billed cloud usage, AI add-ons priced into existing software, and tools expensed by individuals. Nobody adds it up. The result is a bill that is larger than anyone budgeted and a return that nobody can evidence. 78% of IT leaders have been hit by unexpected charges from consumption or AI pricing, AI spend rose 108% in a year and 36% of licences go unused (Zylo, 2026). Gartner warns that scaling costs can be misjudged by 500 to 1,000% (Gartner, 2024). Meanwhile 95% of enterprise generative AI pilots show no measurable return (MIT NANDA, 2025) and only 37% of firms report any effect on earnings (McKinsey, 2026). The fix is not to spend less on AI, and not to slow it down; it is to be able to say yes to more of it with the cost known. It is to run AI spend the way a finance team already runs cloud and SaaS: find every line, tie each to an outcome, use the right model for the right job, set kill criteria for pilots, and report one number a month to the board. This paper explains where the money goes, why the bill surprises, and what that report should contain.

1. Where the money goes

Ask a finance director what the business spends on AI and the answer is usually a single subscription line and a shrug. The real figure is spread across four kinds of spend, each owned by a different part of the business and each invisible to the other three.

The first is **per-seat licences**. These are the visible ones: an AI assistant bolted onto the office suite, a paid chatbot plan for a team, a code assistant for developers. Microsoft 365 Copilot in the UK is listed at £13.80 per user per month on promotion, rising to £16.10 from 1 October 2026 (Expertsure, 2026; confirm on microsoft.com/en-gb). For a 500-person business that is roughly £96,600 a year at the standard price, before anyone asks whether all 500 people use it.

The second is **consumption spend**. When the technology team connects a business process to an AI model through a cloud provider, the bill is metered by usage rather than by seat. It arrives inside a cloud invoice that is already large, and it grows with volume rather than with headcount. This is the category that produces the surprises in the next section.

The third is **vendor add-ons**. The CRM, the HR system, the helpdesk and the accounting package have all added AI features, and most have repriced. That cost appears as a renewal increase rather than a new line, which is why it is rarely counted as AI spend at all.

The fourth is **expensed tools**. Individuals and teams pay for AI tools on company cards and claim them back. Zylo's analysis of SaaS spend found that ChatGPT was the most-expensed application of the year, that business units rather than IT control 81% of software

spend, and that AI spend rose 108% year on year (Zylo, 2026). This is the category the CFO sees last and understands least.

Table 1. Four kinds of AI spend, and who can see them

Kind of spend	Where it appears	Who usually sees it	Typical problem
Per-seat licences	Software budget, IT or department	IT, procurement	Seats bought for everyone, used by some
Consumption (usage-billed)	Inside the cloud invoice	Technology team	Grows with volume; no budget line; late surprise
Vendor AI add-ons	Renewal uplifts on existing software	Contract owner, if anyone	Not counted as AI spend; paid regardless of use
Expensed tools	Expense claims, company cards	Finance, after the fact	Untracked, unapproved, duplicated across teams

UK data shows how quickly this has built up. Two-thirds of UK businesses say they have invested in AI, with 33% spending under £25,000, 18% between £25,000 and £100,000, 8% between £100,000 and £250,000 and 7% more than £250,000 (Lloyds Business Barometer, 2026). The Government's own adoption research found that 76% of businesses rate high cost as a significant barrier (DSIT, 2026). Worldwide, Gartner forecasts AI spending of \$2.59 trillion in 2026, up 47% in a year (Gartner, 2026). The money is real. The question is whether anyone in the business can see all of it.

2. The surprise bill

Consumption pricing is the part of AI spend that most resembles a utility bill, and it behaves like one. Nobody notices the meter until the invoice arrives.

Zylo found that 78% of IT leaders had been hit by unexpected charges from consumption or AI-based pricing, and that 61% had cut or delayed projects because of unplanned software cost rises (Zylo, 2026). Gartner's warning is blunter: chief information officers can miscalculate the cost of scaling AI by 500 to 1,000%, and 90% of CIOs say cost is limiting the value they get from AI (Gartner, 2024). A pilot that costs a few hundred pounds a month at pilot volumes is not a guide to what the same process costs when it handles every customer enquiry.

