zapier

How to Govern AI Spend Without Sacrificing Speed

Moving from tokenmaxxing to maximizing value — a governance framework for capturing the full power of AI without the runaway cost.

AI spend governance

CONTENTS

What this paper covers

SECTION 01

The real question

Every enterprise conversation about AI eventually arrives at cost. And when it does, leaders tend to land on the same question: how do we spend less?

It is a reasonable place to start, but it usually leads to the wrong answers. A sharper question is: how much value are we getting for what we spend? That reframe changes what you measure, what you govern, and what you do when something looks wrong. It moves the goal from spending the least on AI to getting the most out of it.

AI cost problems come in two distinct shapes. Some organizations have sprawl: AI spreads team by team without a front door, and now finance sees a number, security sees a surface, and IT sees access, but nobody has the whole picture. Other organizations have the opposite problem: they have exec buy-in, but teams are still negotiating first approved use cases. The bill is not yet the issue. Durability, adoption, and who owns governance is.

Most AI cost frameworks are written for the first group. This paper is written for both, and it starts one level up, with the mindset that decides whether any governance program actually pays for itself.

SECTION 02

From tokenmaxxing to maximizing value

When AI is not delivering obvious ROI, the first instinct is to try more: bigger models, more tokens, inference on everything. Call it tokenmaxxing. It rarely works, and the reason is structural. The companies selling you tokens are incentivized for you to use more of them. The promise is value, but the product is tokens, and the default path pulls you toward spending more on inference whether or not it is paying for itself.

The thing worth internalizing is that value is not captured at the model. Models are increasingly commoditized — the easy part. Value is created where the work actually happens: when the people doing the work can use AI to reach into your systems, your data, and your processes and move real work forward.

So the right unit of measure is value per token. Tokens are the denominator. When you tokenmax, you pay inference rates for work a deterministic step could have done for a fraction of a cent. Maximizing value is the opposite discipline: use inference where it earns its keep — judgment, ambiguity, language, fuzzy matching — and fast, governed automation everywhere else.

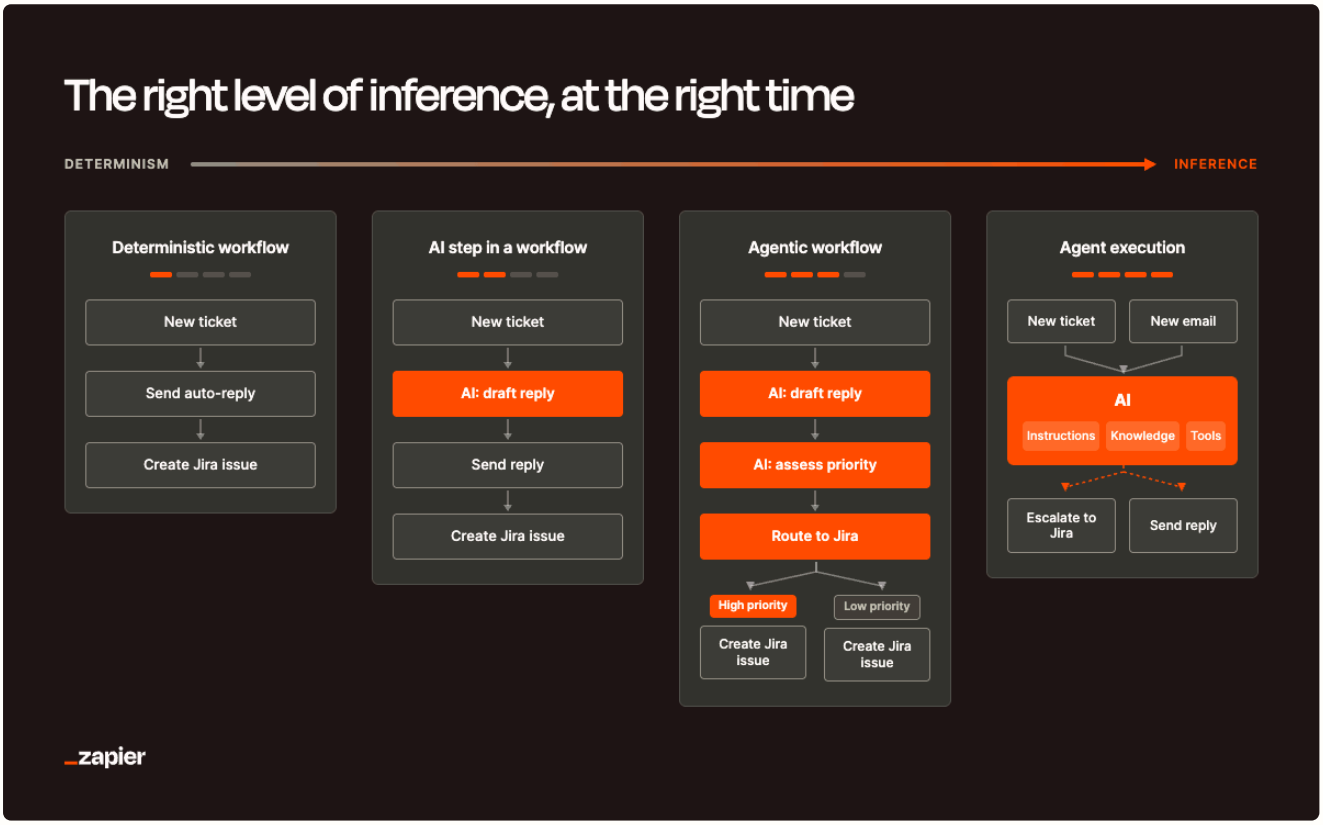
This is not an argument for using less AI. It is an argument for using AI where it creates value and letting predictable, cheaper automation carry the rest. The organizations that win will be the ones that know the difference — and can move fluidly between the two.

