

HOLOKAI

WHITE PAPER · ENTERPRISE AI ECONOMICS

The Coming AI Bill

PAYING FOR INTELLIGENCE: THE HIDDEN COSTS OF ENTERPRISE AI AT SCALE

How to see, plan, optimize, and contain large language model costs across every model and provider, before the discounts disappear.



WRITTEN FOR

Enterprise Technology & Finance Leaders
CFO, CTO, CIO, VP'S Engineering & Architecture

EDITION

June 2026

Holokai · Navigating Enterprise AI

The Reckoning Nobody Has Budgeted For

Nearly every organization that embraced AI over the last two years is carrying a quiet problem on its balance sheet.

The pilots worked. The demos impressed leadership. Teams shipped copilots, assistants, summarizers, and agents—and so far the bill has been comfortable. That is by design. The industry is in a land-grab, and providers are subsidizing adoption with promotional pricing, generous free tiers, startup credits, and steep discounts meant to win your workloads and lock in your habits.

That window is closing. As the market matures, list prices firm up, credits expire, and “intro” tiers convert to standard rates. Usage that felt free quietly compounds into a line item nobody forecasted. Programs that scaled at discounted rates are about to watch that bill come due, and for many the jump will not be incremental. It is the kind of surprise that triggers a hard look from finance and a freeze on the very initiatives that were supposed to drive the next wave of business value.

What makes this frustrating is how foreseeable it is. Almost no one is instrumented to see it coming, and most organizations cannot answer basic questions today:

- *What is each AI capability actually costing us, per request and per month?*
- *Which models and providers are driving the spend—and is that allocation deliberate?*
- *If our usage doubles next year, what does the bill look like?*
- *Could we serve the same quality for less, on a different model?*

This paper is about turning those questions into a dashboard, an experiment, and a plan. It describes how the Holokai platform gives organizations four things they need before the reckoning arrives: the ability to see their AI spend precisely, plan for where it is heading, optimize what they pay per outcome, and contain cost as they scale.

WHY THIS MATTERS AT THE EXECUTIVE LEVEL

AI cost is more than a budget concern—it is a strategic exposure. Spend has to be attributable to a business outcome, defensible in a board review, and governed against the data-protection and auditability requirements your organization already carries. Because Holokai captures and attributes every request, cost can be allocated down to a business unit, a product feature, or an individual team. The same instrumentation that controls cost produces the accountability record leadership and auditors increasingly expect.

THE PROBLEM

Why AI Cost Is Uniquely Hard to Control

Cloud taught a generation of organizations to manage variable infrastructure spend. AI is harder, for a few specific reasons.

01

The unit of cost is invisible

Cloud bills you for servers and storage—things you provisioned and can see. AI bills you for tokens, a unit most stakeholders have never counted and cannot intuit. A single user asking a longer question, attaching a document, or triggering a multi-step agent can cost ten times what a simple query does, and nothing in the experience signals it.

02

Every provider counts differently

OpenAI, Anthropic, Google, and self-hosted options each describe and price usage in their own dialect. The moment you run more than one—which almost every serious organization now does—you lose the ability to compare them on a single ruler.

03

The pricing is a moving target

Rates change. New model generations launch monthly. Prompt caching, batch tiers, and large-context surcharges can swing the real cost of an identical request by an order of magnitude. A spreadsheet built last quarter is already wrong.

04

Cheap is meaningless without quality

The least expensive model that fails your task is the most expensive choice you can make. It costs you the spend and the outcome. Cost only means something next to whether the work was actually good.

Most AI cost management today is a guess dressed up as a spreadsheet. Holokai replaces the guess with measurement.

THE APPROACH

The Platform in One Picture

Holokai is an AI-agnostic control plane that sits between your applications and every model provider you use: a single pane of glass across all models, teams, and use cases.

Every AI request your organization makes flows through Holokai, so Holokai sees everything. More importantly, it translates every provider into one common language before anything is recorded. Whether a request went to OpenAI, Anthropic's Claude, Google's Gemini, or a model running on your own hardware, Holokai captures it in a single normalized format.