The result is that cost becomes a brake on use. McKinsey's 2026 survey found that 20% of organisations say AI operating costs limit how much they use it (McKinsey, 2026). That is the worst of both worlds: the business has paid to build something and then rations it because it cannot predict what running it will cost.

A pilot that costs a few hundred pounds a month is not a guide to what the same process costs when it handles every customer enquiry.

Seat licences fail in the opposite direction. Where consumption spend is under-estimated, seat spend is over-bought. Zylo reports that 36% of software licences go unused (Zylo, 2026). AI seats are especially prone to this, because they are often rolled out to whole departments on the assumption that everyone will want them, and because usage is rarely measured after the launch email. A business paying for 500 AI seats with a third of them idle is spending roughly £32,000 a year, at the Copilot price above, on nothing.

The common thread is that AI spend has no owner. Consumption sits with the technology team, seats with IT or procurement, add-ons with whoever signed the renewal, and expenses with finance. Each sees a fragment. The board sees none of it until something goes wrong.

3. Paying for pilots that go nowhere

The second hidden cost is not the bill for AI that runs. It is the bill for AI that never did.

The evidence here is consistent across every major study. MIT's research found that 95% of enterprise generative AI pilots deliver no measurable return on the profit and loss account, and that only 5% of custom-built AI tools reach production (MIT NANDA, 2025). Gartner predicted that 30% of generative AI projects would be abandoned after proof of concept by the end of 2025, citing poor data quality, inadequate risk controls, escalating costs and unclear business value (Gartner, 2024). BCG found that 74% of companies struggle to achieve and scale value from AI, and that the successful minority put 70% of their effort into people and processes, 20% into technology and only 10% into the algorithms themselves (BCG, 2024). McKinsey's latest survey finds only 37% of organisations reporting any enterprise-level effect on earnings from AI, and only 6% qualifying as high performers (McKinsey, 2026).

These figures describe large enterprises with dedicated data teams. A mid-market business with a smaller bench and less patient shareholders should expect the odds to be no better.

What goes wrong is rarely the technology. It is the absence of a decision structure around it. Pilots start without a stated outcome, so there is nothing to measure them against. They run without a budget ceiling, so the consumption bill drifts. They continue because stopping feels like admitting failure, so a project that should have been closed in week six is still consuming money and attention in month nine. Gartner's four causes of abandonment, data, controls, cost and value, are all things that could have been assessed before the first pound was spent.

The bill for AI that never ran is often larger than the bill for AI that does.

The cost is not only cash. Every pilot that goes nowhere consumes the time of the subject-matter experts who were seconded to it, and it uses up the organisation's willingness to try again. The 74% who struggle to scale (BCG, 2024) are not short of ideas. They are short of a way to tell a good idea from a bad one early enough to matter.

4. Right model, right job

The single largest lever on AI running cost is one most executives have never been told about. AI models are not one thing. They come in sizes, and the price per task between the largest and the smallest differs by orders of magnitude. The most capable models are extraordinary at difficult reasoning and are priced accordingly. The smaller and older models are a fraction of the cost and are perfectly adequate for the majority of routine tasks a business actually runs: classifying an email, extracting fields from a form, summarising a call, drafting a first reply to a standard query.

Most businesses, and many suppliers, default to the most expensive model for everything, because it is the one they have heard of and because nobody is measuring the difference. That is the equivalent of sending every parcel by same-day courier. The right discipline is to decide, for each task, what level of capability the job needs, and to pay for that level and no more.

Cost per task, not cost per month

The unit that makes this manageable is cost per task: what it costs to handle one enquiry, one document, one case. Once a business knows that number for each use case it can compare models on it, set a ceiling for it and watch it over time. A monthly consumption invoice tells you that spend went up. A cost-per-task figure tells you why, and whether the answer is a cheaper model, a shorter prompt or a process that is being run more often than it should.

Tiered use of models

In practice the answer is usually a tier. A small, cheap model handles the routine majority. Anything it cannot handle with confidence is passed to a more capable model. Anything that model cannot resolve goes to a person. Each hand-off is logged. The effect is that the expensive model is used for the minority of cases that need it, and the business pays frontier prices only where frontier capability earns its keep.