> "A lot of what we do that is repetitive could just be a workflow. You get way better cost structure and much more repeatability — you don't have to worry about hallucinations when you go deterministic. A workflow might be 80% fully deterministic, and then at one point it needs a summarization or a judgment call. That's the sweet spot."
>
> **Bryan Helmig** — Co-founder and CTO, Zapier

SECTION 03

The spectrum: deterministic to agentic

It helps to picture the work on a spectrum. On one end, fully deterministic, rule-based flows: predictable, observable, cheap. On the other, fully agentic flows that use inference for nearly every step: flexible, but expensive and harder to predict. In between sits the practical middle: deterministic workflows with AI steps added exactly where they do the kind of work AI is genuinely good at.



The key idea is that a workflow should not sit still on this spectrum. Over time, a healthy workflow tends to get more deterministic — more predictable, more observable, and cheaper with every run — while keeping agentic flexibility for the genuine edge cases where judgment is required. Start agentic to move fast; earn determinism as the pattern becomes clear.

Across the workflows we see built on Zapier, a consistent pattern shows up: most of the steps inside an "agentic" build don't actually need inference at all. Teams reach for the agentic build experience because it's the easiest way to get started — not because every step requires judgment. The result is workflows that call a model to do work a deterministic step could have handled just as well, every single run. That gap, repeated across an organization, is what tokenmaxxing looks like on an invoice.

> “Controversial take: I think automation is more powerful than AI. AI is probabilistic. Automation is deterministic. Deterministic workflows get you there 99.9% of the time.”
>
> **Chris Morrison** — Business Analyst and AI Lead, Erewhon

SECTION 04

What getting the mix right is worth

The savings are not marginal. Using AutomationBench, Zapier's internally-developed benchmark for measuring how models perform on real automation work, we analyzed complex, multi-system business workflows and recomputed their cost when the deterministic steps were automated and AI was reserved for the steps that truly need it.

~71%

Customers with complex business workflows can expect to save roughly 71% of the cost by making the workflow deterministic where they can and using AI only in the steps that need it.

SOURCE: AUTOMATIONBENCH*

**Methodology: In an internal Zapier study, we modeled 24 AI agent workflows to classify which steps required AI versus deterministic automation. Savings reflect the estimated cost reduction if deterministic steps were routed away from the LLM, using modeled LLM pricing of $1.50/M input tokens and $9.00/M output tokens, and a modeled $0.001/action fee based on Zapier's Teams plan — not guaranteed rates for all plans. Not based on live customer workflows. Results will vary.*

The word complex matters. A workflow that just pulls today's meetings from a calendar and summarizes them will save less as there is not much deterministic plumbing to remove. But a workflow that looks up data in one app, applies your business rules, and updates another app based on the result is mostly deterministic plumbing wrapped around a few genuine judgment calls. Those are the workflows where the economics change dramatically — and enterprises run them by the thousands.

