WSO2CON NORTH AMERICA

May 20 - 22, 2026 | Austin, Texas, USA

AI Cost Attribution & Optimization for LLM Consumption

Anusha Jayasundara

Senior Technical Lead

Is AI still Experimental?

AI adoption is past the experimental threshold, and spend is scaling faster than cost governance.

88%

of organizations using AI in 2025

↗ up from 78% in 2024

3.2x

growth in enterprise GenAI spend

↗ $11.5B → $37B in one year

$2.59T

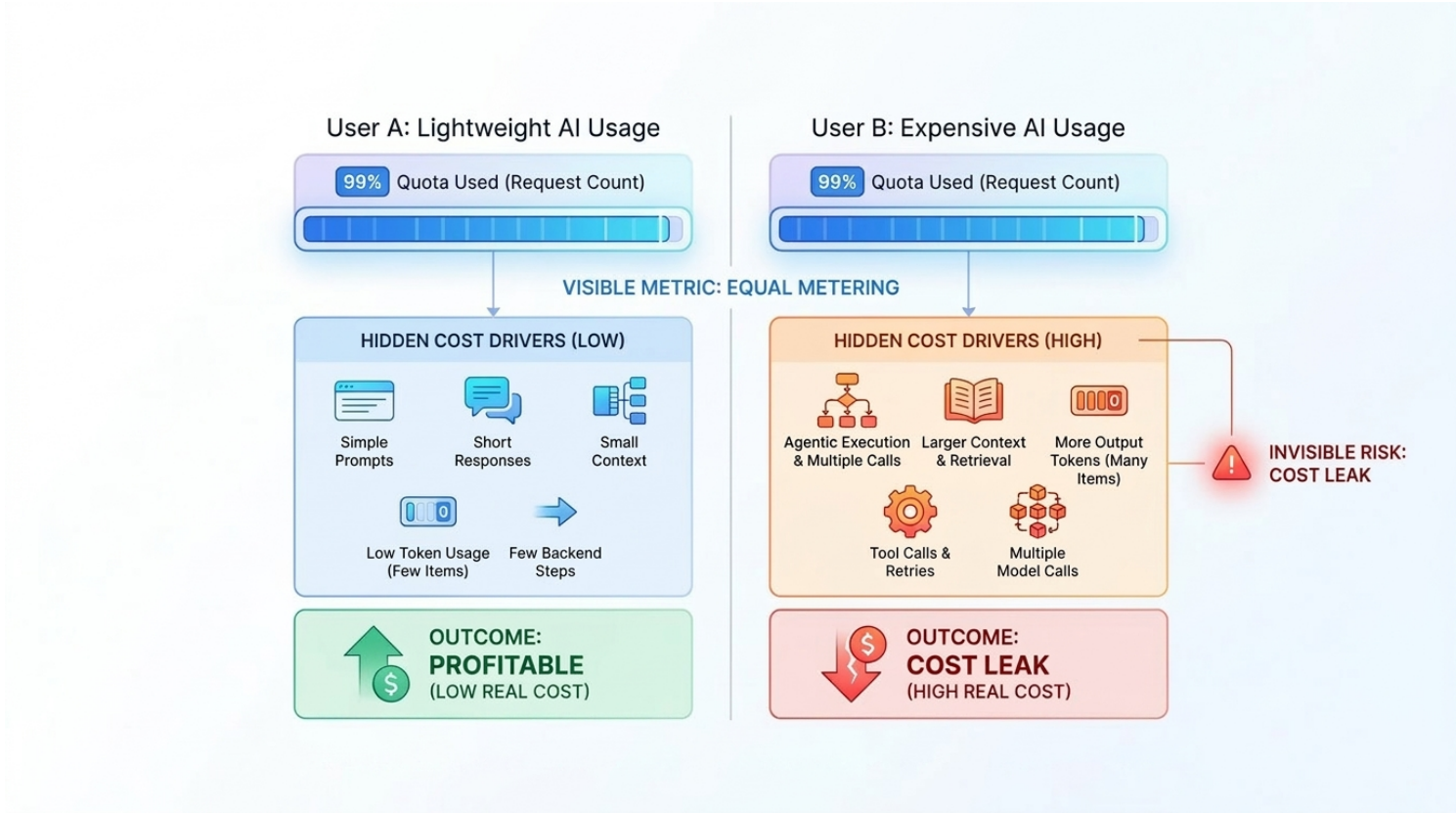
forecast AI spending in 2026

↗ +47% YoY

AI Spending is Scaling, Visibility isn't

- Team Ships AI features fast
- Bill arrives later,
- No one knows whose it is?

Same plan. Same throttle. Opposite outcomes.



That was Cursor

WHAT HAPPENED NEXT

JUN 2025

Pricing redesigned mid-stream