Human-in-the-loop as a cost control

Keeping a person in the loop is usually presented as a safety measure, and it is. It is also a cost measure. A human reviewer at the point where the AI is least certain catches the expensive mistakes before they reach a customer, and generates the ground truth that tells you which model is adequate for which task. Businesses that design the human step in from the start tune their models faster and cheaper than those that add review later because something went wrong.

Kill criteria for pilots

Finally, every pilot should start with the conditions under which it will be stopped: a cost-per-task ceiling, an accuracy floor, a date and an owner. A pilot that has not met its outcome by the date is closed, and the reasons are written down. This is the mechanism that converts the 30% abandonment rate (Gartner, 2024) from a waste into a learning cost, because the money stops early and the lesson is kept. It is also what makes the successful 5% (MIT NANDA, 2025) recognisable: they cleared a bar that was set in advance.

5. One number a month

None of the above works unless someone reports on it. The board does not need a dashboard of model metrics. It needs one page a month that answers four questions: what are we spending on AI, what are we getting for it, what is wasted, and what will next quarter look like. This is the FinOps discipline that cloud teams learned a decade ago, applied to AI; the FinOps Foundation reports that 98% of practitioners now manage AI spend, up from 63% a year earlier (FinOps Foundation, 2026). The mid-market business does not need a FinOps team. It needs the report.

Table 2. What a board-level AI cost and value report contains

Line	What it shows	Why the board needs it
Spend by tool and team	Every AI cost across licences, consumption, add-ons and expenses, allocated to the team that incurred it	The total, for the first time, and who owns each part of it
Cost per outcome	For each live use case, the cost of one handled enquiry, document or case, and the trend	Whether the spend is buying anything, and which use cases are worth scaling
Unused and duplicate licences	Seats not used in 30 days; tools that duplicate each other across teams	Immediate savings, and the rationalisation decisions to take
Forecast	Expected spend next quarter at current volumes, and at planned volumes	No more surprise bills; scaling decisions made with the cost known
Effectiveness measures	Accuracy, escalation rate and time saved per use case, against the target set at the start	Evidence, rather than opinion, on whether AI is working
Pilots and kill criteria	Each pilot's ceiling, floor, date and status	Which pilots to stop, and which to fund
Exceptions	Spend or usage outside policy since the last report	The control picture, in the same document as the cost picture

The value of this report is not the numbers on any given month. It is that the same numbers appear every month, so that trends are visible and decisions are taken while they are still

cheap. A business that sees consumption spend rising 20% a month in month two makes a different decision from one that sees the annual total in month twelve. A business that sees a third of its seats idle in the first quarter cancels them at renewal rather than paying for another year.

The board does not need a dashboard of model metrics. It needs one page a month that says what AI costs, what it returns and what to stop.

The report also changes the conversation with suppliers. A vendor asked for cost per task, accuracy and a kill criterion at the scoping stage is a vendor who has been told that the business intends to measure them. That alone removes a large share of the pilots that would otherwise have gone nowhere.

6. What to do next

The order of work matters, because it is easy to spend money on AI cost management before knowing what the costs are. The sequence below is how GenWA works with clients, and each step produces something the board can use on its own.

Start with an Independent AI Audit. In three weeks, an independent review finds every AI tool and cost line in the business across licences, cloud consumption, vendor add-ons and expenses, along with what data goes into each, who owns it and what it is for. The output is a board-ready report with a risk register and a 90-day plan. For most businesses this is the first time a single figure for AI spend has existed, and it usually contains the unused licences and duplicate tools that pay for the audit.

Then set the rules with AI Policy & Controls. The cost rules sit alongside the data rules: which tools are approved, who may buy AI on expenses, what model tier is the default for routine work, what a pilot must define before it starts and when it must stop. Written for a UK business, aligned to UK GDPR, ISO 42001 and SOC 2, with the training and evidence pack to show it is followed.