That one decision makes the rest possible. When every provider speaks the same language, you can finally:

- **Sum and compare spend across vendors on a common unit of measurement,**
- **Attribute every dollar to a model, an application, a team, or a user,**
- **Run experiments that pit models against each other on identical work, and**
- **Forecast tomorrow's bill from today's data.**

No more per-vendor spreadsheets, reconciliation projects, or blind spots between providers. One platform, one language, one source of truth for AI cost.

CAPABILITY 1

SEE · Total Awareness of AI Spend

You cannot manage what you cannot see. The first thing Holokai delivers is complete, real-time visibility into AI consumption and cost, at a granularity that turns mysterious monthly invoices into an itemized, queryable ledger.

For every request, the platform records what it consumed and what it cost: input and output tokens, the subtler categories that drive modern bills (cache reads, cache writes, and the “thinking” tokens reasoning models burn through), the response time, and the exact dollar figure. That figure is not an estimate. It is calculated against the rate in effect for that model on that day, including the caching discounts and tier surcharges public calculators ignore.

Because Holokai speaks one language across providers, the data rolls up into a unified picture you can slice however the business needs:

BY APPLICATION

Which product features and capabilities are the cost centers?

BY MODEL & PROVIDER

Where is the money actually going—and is that allocation deliberate and optimal?

BY TEAM OR USER

Who is driving consumption—and is it producing value?

OVER TIME

Is cost-per-outcome trending up or down?

This is the difference between receiving a bill and understanding it. When the invoice arrives, you already know what is in it, why, and whether it is reasonable—line by line and model by model. “The AI bill went up and we are not sure why” stops being a sentence anyone has to say.

THE AWARENESS DIVIDEND

Most organizations find, the first time they see this view, that a surprising share of their spend sits in a handful of features or a few power users—often producing marginal value. You cannot act on that until you can see it.

CAPABILITY 2

PLAN · Forecast the Bill Before It Arrives

Awareness tells you where you are. Planning tells you where you are headed. This is where the coming bill stops being a surprise and becomes a number on a forecast.

Because the platform captures real cost and consumption per request, it can project forward with confidence. Knowing what a capability costs today and how fast its usage is growing, you can model what it will cost at two, five, or ten times the volume—and what happens when this quarter’s promotional rate reverts to standard pricing next year. That is the scenario most organizations are walking into blind. Holokai lets you put it on a slide before it puts you in a budget crisis.

It changes the conversation with finance. Instead of “AI cost is variable and hard to predict,” the answer becomes a defensible forecast: current spend, the growth curve, the impact when discounts expire, and a plan to stay ahead of it. AI moves from an unpredictable expense to a managed, modeled line item—which is the prerequisite for ever trusting it with a bigger budget.

FOR THE CFO AND THE BOARD

A forecast that already accounts for expiring promotional rates is exactly the artifact a CFO needs to justify continued AI investment through a budget cycle. It turns “trust us, it is worth it” into a modeled projection that finance and executive leadership can both work from.

CAPABILITY 3

OPTIMIZE · Pay Less for the Same Outcome

This is where Holokai moves from accounting to advantage—and it is the capability most platforms do not have.

Building an AI capability has a natural lifecycle. Early on, the smart move is to build on the best model available: the newest, most capable, and usually the most expensive frontier model. You are not optimizing for cost yet. You are de-risking the work. The premium model gives you headroom to prove the use case, ship it, and get it stable in production without fighting model limitations while you are still fighting the problem itself.

Once a capability is stable, though, paying frontier prices on every request is rarely the right answer. The work is understood, the prompts are settled, and the quality bar is well defined. That is the moment to optimize—to find out whether a cheaper model clears the same bar. You can only know by proving it on your own workload, not a public benchmark. The platform’s Experiments capability is built for that.

Experiment with Side-by-side Model Comparisons

An Experiment takes a representative set of your own tasks—ideally captured from real production traffic—and runs them across a lineup of candidate models and providers at once. Same inputs, same grading, sent to GPT, Claude, Gemini, an open model, whatever you want to weigh. The platform handles the orchestration and hands you one comparison table.

For each model in the lineup, you see side by side identical work: quality (how often it produced an acceptable result, scored consistently), cost (what it cost per task and in total), speed (how fast it responded), and token consumption (how much it used to get there).

