FOXTROT

AI Adoption That Scales. Value That Lasts.

WHITE PAPER

IS YOUR AI GETTING MORE EXPENSIVE FOR THE RIGHT REASONS?

Rising AI spend can be a healthy sign of adoption. It can also hide costs that get harder to remove with every workflow built on top of them. Here is how to tell the difference, and how to measure what AI really costs to deliver a result.

AI & Data Solutions | www.foxtrot-services.com

October 2026

AT A GLANCE

The cost of running AI is falling, but the AI bill is still rising.

>90%

Projected drop, from 2025 to 2030, in what it will cost GenAI providers to run inference on a one-trillion-parameter model

5 to 30×

More tokens per task that agentic models require compared with a standard GenAI chatbot

→ **"As token consumption rises faster than token costs fall, overall inference costs are expected to increase."**

Source: Gartner, "Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025," press release, March 25, 2026.

KEY POINTS

- **Rising spend can be good news.** When AI takes on more valuable work across the business, higher cost may simply reflect higher adoption. The harder question is whether the economics improve as the system scales.
- **Meaningful savings can appear earlier than teams expect.** One construction technology company cut compute costs by 90% before its broader architecture was fully deployed.
- **Token counts tell only part of the story.** A single agent request can trigger model calls, data queries, tools and actions. The real cost sits in the whole workflow.
- **A more useful unit is the cost of completing the work successfully,** including retries and the moments a person has to step in.
- **Model flexibility has economic value.** The right model for a workflow today may not be the right one a year from now.

CONTENTS

01 A rising AI bill can be a healthy sign

02 Meaningful savings can show up earlier than expected

03 What looks small in a pilot gets expensive at scale

04 Follow the cost back to the work

05 Measure the cost of finished work

06 Know what each agent costs and what it returns

07 Flexibility has economic value too

08 Know what you're paying for

01 A rising AI bill can be a healthy sign

Much of that increase can reflect healthy adoption.

As AI moves from experimentation into day-to-day operations, spend often rises for a good reason. More people are using it, and they're using it across more of the business.

A workflow that began with one team gets picked up by another. A manual task gets automated. New use cases are added, more data moves through the system, and work that once happened occasionally starts running all day.

If AI is doing more valuable work across the business, higher spend may simply reflect higher adoption. The test is whether the economics are improving as the system scales.

Industry forecasts point the same way. Gartner expects the cost for providers to run inference on a one-trillion-parameter model to fall by more than 90% between 2025 and 2030. It also expects overall inference costs to rise as token consumption grows faster than token costs fall; agentic models, for example, can require 5 to 30 times more tokens per task than a standard GenAI chatbot. For enterprises, the issue is whether that extra consumption delivers enough value to justify it.

When costs start climbing, the instinct is often to look for places to cut. We think it's worth understanding what's driving the increase first.

02 Meaningful savings can show up earlier than expected

You don't need a fully scaled AI environment to find meaningful savings.

We saw this firsthand with a construction technology company. Before the broader architecture was fully deployed, our team analyzed the company's existing data infrastructure and reduced compute costs by 90%.

Just as notable was how early the savings showed up. The company didn't need to wait for a major redesign or a fully scaled AI environment to uncover them. There was already enough happening inside the existing infrastructure to materially improve the economics.

90%

reduction in compute costs before the broader architecture was fully deployed

EARLY RESULT

CASE IN POINT | CONSTRUCTION TECHNOLOGY

A company using autonomous systems to build homes had advanced technology, but the operation around it was still reactive. Equipment downtime was frequent, machine data sat across disconnected systems, and the field knowledge that explained failures stayed with the crews who experienced them. Foxtrot is bringing real-time machine telemetry and field insight together in one operating foundation on Palantir Foundry, revealing patterns across maintenance, equipment failures, materials and personnel.

That's one reason we think cost deserves attention earlier than it often gets. By the time AI is widely deployed and the expense becomes uncomfortable, inefficient choices may already be embedded in workflows the business relies on. Each new workflow built on an inefficient pattern makes that pattern harder to change.

The 90% reduction was one part of a broader engagement. Cost efficiency sat alongside work to improve equipment reliability and build the technology needed for where the company wanted to go next. The goal was to create an operating foundation that could support the business as it grew.

Cost was part of that conversation from the beginning.

03 What looks small in a pilot gets expensive at scale

In a pilot, inefficiency is easy to miss.