Then connect it all through Business Watch. Business Watch is the central hub for all your AI systems: the one place every AI tool, model, agent and vendor connects through, gets its rules, is costed and is measured. It shows spend by tool and team, cost per outcome, unused licences, forecast and effectiveness alongside usage against your rules and data in and out, and it produces the monthly board report described above with risk and return side by side, without anyone compiling it by hand. It flags the consumption spike or the idle seats while there is still time to act, and it gives the board the cost picture it needs to fund more AI with confidence.

Build with control when there is something worth building. When the audit and the report show a use case that clears the bar, it should be built with the right model tier, human review at the right points, cost per task measured from day one and kill criteria agreed

before the first invoice. That is how a pilot becomes one of the 5% rather than one of the 95%.

If the business cannot state its total AI spend to within 20%, or cannot name the cost per task of its most-used AI process, the practical first step is a 30-minute conversation about the audit. Write to hello@genwa.co.uk.

Sources

1. Zylo, 2026 SaaS Management Index, January 2026. <https://zylo.com/news/2026-saas-management-index>
2. Expertsure, Microsoft Copilot review with UK pricing, August 2026 (secondary; confirm on microsoft.com/en-gb). <https://www.expertsure.com/uk/ai-tools/microsoft-copilot-review/>
3. Lloyds Bank Business Barometer, "Impact of AI adoption on business", March 2026. <https://www.lloydsbankinggroup.com/media/press-releases/2026/lloyds/impact-of-ai-adoption-on-business.html>
4. Department for Science, Innovation and Technology, AI Adoption Research, February 2026. <https://www.gov.uk/government/publications/ai-adoption-research/ai-adoption-research/>
5. Gartner, "Gartner forecasts worldwide AI spending to grow 47% in 2026", May 2026. <https://www.gartner.com/en/newsroom/press-releases/2026-05-19-gartner-forecasts-worldwide-ai-spending-to-grow-47-percent-in-2026>
6. Gartner via CIO.com, "CIOs view cost management as possible AI value killer", November 2024. <https://www.cio.com/article/3608214/cios-view-cost-management-as-possible-ai-value-killer.html>
7. McKinsey, The State of AI 2026, August 2026. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
8. MIT NANDA, "The GenAI Divide", August 2025 (via Virtualization Review). <https://virtualizationreview.com/articles/2025/08/19/mit-report-finds-most-ai-business-investments-fail-reveals-genai-divide.aspx>
9. Gartner, "Gartner predicts 30% of generative AI projects will be abandoned after proof of concept by end of 2025", July 2024. <https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025>
10. BCG, "AI adoption in 2024: 74% of companies struggle to achieve and scale value", October 2024. <https://www.bcg.com/press/24october2024-ai-adoption-in-2024-74-of-companies-struggle-to-achieve-and-scale-value>
11. FinOps Foundation, State of FinOps 2026, February 2026. <https://www.linuxfoundation.org/press/state-of-finops-survey-ai-value-and-skills-top-priorities-as-finops-matures-across-technology-value-98-manage-ai-90-saas-64-licensing-48-data-center-1>

About GenWA

All the upside of AI. None of the surprises.

GenWA helps UK mid-market businesses take the upside of AI with confidence. We show you what AI is doing in your organisation, put business rules around it and make it pay. Our services are the Independent AI Audit, AI Policy & Controls, Business Watch (the central hub for all your AI systems), Build with control and Managed AI Ops. GenWA is an AWS Marketplace seller and our work is aligned to UK GDPR, ISO 42001 and SOC 2; we prepare you for certification.

John White, Founder & CEO. 30+ years driving digital transformation across financial services and fintech. Former CTO at Vanquis Bank (FTSE 250), Chetwood Bank, and Arrow Global/Intrum. Head of Solutions at Capital One UK, establishing their first international operation. Led major transformation initiatives at Barclays/Barclaycard. Previously CEO (SMF1) at Shoreditch fintech, achieving regulatory approval and launching credit products from concept to market in nine months.

Lukasz Piotrowski, CTO & Co-founder. 20+ years building and scaling engineering organisations across financial services and capital markets. Former Head of Engineering at Intrum UK and Chetwood Financial, delivering AWS-native contact centre, collections, and cloud-native banking platforms at scale. Deep capital markets experience from Numis Securities.

Contact: hello@genwa.co.uk · genwa.co.uk · linkedin.com/company/genwa