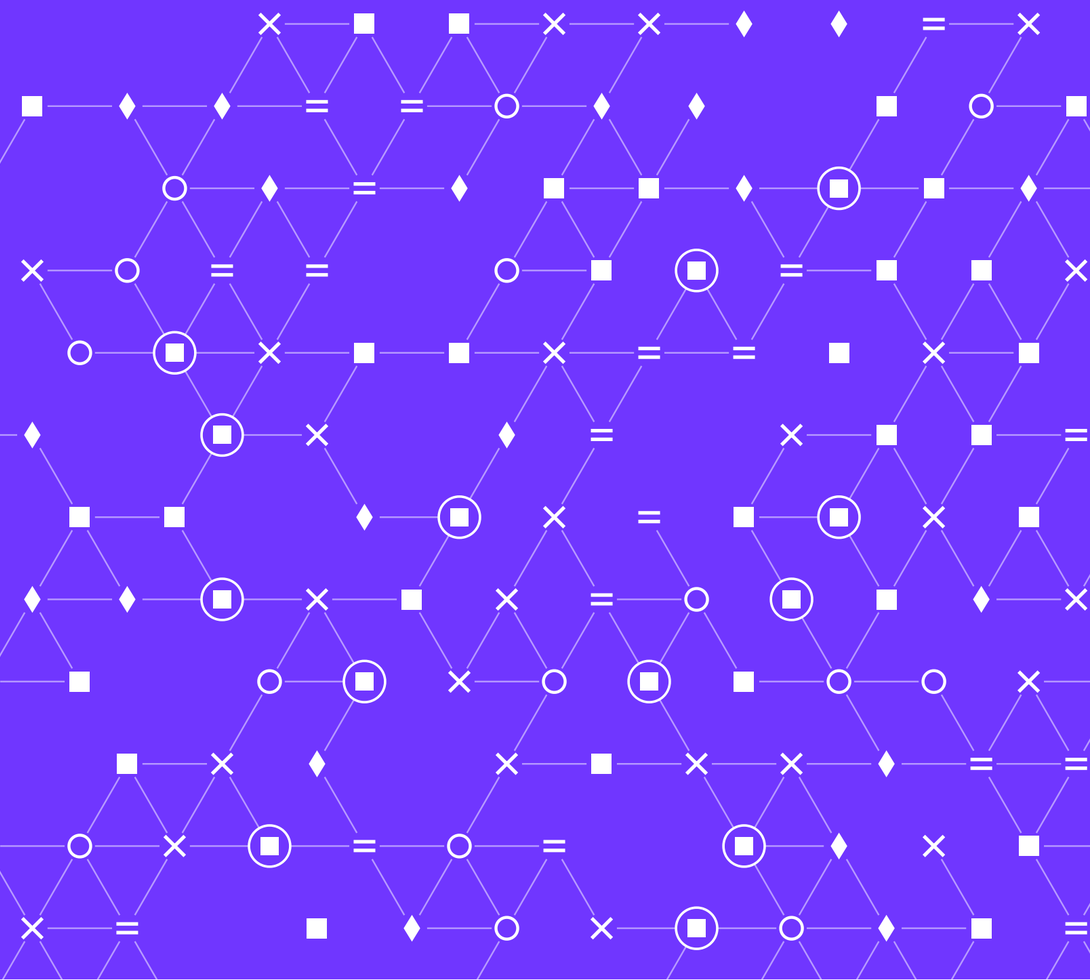
PLAYBOOK

HOW TO ASSESS AI SPEND

Three questions to answer before you tighten team budgets



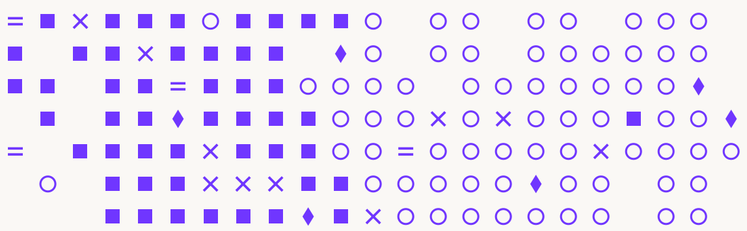
Numeric

INTRODUCTION

Your AI bill went up. Why?