This is the table that ends religious debates about model choice. An Experiment can show that a model costing a fraction of the incumbent clears the same quality bar at near-identical rates—which means you can move the workload and cut its bill by 60 to 80 percent with no degradation a user would notice. Just as often, it shows the premium model really is worth it for one critical workflow and overkill for ten others. In both cases you are deciding on evidence rather than habit.

And because every model speaks through the Holokai gateway in the same normalized format, acting on that evidence costs nothing operationally. Moving a workload to the winning model is a configuration change, not a migration—no refactoring agents, no re-coding applications, no retraining users on a new integration. The switch that used to be a quarter-long project becomes a decision you can make the day the data supports it.

Because new models launch constantly, an Experiment is not a one-time exercise. It becomes a benchmark you re-run whenever a cheaper or better option appears, so your model choices keep pace with the market instead of ossifying around whatever you picked first.

CAPABILITY 4

CONTAIN · Keep Cost Under Control as You Scale

Optimization is a moment. Containment is a discipline: keeping cost under control continuously, as usage grows and the organization's AI footprint expands.

Because every request flows through one gateway and is measured against one source of truth, the platform becomes the natural control point for governance. Costs are attributed, so accountability is real. Teams see their own consumption, and spend can be allocated rather than absorbed into one anonymous bill. Trends surface the moment they start, so a runaway feature or a misbehaving integration shows up as an anomaly on a chart instead of a shock on next month's invoice. And because switching models is a configuration change rather than a rebuild, the same gateway that gives you visibility also gives you a lever you can pull immediately—moving a workload to a better-fit model the moment Experiments or a cost anomaly makes the case for it.

There is a shift underway. The discount era rewarded adoption. The era after it will reward efficiency. The organizations that come out ahead are the ones that build the muscle—and the instrumentation—to keep cost-per-outcome flat or falling even as AI usage climbs. Containment is what lets you keep saying yes to new AI work, because each initiative arrives with a known, managed cost instead of an open-ended risk.

FOR THE CFO

The control point that contains cost is the same one that enforces policy governance. Routing every exchange through a single governed plane means data-handling rules, logging, and access controls apply uniformly, and the audit trail needed to protect IP, PII, and sensitive data is produced in the course of normal operation rather than as a separate compliance project.

PUTTING IT TOGETHER

What This Looks Like in Practice

Picture an organization six months after adopting Holokai.

The head of engineering opens a dashboard and sees AI spend broken out by feature, model, and team—updated in real time and accurate to the day's true rates. Finance has a forecast that already accounts for promotional rates expiring next year, so nothing is lurking. When a new model launches, a product owner runs last month's real traffic through an Experiment overnight and shows up the next morning with a table proving it matches quality at 40% of the cost; the migration gets approved on evidence, not opinion. When one team's usage spikes, it surfaces as an anomaly that same week, gets investigated, and gets fixed before it ever reaches the invoice.

That organization is not bracing for the AI bill. It has already seen it, planned for it, optimized against it, and put guardrails around it. While peers freeze AI budgets in a panic, this one expands its budget with confidence—because every dollar is measured, justified, and working.

LOOKING OVER THE HORIZON

From Token Management to Efficient AI Operations

See, Plan, Optimize, and Contain solve the cost problem you have today. But the economics of AI are about to shift again—and the next wave of cost will not come from the models themselves. It will come from how organizations use them.

The most expensive pattern in enterprise AI is not the price-per-token on a frontier model. It is the architecture that keeps a model in the loop on every step of every workflow. Today's agent frameworks let a model reason, pick a tool, read the result, reason again—and each turn re-sends a growing context window you pay for. The orchestration itself becomes the dominant token cost, and it scales with the number of steps rather than the amount of real work being done.

Holokai's roadmap addresses this directly with three capabilities that move cost management from the token level to the operational level:

Deterministic workflows. Instead of letting a model improvise each step, BigBrain—Holokai's agentic workflow engine—compiles a user's goal into an explicit, structured plan that executes without consulting a model to decide what happens next. Models are called only where generating something is the actual job. The thousands of control-flow decisions a deterministic graph makes for free are no longer billable events.

Intelligent routing. Not every request deserves a frontier model, and today that judgment call is manual—made by a person, informed by Experiments. Holokai's roadmap moves that judgment into the

gateway itself: rule-based and dynamic routing that sends simple, high-volume tasks to internal or lower-cost models automatically, reserving frontier spend for the work that actually needs it. A simple lookup never touches the premium meter. A nuanced analysis gets the model it needs—without a person in the loop deciding which is which."

