Archera × Cloudelligent

2026

# Optimizing GenAI Workloads in the Cloud

A Joint Cost Optimization Playbook

# The GenAI Cost Problem

GenAI workloads don’t behave like anything your FinOps stack was built for — and the cloud bills are starting to show it.

Archera and Cloudelligent have partnered to give teams flexible, intelligent guardrails — without locking you into commitments you'll regret.



Unpredictable, spiky workloads that resist standard commitment models



Token-based pricing that's opaque and hard to forecast



Rapidly evolving models that make long-term commitments risky



Infrastructure decisions made at design time that drive the entire bill

Archera

# Three Core Challenges

Why traditional FinOps tools fall short for GenAI

Challenge 1

## Spiky Usage Patterns

GenAI inference doesn't follow predictable curves. Batch jobs, user-driven prompts, and agentic workflows create volatile consumption that traditional Reserved Instances and Savings Plans can't accommodate without overpaying.

Challenge 2

## Capacity Prediction Issues

Forecasting capacity blocks for GenAI is inherently difficult. Committing too early means paying for unused resources; committing too late means throttling and degraded performance.

Challenge 3

## Provisioned Throughput Complexity

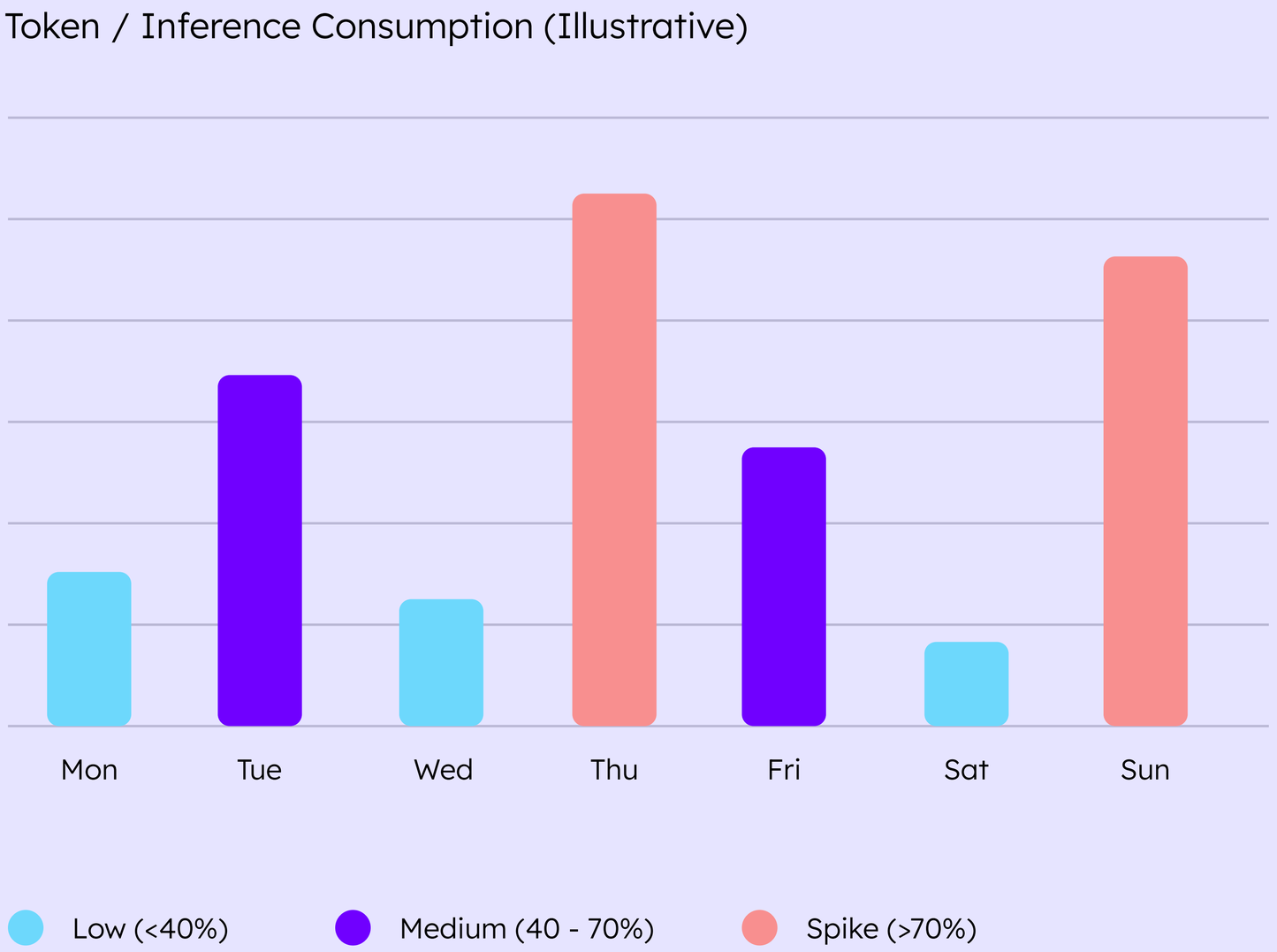
Scaling needs are uncertain, SLA requirements are strict, and new model versions emerge constantly — making provisioned throughput commitments a moving target with real financial consequences.

# Spiky Usage Patterns

**When usage spikes without warning, flat-rate commitments break**

GenAI workloads spike unpredictably — and every spike you didn’t plan for becomes a cost you can’t explain.