The pattern underneath the number is consistent. The steps that genuinely need AI are interpreting free text, fuzzy-matching records, deciding under ambiguity, and composing language. Almost everything else, endpoint discovery, reads, structured writes, is deterministic work that does not need to be paid for at inference rates.

SECTION 05

Where you are on the curve changes everything

Before you set a governance policy, be honest about your actual starting point. Two common pitfalls pull in opposite directions, each requiring a different response.

PITFALL ONE

Governance arrives after the fact

AI spread fast, shadow tools multiplied, and leadership is now trying to draw a map of something that has already been built without one. Finance sees accelerating SaaS line items, IT has a backlog of access reviews, and Security wants a full audit. The instinct is to lock down. That instinct is usually wrong — not because the risk isn't real, but because restriction without visibility just pushes the work underground.

PITFALL TWO

Governance becomes the bottleneck

Leadership has an AI mandate, exec buy-in exists, but teams are still working through IT queues, security reviews, and approval processes that weren't designed for AI-speed iteration. Adoption stalls not because teams don't want to use AI, but because the path to an approved first use case takes three months. The risk here is different: not data exposure, but adoption failure and the reputational cost of being the organization that "has AI everywhere except where it matters."

The governance model that works in both situations has the same three parts: visibility, policy, and action at the workflow layer. The priority order and the calibration of controls is what changes.

SECTION 06

The common mistake: managing the invoice, not the work

AI spending does not live in one place. It shows up in model provider bills, SaaS renewals, developer tools, embedded AI features in existing contracts, internal agents, and workflows running in the background. A single Zap can call a premium model inside a loop. A single agent can take actions across Slack, Google Drive, Jira, and your CRM. The invoice does not tell you which workflow generated which cost, or whether the cost was worth it.

That context gap is the core problem. A high bill might mean a team found a workflow that saves hundreds of hours. It might also mean someone left an agent running on a premium model all weekend with no retry limit. Both look the same on the invoice.

This is why connector sprawl compounds the cost problem. When every team wires AI into the tools they need independently, IT loses the map: not just of spend, but of data flow, access, and accountability. The same workflow that wastes money can also expose data, create duplicate work, or take the wrong action at scale. Cost governance and security governance are the same conversation.

The way out is not a bigger dashboard. Dashboards tell you what happened after it happened. Governance needs action at the workflow layer, where policy becomes a real-time nudge, an approval route, a pause, or a coaching message before the spend lands on an invoice nobody can explain.

> “Integration plans don't fail because they're badly written. They fail because they're written too far from the work. The people who live inside the workflow see the failure modes everyone else underestimates.”
>
> **William Ascough** — Founder, Voltt

SECTION 07

Governance is a steering wheel, not a brake

The most important framing shift in AI cost governance is this: high usage is not automatically a problem.

Some people are excellent AI operators. They are automating repetitive work, compressing research cycles, and producing better outputs faster. These people should not be slowed down. They should be studied and scaled.

Other usage is genuinely wasteful. A premium reasoning model handles a routine summary. A background workflow retries without limits. A team buys three tools that do the same job. That is tokenmaxxing in miniature. It should be coached, routed, or stopped.

The job is to tell the difference — in real time, at the workflow layer, not three weeks later when the invoice arrives. The question is never just "how many tokens?" It is "how much value per token?"

FOR ORGANIZATIONS IN EARLY ADOPTION

This framing is especially important. The first governance failure in organizations still building their first approved use cases is almost never runaway spend. It is premature restriction that kills momentum before adoption establishes itself. The governance posture at this stage is: make the approved path easy, make the unapproved path visible, and move fast to unblock the first ten use cases before teams lose patience and find workarounds.

> “One of the bigger things we're going to automate next is Zapier approvals, which provides much-needed control and flexibility. That will allow us to further democratize access to Zapier to individuals who don't have access today.”
>
> **Marcus Saito** — Head of IT and AI Automation, Remote

FOR ORGANIZATIONS MANAGING SPRAWL

The governance posture shifts toward coaching and consolidation. Not a crackdown, but rather a systematic effort to route the highest-cost behaviors toward better defaults, consolidate duplicate tools before renewal, and give IT a clean map of what is running and who owns it.

SECTION 08

The operating model: watch, decide, act

A practical AI spend governance program has three parts. Most companies complete only the first.