An extra query may barely register when a workflow runs twenty times. Carrying more context than necessary may not seem consequential. Using a highly capable model for a relatively simple job may be perfectly manageable. Repeat those choices thousands of times a day and the economics change.

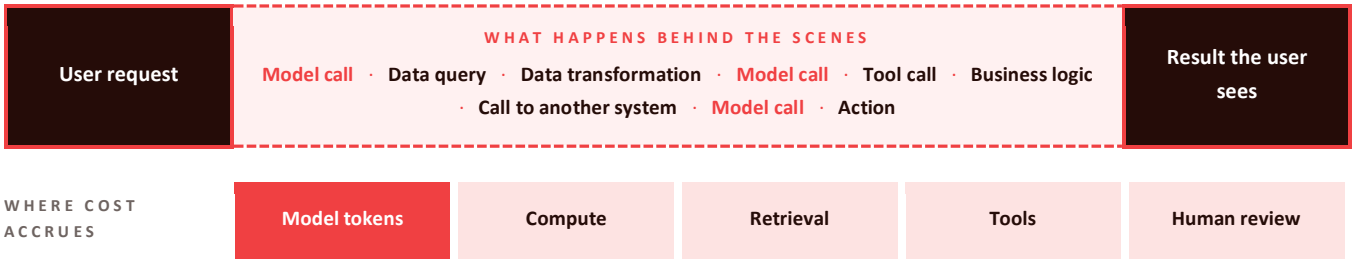
EXHIBIT 1 THE SAME CHOICES AT TWO SCALES

DESIGN CHOICE	IN A PILOT	IN PRODUCTION
One extra query per request	Barely registers across twenty runs	Repeats thousands of times a day
More context than the task needs	Seems inconsequential	Travels through every interaction and adds work each time
A highly capable model on a simple job	Perfectly manageable	Paid for on every run, even where a lighter model may perform just as well

Agents multiply the work behind each request

A single request may kick off several model calls, data queries, tools and actions before the user ever sees the result. All of that activity may be justified. It is worth checking whether the work happening behind the scenes is necessary for the outcome you need.

EXHIBIT 2 ANATOMY OF A SINGLE AGENT REQUEST



Token usage captures the model layer. The full cost of a request can reach all five.

That's also why token usage alone can be misleading. The model is only one part of what's happening. Data may need to be queried and transformed. Logic runs. Context moves through the workflow. Other systems may be called along the way.

A better measure is what the entire workflow costs to run and what the business gets in return.

04 Follow the cost back to the work

When compute spend starts moving, we want to understand what changed.

EXHIBIT 3 SIX QUESTIONS TO ASK WHEN AI SPEND MOVES

01	Did adoption genuinely grow?
02	Did one workflow become much more active?
03	Is the system repeatedly processing information it could reuse?
04	Is it carrying more context than the task requires?
05	Is the model appropriate for the work it's doing?
06	Can the team trace the spend back to the workflows responsible for it?

In a large enterprise, these questions are harder to answer than they look. We look at the whole workflow because the savings can sit in very different places. Depending on what the system is doing, the opportunity may sit in any of six places.

EXHIBIT 4 SIX PLACES TO LOOK FOR SAVINGS

Context management

Carry the information a task needs through the workflow, and leave behind what it doesn't.

Caching

Reuse work the system has already done, from repeated prompt context to answers for repeat requests.

Model routing

Send each task to a model suited to it, lighter for simple work and more capable for complex reasoning.

Workload design

Shape how and when work runs to fit its real speed, reliability and volume requirements.

Repeated processing

Find the data the system queries and transforms again and again when it could reuse it.

Cost attribution

Trace spend back to the individual workflows and agents responsible for it.

No single architecture is cheapest for every AI workload, because workloads vary widely. A high-stakes decision that happens a few hundred times a month may justify a very different setup from a repetitive task running thousands of times an hour. A customer-facing workflow may have speed and reliability requirements that an overnight process doesn't. Some work warrants the most capable model available. Other work may perform just as well with something lighter.

EXHIBIT 5 WHAT DRIVES COST, BY TYPE OF WORKLOAD

WORKLOAD	WHAT TENDS TO SHAPE ITS ECONOMICS
High-stakes decision, a few hundred times a month	Accuracy and consequence. It may justify the most capable model and an extra check.
Repetitive task, thousands of times an hour	Volume. Small per-run savings compound, and a lighter model may perform just as well.
Customer-facing workflow	Speed and reliability requirements the business can't compromise on.
Overnight process	Looser latency requirements, which leave more room to trade speed for cost.