Traditional RI/SP commitments are designed for flat, predictable workloads — not this.




Archera

# Capacity Prediction Issues

## Forecasting Is Guesswork

GenAI capacity needs vary wildly based on prompt complexity, context length, and concurrent users. Historical data rarely predicts future demand accurately.

## Unused Capacity = Wasted Spend

Organizations paying for provisioned capacity they don't use can see 30–50% waste in early-stage GenAI deployments.

## The Commitment Trap

Long-term commitments lock teams into specific model versions and instance types at the exact moment the technology is evolving fastest.

## Archera's Answer

Short-duration, flexible guarantees that let you capture discounts without multi-year lock-in — and insurance-backed protection if usage doesn't materialize.

Archera

# Provisioned Throughput Complexity

## Scaling Uncertainty

Agentic systems and multi-model pipelines make it nearly impossible to predict throughput requirements at commitment time.

## SLA Pressure

Teams need consistent latency and throughput guarantees — but over-provisioning to meet SLAs creates significant cost overhead.

## Rapid Model Evolution

New model versions frequently change cost-performance profiles, making yesterday's commitment a poor fit for tomorrow's workload.

## Cloudelligent's Approach

Continuous model re-evaluation and architecture reviews ensure your provisioned throughput commitments remain aligned as the landscape shifts.

# The Joint Solution

**Spend less. Ship faster. Never get locked in.**

## Archera

- Flexible, short-duration compute guarantees
- Insurance-backed commitments (pay only for what you use)
- Automated rightsizing and commitment optimization
- Cost avoidance without long-term lock-in

+

## Cloudelligent

- FinOps-first GenAI architecture balances innovation with cost-consciousness
- Model selection & routing frameworks
- Custom observability dashboards
- Ongoing model re-evaluation as tech evolves

# Cost Reduction for FinOps Teams

Align every dollar of spend to actual usage.

01

## Usage-to-Cost Alignment

Archera's platform surfaces token consumption, inference costs, and infrastructure spend in a unified view — so FinOps knows exactly what's driving the bill.

02

## Optimized Effective Rates

Capture commitment discounts on the resources you're confident about, while Archera's flexible guarantees cover the unpredictable tail.

03

## Cloudelligent's FinOps Program

Free-of-cost FinOps program included with every engagement — baked into your project from day one, not bolted on after a surprise bill.

04

## Automate Tagging

Cost attribution starts with tagging. Cloudelligent embeds automated tagging policies so your cost data is always clean and actionable.

Archera

# Cost Avoidance for Finance

Commitments that come with a safety net.

### Insurance-Backed Guarantees

Archera's unique model acts like a financial insurance policy — if your actual usage doesn't match your commitment, you're reimbursed for the delta.

### No More Commitment Anxiety

Finance teams can approve commitments confidently, knowing there's a backstop against overpayment if forecasts miss.

### POC vs. Production Clarity

Cloudelligent helps teams define clear cost baselines early, so finance can distinguish experimental R&D spend from production scaling costs.

### Predictable Budgeting

Custom Gen AI dashboards give finance real-time visibility into spend trends, so there are no end-of-month surprises.

Archera

# Simplified Management for Engineering

Build fast. Stay in budget.
No friction with finance.

01

## Clear Resource Usage Tracking

rchera and Cloudviz provide engineering teams with granular visibility into what's consuming compute — down to the model, endpoint, and workload.

02

## Prevent Budget Overruns

Alerting and guardrails prevent the runaway spend that kills GenAI projects. Engineering ships features; finance doesn't shut them down.

03

## Model Selection Frameworks

Cloudelligent's structured frameworks help engineering choose the right model for each use case — balancing accuracy, latency, and cost from the start.

04

## Governance Without Bureaucracy

FinOps compliance embedded in CI/CD workflows means cost controls don't slow down development velocity.

Archera

# Architecting for GenAI Efficiency

Design decisions that drive the bill — and the savings

Phase 1

## Design

- Choose optimal model per use case
- Define cost baselines & forecast usage
- Implement model routing for workload optimization

Phase 2

## Deploy

- Minimize token burn & infrastructure costs
- Optimize inference workflows
- Embed FinOps guardrails from day 1

Phase 3

## Govern

- Track latency, throughput, and model performance metrics
- Automate alerting & dashboards
- Continuous model re-evaluation as costs evolve

Phase 4

## Optimize

- Archera flexible commitments + insurance
- Rightsizing recommendations
- Ongoing architecture reviews with Cloudelligent

Archera

# Your Next Steps

Three steps to measurable GenAI cost savings

## See what GenAI cost optimization looks like for your environment.

Book a free 30-minute demo with Archera + Cloudelligent. No contracts. No commitments. Just answers.

archera.io/demo

Free 4-hour consultation available — no long-term contracts required.

Step 1

### Consultation Call (Cloudelligent)

Meet with Cloudelligent's Solutions Architects to align on your goals, cloud setup, and cost priorities.

Step 2

### Tool Onboarding (Archera + Cloudviz)

Gain access to Archera and Cloudviz for commitment flexibility, visibility, and real-time cost insights.

Step 3

### Cost Optimization Assessment (Joint)

Receive a full analysis of usage, rates, tagging, and GenAI architecture to uncover savings opportunities

Archera



Your GenAI costs are solvable. Archera and Cloudelligent are ready to show you how — book your free assessment today.

Archera: archera.io
Cloudelligent: cloudelligent.com