Did the team get more work done, or did an agent reread the same files? Is someone using a high-powered model for simple arithmetic? Is spend climbing inside a tool whose AI features nobody tracks? The invoice won't tell you.

At Numeric's Inflection 2026 summit, Glean VP of Finance Michael Miao and Figma CFO Praveer Melwani discussed how they make these calls. In separate conversations, Numeric Head of Finance Sam Isaac and Solutions and AI Operations Lead Geoff Lew explained how they budget for AI and assess what it costs to run.



THREE QUESTIONS FOR AN AI SPEND REVIEW

1 How do we find out what's driving AI spend?

Track every charge and find what changed.

2 How much should we be willing to spend?

Set allowances based on what the work delivers.

3 How can we make useful AI work cost less?

Benchmark a workflow, change one thing, and compare.

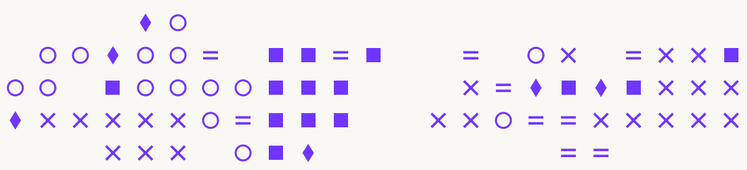
QUESTION 1 OF 3

How do we find out what's driving AI spend?

Track recurring AI work before it becomes a large expense. Geoff Lew gives two reasons:

1. As software vendors tie more pricing to usage, costs that a flat subscription hides will show up more directly.
2. Model providers like Anthropic and OpenAI may begin charging more for usage.

Start with an inventory. "Track all the places where you are spending tokens," Sam Isaac advises. Variable AI charges also turn up inside support tools, knowledge platforms, and other software you already pay for.



MONTHLY AI SPEND INVENTORY

Put every charge in one report. List fixed fees and usage separately, then check the total against what you were billed.

Source	Billing period	Fixed fee	Usage charges	Team / owner
Model provider bills				
AI charges in existing software				
Total vs. billed				

WHAT TO NOTE

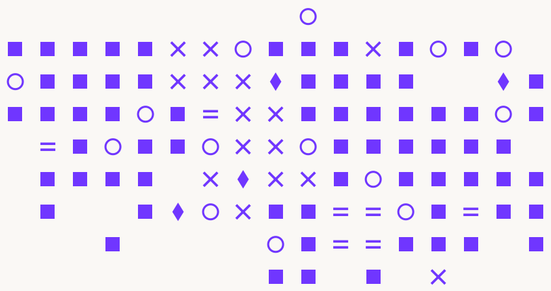
The more granular you can go, the better.

Break spend down by team and, where possible, by person. Keep the model and token types if you can see them. Miao checks a daily dashboard of tokens by model and type at Glean, then pairs usage with the rates Glean pays.

When a charge jumps, trace it to the work

For recurring jobs that make up a meaningful share of the bill, track the workflow, its owner, the number of runs, and what those runs cost. That record is what you compare against when the bill moves.

A department's spend might rise because more people started using a shared tool. Or the default model behind a workflow changed, on purpose or not.



DIAGNOSING AN INCREASE

1

Compare to a similar period

Pull the workflow's run count and average cost per run.

2

Did run count rise?

More people or more jobs. Ask whether the extra runs did useful work.

3

Did cost per run rise?

Check which model ran and how much data it read.

4

Is it one account?

Ask what that person is running before you call it waste.

More runs, same cost per run

Usage grew. The question is whether the work it produced is work the business needs.

Higher cost per run

Something about how the workflow runs changed: the model, the amount of context, or the output it produces.

QUESTION 2 OF 3

How much should we be willing to spend?

“Tokens are just dollars in another currency,” Miao says. Sam Isaac asks that token spend tie back to a real outcome: a better product, more effective selling, or better results for customers.