Watch — build a map, not just a dashboard

The first step is visibility, but the right kind. Leaders need to see AI spend by employee, team, cost center, tool, model, workflow, business process, data sensitivity, and, critically, outcome.

Outcome is the variable most dashboards miss. If support AI usage increases 40% and ticket resolution improves 25%, that is likely a good investment. If sales research spend doubles but the pipeline does not move, that needs review. The goal is not to shame high users, but to understand the work.

> FOR EARLY-STAGE ORGANIZATIONS
>
> The watch layer starts simpler. Map what tools exist, who owns them, what use cases are live, and which workflows touch sensitive data. A first AI spend inventory by tool, team, and owner is more valuable at this stage than a sophisticated analytics platform.

> FOR SPRAWL-STAGE ORGANIZATIONS
>
> The map needs to capture workflow metadata, not just token counts. What triggered the usage? Which systems did it touch? Was a human in the loop? Did it complete successfully? These are the questions a vendor invoice cannot answer.

Zapier is useful here because it sits at the workflow layer where these signals live. It connects billing data, identity systems, approval workflows, collaboration tools, and AI usage signals into one governed operating layer — so visibility is a byproduct of how work runs, not a separate reporting project.

Decide — policy that people can use

Policy only works if teams can remember it and apply it without a compliance review for every workflow.

The most important policy decision is model defaults. For most organizations, the biggest cost lever is not whether people use AI, but which model they reach for by default. A clear policy looks like: use the cheapest model that gets the job done; premium reasoning models are for high-stakes judgment, not routine formatting; every production agent needs an owner; every background workflow needs a run limit.

> "More reasoning isn't always better. There tends to be a point of diminishing returns for most tasks. You've got to match the reasoning level to the complexity of the task. Otherwise, it's just sitting there spinning — thinking for no real reason."
>
> **Lucas Bergstrom** — Lead Product Manager for AutomationBench, Zapier

The second lever is determinism. For any workflow that runs often, ask which steps actually need inference and which are deterministic plumbing, then move the plumbing off the model. This is where the value-per-token discipline turns into real savings: roughly 71% on the complex, multi-system workflows enterprises run most.

FOR EARLY-STAGE ORGANIZATIONS

Policy is mostly about establishing the approved path clearly enough that teams use it. The failure mode is not policy that is too loose. Instead, it is a policy document that lives in a wiki and gets ignored because the approved path is slower than the workaround. Make the approved workflow the easy workflow. That is the governance win at this stage.

FOR SPRAWL-STAGE ORGANIZATIONS

Policy needs thresholds. Budget limits by team or cost center. Daily and monthly spend limits by workflow. Run limits for agents. Approval gates for premium model access. Review requirements for sensitive data workflows. Required owners for every workflow in production.

Speed without accountability becomes chaos. Accountability without speed becomes bureaucracy. Good policy holds both. Define what good looks like, not just what is forbidden.

Act — turn policy into workflow

This is where most governance programs break. Policy exists with limited control. Leaders know what good looks like. No mechanism routes behavior toward it.

Act means converting policy into workflows that coach, route, approve, pause, or escalate — in real time, at the moment the decision can still change. A few examples:

Coach the user in the moment

If a user runs a routine summarization task on a premium model, send a Slack DM while the pattern is fresh. A monthly report does not create that learning. A real-time message does.

> "You used a premium model for a summary. The default model handles this well and costs less. Want to switch future runs?"

Route premium usage through approval

If a workflow crosses a spend threshold, route it to the team lead. The approver does not need to understand every token. They need a clear decision.

> "This workflow is projected to spend $2,000 this month on premium model access. Approve, switch to the default model, or pause for review."

Pause runaway agents

If an agent runs more than expected and exceeds a retry threshold, pause it automatically. One control prevents a surprise bill.

> "This agent has run 47 times in six hours with 18 failures. It is paused until the owner reviews cost and retry behavior."

Escalate sensitive workflows

If a workflow sends customer data to an external AI service, route it for security review before activation. Instead of a permanent block, this is a human-in-the-loop gate for the specific risk that exists.

FOR EARLY-STAGE ORGANIZATIONS

The most valuable act-layer workflow is one that makes approval fast. A new use case request routed to IT, responded to in 24 hours, with a templated approval path. That is governance that accelerates adoption rather than stalling it.

SECTION 09

Agents change the risk profile

The next phase of AI spend will not come primarily from humans typing prompts. It will come from agents doing work: researching, summarizing, classifying, updating records, creating tickets, sending messages, and triggering follow-up steps across systems.