Switched from request-based limits to dollar-denominated credit pool. Cost finally entered the throttle, but users were caught off guard.

JUL 2025

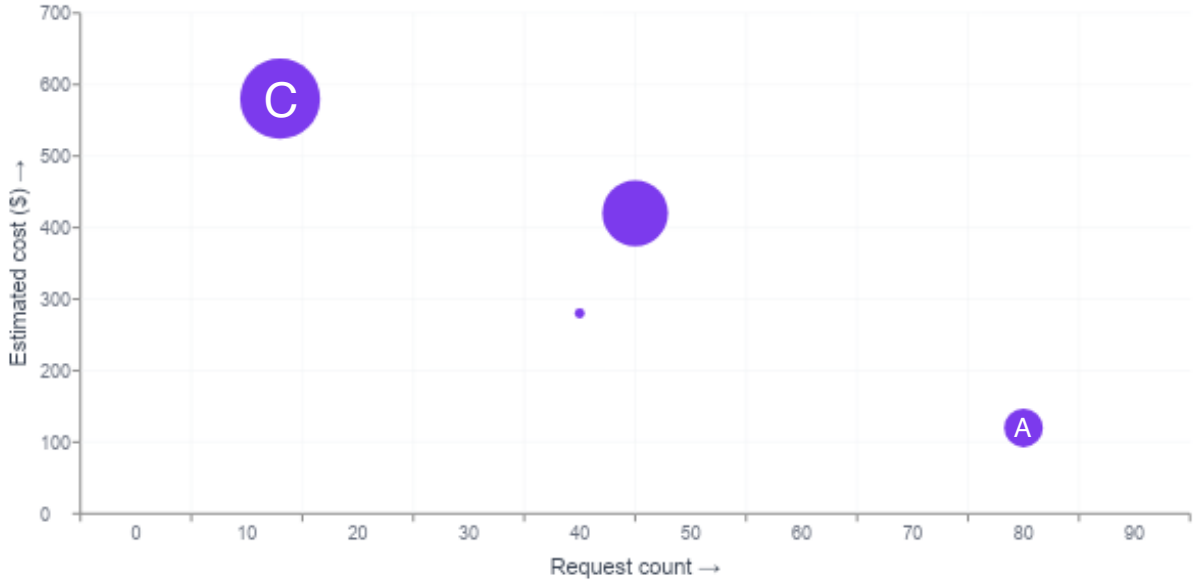
CEO issued public apology

"These changes hurt trust." The company acknowledged pricing changes were not communicated clearly.

Throttling by request count is blind to 100x cost variance per request. You cannot price, route, or govern what you cannot attribute.

The biggest user is not always the most expensive

Traditional API analytics may point to the wrong optimization target.



CUSTOMER PROFILES

- **A** High requests, low cost
- **B** Medium req., high cost
- **C** Low req., very high cost
- **D** Medium req., medium cost

Bubble size = business value

REST API vs AI agent: a tale of two flows

One request in, one response out. The same surface — very different internals.