Numeric runs a collections agent that follows up on invoices and escalates exceptions. Sam will spend more on it because it covers work Numeric might otherwise pay an AR contractor to do. At Figma, when someone on the credit and collections team changed roles, the team lead automated enough of the work to defer filling the seat.

Still, every leader believes that teams need enough room and budget to experiment.

At Glean, different roles get different budgets, and alerts show people what their choices cost. Reaching a limit starts a conversation about whether a team needs more.





QUESTION 3 OF 3

How can we make useful AI work cost less?

Geoff Lew's method: benchmark a recurring workflow, change one thing, and run the same inputs again.

1

Script predictable steps

If a step follows fixed rules, like totaling rows in a spreadsheet, test whether code can do it. Save model calls for interpretation and judgment.

2

Match the model to the task

Run a less expensive model on the same inputs and check what reviewers have to fix. Bigger models can churn and validate things they don't need to, Geoff says.

3

Keep context tight

A step that checks one entity's balance may only need that entity's records. Long sessions and repeated output drove up cost and lowered quality.

4

Cache reusable context

If every request resends the same instructions or reference material, see whether the provider can reuse it. Writes and reads may be priced differently.

5

Batch work that can wait

Submit jobs you'll review later, like extracting data ahead of close, as an asynchronous batch. Anthropic's Message Batches API charges half its standard rates.

6

Run shared work once

A shared dashboard can answer the same questions for everyone. Decide who owns it, how often it refreshes, and what data it needs.

"There are tasks that you're trying to … solve with a sledgehammer and you probably could do with a chisel."

Praveer Melwani, Figma

WORKSHEET

AI workflow cost checklist

Pick one recurring workflow. Establish a baseline, test a change, and compare.

1 Establish the baseline

- [ ] Identify the spending source, model, owner, and runs per month.
- [ ] Define an acceptable result, the review it needs, and the work it enables.

2 Choose a change to test

- [ ] Script predictable steps
- [ ] Match the model to the task
- [ ] Keep context tight
- [ ] Cache reusable context
- [ ] Batch work that can wait
- [ ] Run shared work once

Workflow _____ Owner _____

Business purpose _____ Review date _____

Record	Before	After
Model / API cost per run		
Expected runs per month		
Review / rework minutes per run		
Quality check passed? / Evidence		

3 Compare the result and decide

- [ ] Use comparable inputs. Recheck quality, review time, and total cost.
- [ ] Estimate recurring spend: average cost per run × expected runs, plus other charges.