A person might run one expensive prompt. An agent might run hundreds. A person might make one mistake. An agent with tool access can repeat a mistake across systems before anyone notices.

Every production agent should have a clear business owner, a defined purpose, approved tools and data sources, model defaults, spend limits, run limits, retry limits, human review points, action logs, and a shutdown path.

That may sound heavy, but it does not need to be slow. Most of it can be templated and applied consistently across many agents from one governance layer. The mistake is treating agents as experiments after they are already in production. Once an agent can act across tools, it needs an operating model with a running set of controls.

> **FOR REGULATED INDUSTRIES**
>
> This is where the audit trail matters most. Every action an agent takes needs to be attributable, reversible where possible, and reviewable after the fact. Governance that cannot answer "what did this agent do and why" is a liability.

SECTION 10

How Zapier runs this internally

It is reasonable to ask whether this framework holds up in practice. Here is how Zapier runs it on ourselves.

We make AI spend personal, not just visible

Every Zapier employee receives a monthly direct message summarizing their own AI tool usage and spend — not a team rollup, not an anonymized average. The reasoning is simple: the person who can change a usage pattern is the person creating it. A finance dashboard cannot coach an individual toward a cheaper model. A monthly note that says "here is what your AI use cost this month, and here is where it went" puts the signal exactly where a decision can be made.

We set a default model and reserve premium models for expensive problems

When we updated internal defaults, the rationale was explicit: the efficient model is the right choice for most everyday tasks. Premium reasoning models cost significantly more per token and belong to complex, high-stakes work. We did not block the premium models. We made the cheap path the easy path and the expensive path the deliberate choice.

We define the threshold that turns "use the best model" into "let's talk"

Our internal trigger is concrete: a sustained, significant cost increase over several consecutive months prompts a real review of model usage. That one rule does a lot of quiet work: it tells teams they have room to experiment, it tells finance that growth alone is not an alarm, and it names the specific pattern that warrants a harder look.

We track cost at the unit level, not just the total

Per-task model cost, cost per request trend over time, average model requests per tool call — these are the metrics that tell you whether a bill went up because of more usage, pricier models, or chattier workflows. That distinction changes the response.

We are still building toward a governance layer that acts automatically

Today, much of this runs on humans reading reports and sending messages. The direction is toward automating the watch-to-act loop: workflows that catch usage patterns and route the response without waiting for someone to notice. We recommend that path to every enterprise we work with, and we are walking it ourselves.

SECTION 11

What this looks like in practice

SCENARIO A · MANAGING SPRAWL

A support organization rolling out AI triage at scale

The team uses AI to classify tickets, summarize customer history, suggest replies, and escalate complex issues. Usage climbs quickly. With a governance layer in place, the operating view includes:

- Support AI spend by queue, team, and workflow
- Cost per resolved ticket
- Premium model usage by exception type
- Agent run counts and failure rates
- Customer data fields used in AI workflows

Actions are connected to those signals: routine triage defaults to a lower-cost model; complex tickets can escalate; sensitive field workflows require review before activation; agents pause when retry rates spike; managers receive alerts when spend crosses thresholds; finance sees cost per outcome, not just total cost.

SCENARIO B · EARLY ADOPTION

A professional services firm still getting AI off the ground

The challenge is not reining in runaway spend. Instead, it is getting durable adoption with the right risk posture. Governance here looks different.

- The IT team defines an approved use case list with a clear on-ramp: request, review, approve in 48 hours or less.
- The first ten approved workflows have owners, model defaults, run limits, and a check-in at 30 days.
- Sensitive data workflows have a separate path with a security review.
- The executive team receives a monthly adoption report: how many workflows are in production, which teams have adopted, what value is visible.

Cost governance comes later, once usage is established and the map is real.

SECTION 12

A 30-day first version

You don't need a six-month program. You need a focused first version.

Days 1–7

Map the workflows, not just the spend

Identify your highest-volume and highest-cost agentic workflows first — ranked by how often they run and what they cost per run, not by team or tool. For each one, estimate how much of it genuinely requires inference versus how much is doing repeatable, rule-based work. You're looking for the workflows people built agentically because it was easy, not because every step needed judgment.

DELIVERABLE A ranked list of your highest-volume agentic workflows, each tagged with an estimated share of steps that could run deterministically.