Spend should follow what each workload actually requires.

05 Measure the cost of finished work

Good AI economics are about spending appropriately for the job.

Using fewer tokens can lower costs. It can also create a poor trade if quality drops, people have to redo the work, or a workflow needs additional retries and human intervention to finish successfully.

There are times when spending more makes sense too. A simpler task may be well served by a faster, less expensive model. Complex reasoning or a high-value decision may warrant heavier compute. In a higher-risk workflow, an additional model check may justify the extra token spend if it meaningfully improves quality or safety.

This is why optimizing consumption in isolation only gets you so far. **A more useful measure is the cost of completing the work successfully.**

EXHIBIT 6 COST PER COMPLETED TASK, BY MODEL AND TASK TYPE

FOR EVERY 100 TASKS	COMPLEX TASK: LIGHTER	COMPLEX TASK: MORE CAPABLE	SIMPLE TASK: LIGHTER	SIMPLE TASK: MORE CAPABLE
Model cost per run	$0.10	$0.60	$0.10	$0.60
Completed correctly on the first pass	50	95	95	97
Model cost for 100 runs	$10.00	$60.00	$10.00	$60.00
Rework on the rest, at $4.00 per task	$200.00	$20.00	$20.00	$12.00
Total cost to finish all 100	**$210.00**	**$80.00**	**$30.00**	**$72.00**
Cost per completed task	**$2.10**	**$0.80**	**$0.30**	**$0.72**

Hypothetical figures, for illustration only. Assumes every task that misses on the first pass is caught and finished by a person at $4.00. On the complex task, the model that costs six times more per run finishes the work at about 62% lower cost. On the simple task, where both models succeed at similar rates, the lighter model finishes it at about 58% lower cost. Measuring at the task level is how a team knows which case it is in.

Three questions get closer to what the business actually cares about. What does it cost to run the workflow from beginning to end? How often does someone need to step in? What happens to that cost as volume grows?

A lower-cost model can still produce poor economics if it needs several attempts to get to the right answer. A more expensive model may be worth the additional spend when the work is complex or consequential. More

context can improve a result when the information matters. Carrying unnecessary context through every interaction simply creates more work.

That's why we increasingly think about cost at the level of the task, workflow or decision. It gives teams a much clearer way to judge whether AI is becoming more economical as it scales.

06 Know what each agent costs and what it returns

As agentic systems scale, cost has to be visible where decisions get made.

As AI adoption scales, companies need visibility into the full cost of agent execution and whether that spend translates into measurable business value. In its June 2026 forecast on AI coding costs, Gartner observed that "token discipline will not emerge through developer choice alone."

Agent-level attribution

Foxtrot can build an agent-level economics layer using execution traces, token metering and infrastructure telemetry to attribute model, compute, retrieval, tool and human-review costs to individual agents and workflows. Connecting that telemetry to business outcome data makes it possible to measure cost per successful task, intervention rates and net value against an agreed baseline.

EXHIBIT 7 WHAT THE AGENT-LEVEL ECONOMICS LAYER DOES

Capture	Attribute	Connect	Measure	Decide
Execution traces, token metering and infrastructure telemetry	Model, compute, retrieval, tool and human-review costs to each agent and workflow	Cost data to business outcome data	Cost per successful task, intervention rates and net value against an agreed baseline	Which agents to scale, which to optimize and which may no longer justify their cost

Shared-cost allocation

Multi-agent workflows add another layer of complexity. Several agents may share infrastructure and tools, or contribute to the same outcome, so understanding ROI can require shared-cost allocation and controls that prevent the same value from being counted more than once.

That gives teams a much clearer view of which agents are creating enough value to justify their cost. They can see which are worth scaling, which need to be optimized, and which may no longer justify the expense, while still maintaining quality, latency and reliability.

Our AI Spend & Control accelerator targets inefficiencies across data processing, context management, caching, model routing and workflow execution. The construction technology example shows how much value can sometimes be uncovered in infrastructure that already exists.