TRADITIONAL REST API

Client sends HTTP request
GET /products?id=42

Route + auth check
Validate token, match route handler

Single DB query
SELECT * FROM products WHERE id = 42

Return JSON response
Predictable. Fixed compute. ~$0.000

4 steps • 0 LLM call: none • Cost: **~fractions of a cent**
Cost is fixed and deterministic per request

AI AGENT REQUEST

Setup: load prompt + RAG retrieval
Embedding call + planning LLM ($0.034)

Reasoning LLM call (18K tokens)
Full codebase in context — 73% of total cost ($0.270)

Tool calls (read / write file)
No LLM cost — but expands context window for next call

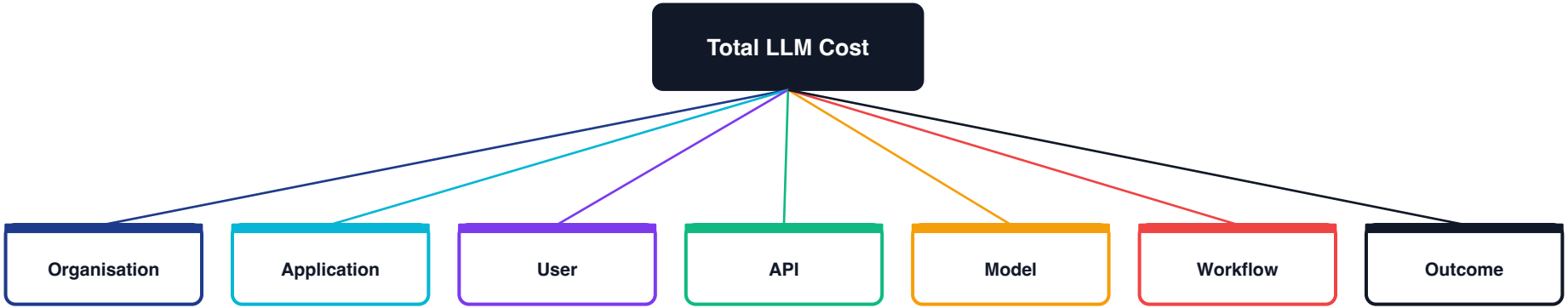
Safety + verify + retry if needed
Each retry = another full LLM call ($0.066)

4 internal events • 3 LLM calls • Cost: **$0.37 (~100x more)**
Cost is variable, scales with context size & steps

REST cost = compute. AI agent cost = compute + tokens + context size + steps + retries. Request-count throttling misses all of that.

Cost is shaped by behavior, not just traffic volume.

Who caused the cost, and why?



Cost attribution connects technical usage to business ownership.

One intent. Many billable events

TRADITIONAL FLOW

AGENTIC FLOW

1 visible request → many billable events

23% scaling agentic AI

39% experimenting with agents

40%+ of agent projects may be scrapped by 2027

Sources: McKinsey State of AI 2025 (adoption); Gartner via Reuters 2025 (risk).