Prompt and workflow intelligence. Before a request ever reaches a model, the Holokai Desktop Application will be able to suggest a more effective prompt, recommend an existing workflow a colleague already built and validated, or point to a curated Workspace with the right data and policies already in place. Each intervention is small. Multiplied across every employee, every day, they represent one of the largest cost levers an organization has.

These capabilities share a common principle: spend tokens on intelligence where intelligence is required, and on nothing else. That is cheaper, more predictable, and more auditable—which is exactly what the governance story needs as well. Efficiency and control are not trade-offs. They are the same design decision viewed from two angles.

COMPANION PAPERS

Holokai publishes companion papers on BigBrain (enterprise-aware agentic workflows) and Shield (distributed governance for enterprise AI) for readers who want the full architecture behind these capabilities.

THE BOTTOM LINE

The Next Two Years Are an Economics Problem

The defining question of the next two years is not whether you adopted AI. It is whether you can afford to keep scaling it as the discounts disappear and the bills come due.

The organizations that win this phase will be the ones that treated AI cost as a measurement problem and solved it early: that could see every dollar, plan for where spend was heading, optimize toward the best outcome per dollar, and contain cost as an ongoing discipline rather than a quarterly fire drill.

Holokai makes all four possible—across every model and every provider, on a single source of truth, in language a technical leader and a budget owner can both act on. One platform to see it, one control plane to manage it, one source of truth for what is becoming the most consequential new line item on your budget.

The Numbers Behind the Reckoning

The macro data is unambiguous. Global AI compute spending grew **166% year-over-year in Q2 2025 alone**, reaching approximately \$82 billion in a single quarter (IDC, 2025). Enterprise AI budgets are expected to more than double as a share of revenue — from 0.8% in 2025 to **1.7% of revenue in 2026** — and analysts project \$3–4 trillion in cumulative global AI infrastructure capex by the end of the decade (BCG AI Radar, Jan 2026). That is not spending that slows down. That is spending that compounds.

Token costs are falling — but total bills are rising. LLM API prices have dropped roughly **80% from 2025 to 2026** — GPT-4-class capability now available at approximately \$0.40 per million tokens versus \$30 per million in March 2023. That sounds like good news. It is not, for one reason: consumption is growing faster than price is falling. Gartner's March 2026 analysis found that agentic AI workloads require **5–30× more tokens per task** than standard chatbots. A single chatbot API call may cost \$0.001. A multi-step agent that plans, retrieves, reflects, and self-corrects runs \$0.10 to \$1.00 per task completion — a 100× to 1,000× multiplier. At production scale, that compounds into monthly infrastructure bills in the tens of millions for Fortune 500 firms.

The hardware layer is shifting the equation further. NVIDIA's Blackwell architecture delivers roughly **15× lower cost per token** than the previous Hopper generation and more than 10× better tokens-per-watt — meaning the inference market is about to absorb dramatically more capacity at dramatically lower per-unit cost (SemiAnalysis InferenceX, 2026). The inference market, valued at \$103 billion in 2025, is projected to reach **\$255 billion by 2030**. The number of inference providers grew from 27 in early 2025 to approximately 90 by end-2025. More supply and lower unit costs do not reduce your bill if the volume of work grows faster — which, in every enterprise AI program on record, it does.

The forecasting problem is already acute. 80–85% of enterprises miss their AI infrastructure forecasts by more than 25% (Mavvrik/BenchmarkIT State of AI Cost Governance, 2025). Less than 1% of executives report ROI of 20% or greater; 53% report only 1–5% ROI (Forbes Research, 2025). 30% of generative AI projects are abandoned after proof of concept — and Gartner projects that more than **40% of agentic AI projects will be cancelled by the end of 2027** — not because the technology failed, but because the cost basis was never understood. The organizations that survive this phase will not be the ones that spent the most. They will be the ones that measured the most.

The bill is coming. Whether it lands as a shock or as a number you already planned for is up to you.

To see Holokai run against your own workloads—including a live cost breakdown and a model Experiment on your real traffic—contact the Holokai team at holokai.ai