Days 8–14

Pick your first hardening targets

Take the top two or three candidates from week one and define what "hardened" looks like for each: which steps convert to deterministic rules, which stay agentic because they genuinely need judgment, and what the fallback path is for edge cases. Pair this with a simple default-model policy for whatever inference remains: cheapest model that clears the bar, premium reserved for real judgment calls. Treat that policy as the supporting lever, not the centerpiece.

DELIVERABLE A hardening plan for 2-3 workflows, plus a one-page model-default policy for what's left.

Days 15–21

Build the mixed-determinism workflows

Turn the plan into working automation. Convert the identified steps into deterministic actions, keep agentic steps only where judgment is truly required, and add the controls that make it safe to run unattended — run limits, approval routing, action logs.

DELIVERABLE 2-3 live workflows running in mixed-determinism mode, each with a logged before/after on cost and reliability.

Days 22–30

Measure the payoff, queue the next batch

Ask what actually changed for the workflows you hardened: cost per run, run time, error rate, how often a human still has to step in. Use that evidence to pick the next batch from week one's ranked list, and check whether the model-default policy is holding for whatever is still agentic.

DELIVERABLE An executive readout showing cost and reliability before and after hardening, plus the next batch of workflows queued for the same treatment..

The first 30 days should produce a working example of maximizing value, not just a map and a policy. Leave with a couple of hardened workflows, real before/after numbers, and a queue of what's next

SECTION 13

The leadership frame

AI spend governance is not about making AI smaller. It is about making AI accountable, and there is a difference. The goal is not to tokenmax, and it is not simply to spend the least. It is to maximize value: to get the most out of every dollar of AI spend.

The organizations that win the next several years will not be the ones with the strictest bans or the most granular dashboards. They will be the ones that can see the work, understand the cost, and guide teams toward better decisions while the work is still happening. They will know when a high bill is a sign of value and when it is waste. They will know which agents can act autonomously, which workflows need human review, and which tools should be consolidated before the next renewal cycle.

They will also know where they are on the adoption curve and govern accordingly. Early-stage buyers need governance that accelerates adoption, not governance that resembles the IT queue they are trying to move past. Sprawl-stage buyers need governance that creates accountability without killing momentum.

The watch-decide-act model works for both.

The cost of not acting now is real. AI is moving faster than monthly invoice cycles. Every month without a watch layer is a month of spend you cannot explain. Every month without an act layer is a month of policy that exists on paper but changes nothing.

Start with visibility and add clear defaults. Close the loop with workflows that coach, route, and act in real time. That is the governance that compounds — both controlling cost, and building the operating intelligence to know when to accelerate.

SECTION 14

Where Zapier fits

Maximizing value only works if you can actually move between deterministic and agentic work on one platform, while governing all of it in one place. That is the shape of Zapier.

Build across the full spectrum

Teams can build agentically, describing what they want and letting Zapier Copilot assemble it. Deployment happens across the complete range from fully deterministic to fully agentic, adding AI steps only where they earn their keep. The same workflow can carry a premium model on the one step that needs judgment and run everything else as cheap, predictable automation.

Connect to the systems where work happens

Zapier maintains pre-built integrations across more than 9,000 apps, with tens of thousands of triggers and actions kept up to date. Everything is exposed to agents through Zapier MCP and available in code through the Zapier SDK. That is the connective tissue that lets AI reach into your real systems and data instead of stalling at the model.

> "Before Zapier MCP, if Zendesk didn't have a trigger for something, the workflow just didn't exist. Now, as long as I have the data, I can build exactly what I need."
>
> **Corey Smith** — Senior Technical Support Engineer, ClickUp

Govern from a single control plane

The building can be distributed to the citizen builders across your company; the governance should be centralized. Who is allowed to do what, how data is allowed to flow, and the visibility to see what is actually happening, in one place. That is the layer that turns "AI everywhere" into AI you can stand behind.

SECTION 15

Next steps with Zapier

Zapier sits at the workflow layer where finance, IT, security, and business teams meet — the one place AI spend governance actually comes together. We connect the systems of record, watch for events, route approvals, send coaching messages, pause or change workflows, and give leaders a business view of AI usage across every surface their teams use.

See how Zapier maps to your stack and your maturity stage.

Whether you're reining in sprawl or getting your first ten use cases off the ground, we can help you build the watch-decide-act loop.

Talk to our team