Reduce waste without blocking value

QUICK WINS

- Response caching
- Cost alerts
- Retry controls
- Model routing
- Token limits

STRATEGIC INVESTMENTS

- Prompt optimization
- Quota policies
- Tool-call limits
- Workflow budgets

From AI API exposure to AI API economics

1

Expose

managed AI APIs

2

Measure

tokens, models, workflows

3

Attribute

users, apps, tenants, outcomes

4

Optimize

routing, limits, caching

5

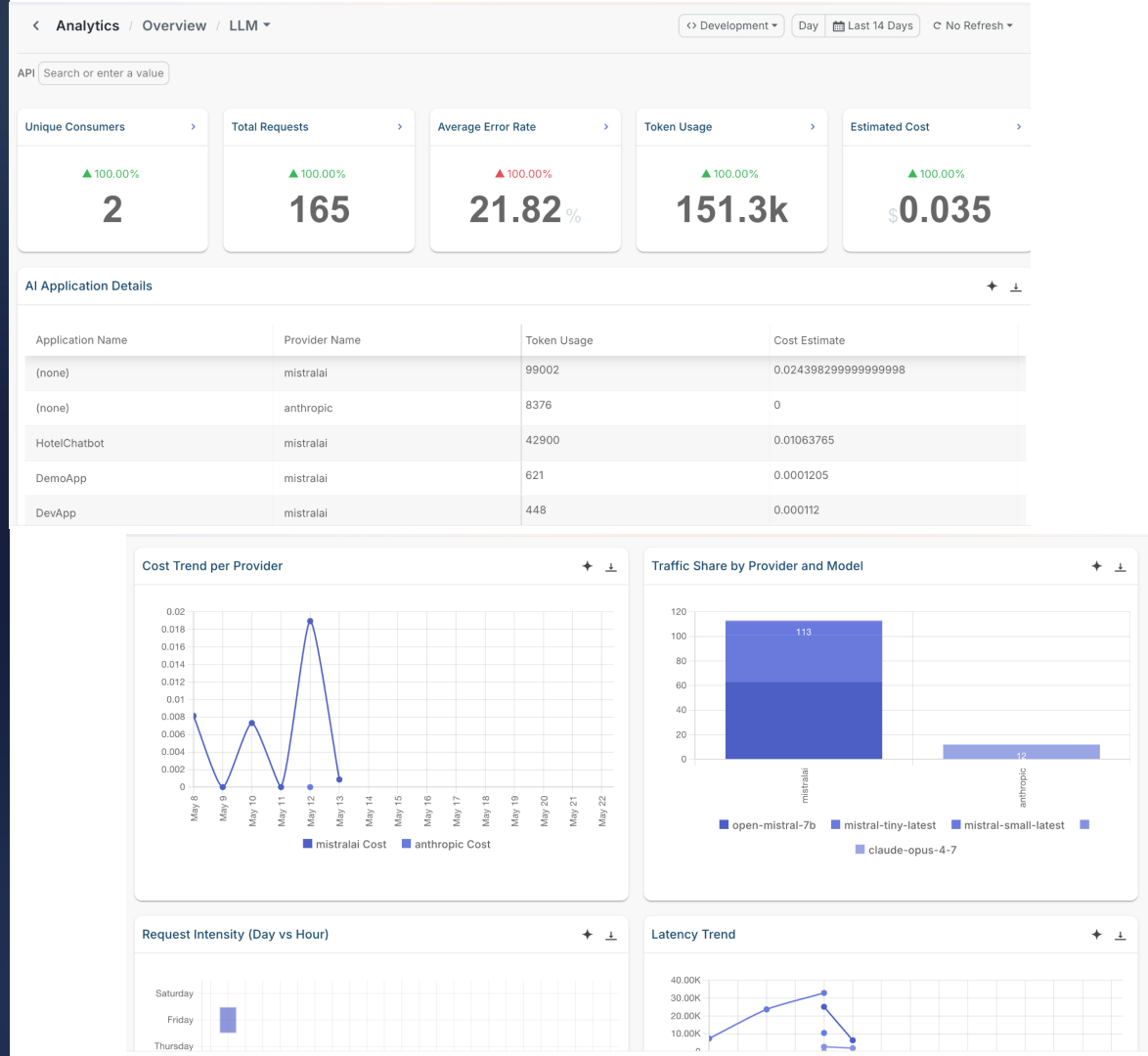
Operationalize

Dashboards, alerts, Reviews

Visibility first, then control, then continuous optimization.

API Analytics Solution

moesif



Five things to take with you

1. LLM cost is highly variable and often hidden inside the execution path.
2. Request count alone is not enough to manage AI API consumption.
3. AI-specific metering must cover tokens, models, context, retries, tools, retrieval, and agent steps.
4. Cost attribution exposes who and what is actually driving spend.
5. Governance and optimization are what make AI APIs economically sustainable.

CLOSING THOUGHT

AI cost optimization starts with visibility.

The winning AI platforms will not simply expose LLMs through APIs.
They will measure, govern, optimize, and monetize AI consumption with discipline.

WSO2

WSO2CON NORTH AMERICA

May 20 - 22, 2026 | Austin, Texas, USA

Thank You!

Anusha Jayasundara

Senior Technical Lead