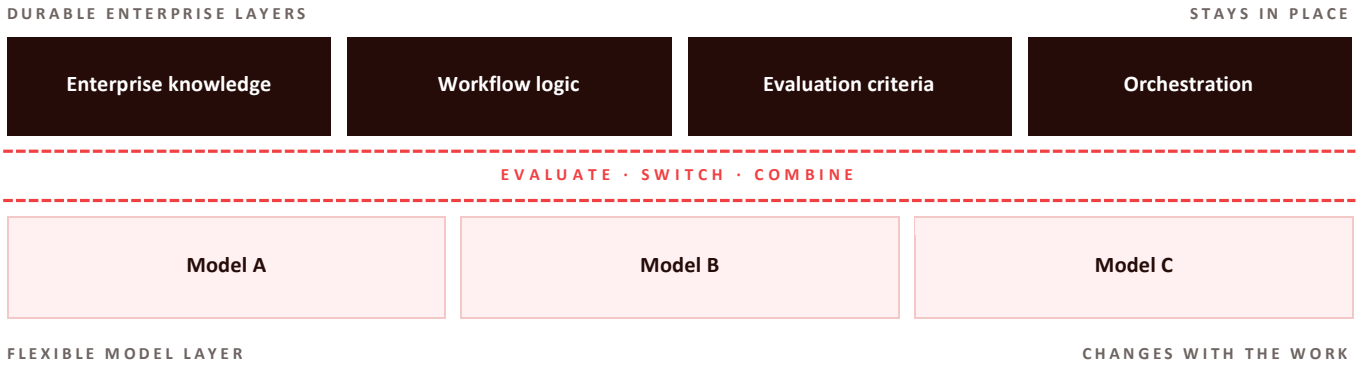
07 Flexibility has economic value too

A model that makes sense for a workflow today may not be the best option a year from now.

Another model may become faster, more capable or more economical for the job. Companies should be able to take advantage of that without rebuilding the business logic, workflows and context around the model.

That flexibility is part of AI sovereignty. Our approach separates enterprise knowledge, workflow logic, evaluation criteria and orchestration from the underlying model, so organizations can evaluate, switch or combine models based on what the work requires, including performance, security, latency, regulatory requirements and cost.

EXHIBIT 8 KEEP WHAT THE ENTERPRISE OWNS SEPARATE FROM THE MODEL LAYER



That flexibility matters economically too. Gartner expects that falling provider token costs “will not be fully passed on to enterprise customers.” As prices and capabilities move, the ability to evaluate, switch or combine models gives companies more room to respond.

08 Know what you’re paying for

The healthiest AI environment won’t always have the lowest compute bill.

A company putting AI to work across meaningful parts of its business may spend considerably more than one running a handful of pilots. That can be a good trade, as long as the company can explain why it is spending more.

Teams should know which workflows are driving spend, what those workflows cost to complete successfully and whether the value they’re creating justifies the expense. Some growth will reflect healthy adoption. Other growth may point to unnecessary work or an architecture that needs to evolve. Teams should be able to tell the difference.

As AI becomes more embedded in the business, its economics should become easier to understand and manage.

If the system is becoming more useful, you should be able to explain what you're paying for and why.

SELF-ASSESSMENT

Can you explain your AI spend?

Answer for the AI workflow that costs your organization the most today.

01	We can trace AI and compute spend back to the individual workflows and agents responsible for it.	☐ YES ☐ NOT YET
02	We know what the workflow costs to complete successfully, from beginning to end.	☐ YES ☐ NOT YET
03	We know how often a person has to step in to finish the work.	☐ YES ☐ NOT YET
04	The model matches the work, given its complexity, speed requirements and risk.	☐ YES ☐ NOT YET
05	The workflow carries only the context the task requires, and reuses what it has already processed.	☐ YES ☐ NOT YET
06	We could switch or combine models without rebuilding the business logic and context around them.	☐ YES ☐ NOT YET

Mostly "yes"

You have a stronger starting point for explaining spend today. The next test is whether that visibility holds as volume grows and new agents come online.

Several "not yet"

Your spend may contain blind spots worth investigating. Start with the highest-cost workflow and trace its cost, intervention, model fit and context from beginning to end.

If you can't explain where your AI spend is going, start with the workflow costing you the most.

Talk to Foxtrot about AI Spend Optimization.

About Foxtrot. Foxtrot Professional Services is a Palantir partner founded by Palantir alumni and enterprise operators. Forged in Foundry, Foxtrot builds production-grade, governed AI systems for commercial and government organizations that are designed for adoption, flexible across models and accountable for what they cost.