Decision

☐ Keep ☐ Expand ☐ Optimize further ☐ Investigate ☐ Stop

Next action / owner _____

Next review _____

BEFORE YOU TIGHTEN BUDGETS

AI spend – a cause for conversation, not always concern

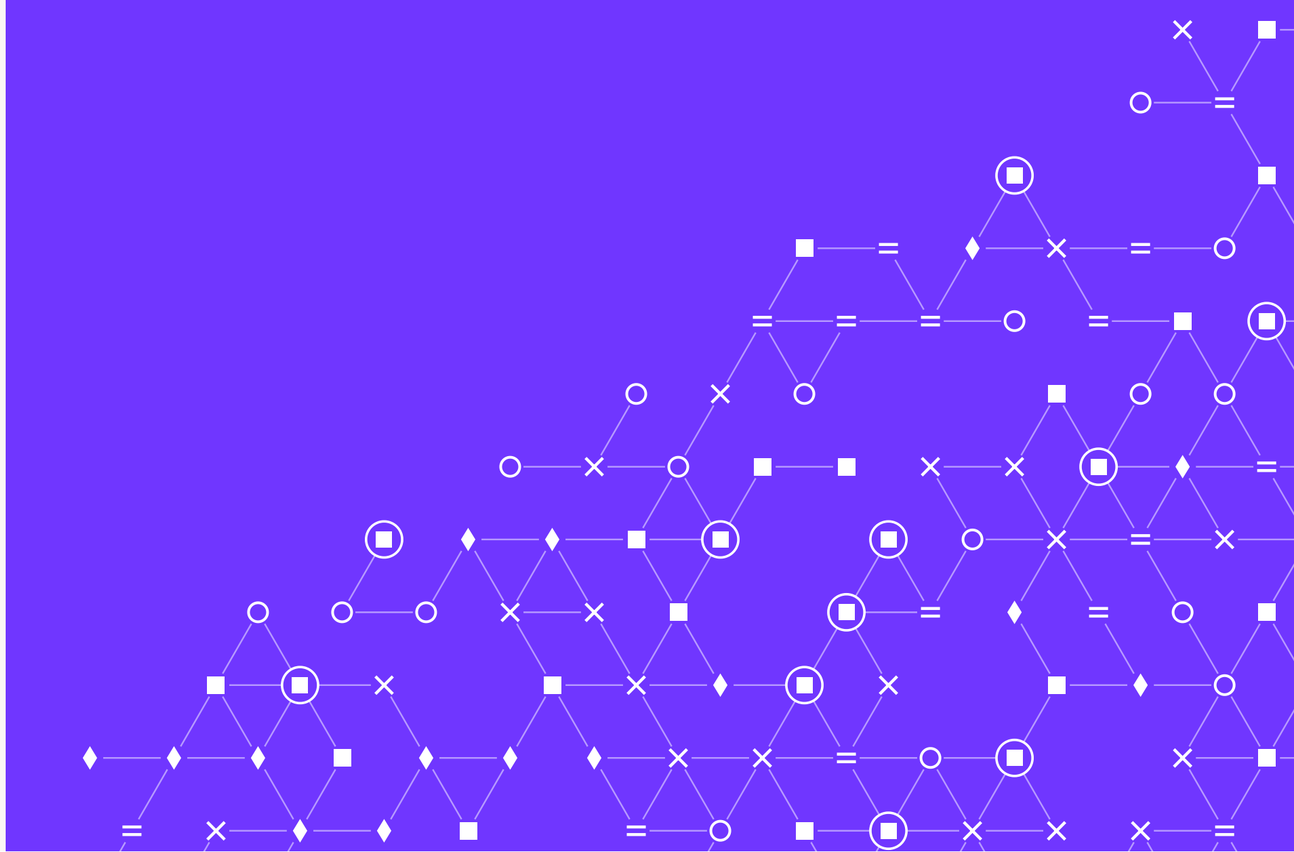
A growing AI bill often means people have found work AI does well. Numeric and Figma have each found workflows worth paying more for. Glean has found a solid way to monitor employee spend via alerts and conversations.

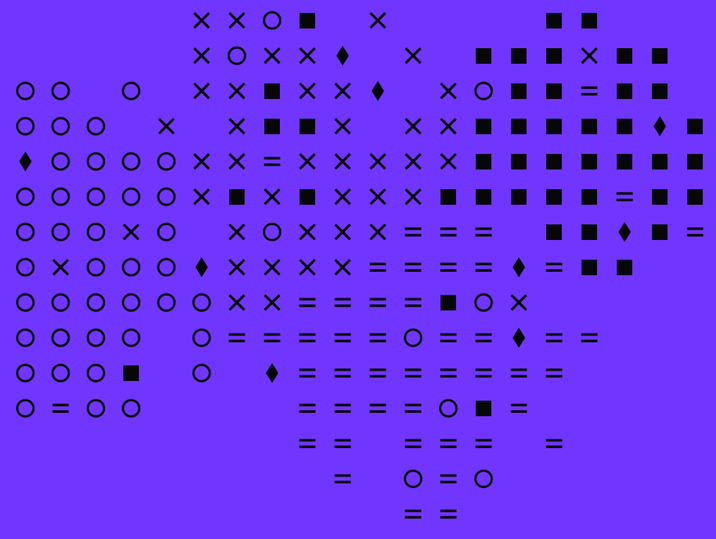
When it is necessary to review an increase in spend, treat the review as a conversation with the people doing the work. Ask what they're running, what it produces, and what they would change. Some workflows will earn a bigger allowance. Some will get cheaper with a smaller model or less context. A few may be worth rethinking or retiring, and the team running them can help you decide which.

Like any other resource, AI spend is a lever with variables: take stock of each, and use consistent evaluation to make decisions.

Hear how Glean and Figma make these calls

Watch the full Inflection panel





Numeric

