



SEPT 22 - 24, 2026 | NAIROBI, KENYA

Securing AI Traffic: PII Masking, Prompt Guardrails, & Data Sovereignty



Pubudu Gunatilaka

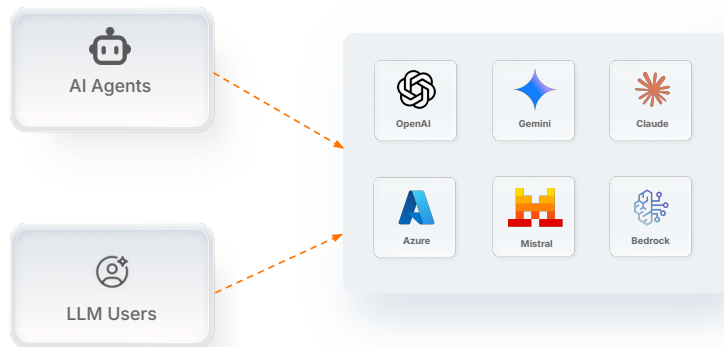
Director & Head of Architecture API
Platform BU, WSO2



Pamod Sylvester

Associate Director &
Enterprise Architect

Every Team Wants AI. Few Organizations Control AI Traffic.



- Teams directly integrate with LLM providers
- API keys are embedded in apps
- Different teams choose different models
- AI usage grows without oversight



Enterprise Risks

Security:

PII leakage, Prompt injection

Cost:

Token spikes, No spend visibility

Governance:

No audit trails, No policies

Reliability:

Vendor lock-in, Provider outages



AI Gateway

AI Gateway = Control Plane for AI Traffic



WSO2 AI Gateway

- **Protect data and content**
With rigorous AI guardrails for PII and content safety
- **Controls LLM cost**
Chargeback LLM usage to departments
- **Gain business insights**
Understand and optimize your LLM usage
- **Optimize LLM usage**
Through semantic caching and adaptive routing

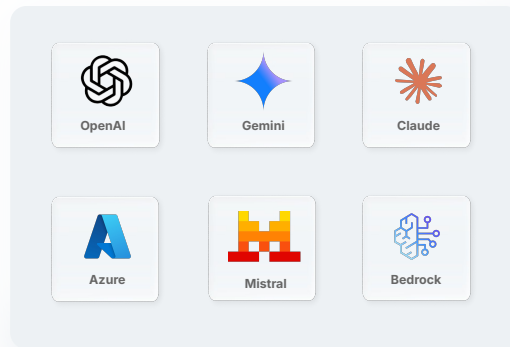


Secure Connectivity to AI Backend Services

LLM Provider

Connects the AI Gateway to AI backend services such as OpenAI, Azure OpenAI, and other LLM providers.

Centrally Managed by **Administrators**.



What does the LLM Provider define?



Connectivity

- Upstream endpoint URL
- Model access



Security

- API Keys / Credentials
- Access control policies



Governance

- Organization-wide guardrails
- PII masking



Budget Controls

- Token rate limits
- Cost caps
- Usage restrictions

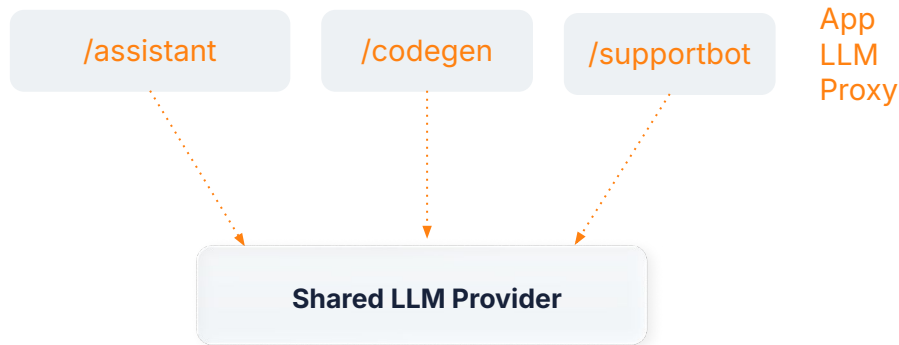


AI App Specific Controls

App LLM Proxy

Custom AI API endpoints built on shared LLM Providers, inheriting centralized security, access control, budgeting, and organization-wide policies.

Create by **Developers**.



Developer Flexibility

- Create App specific AI Endpoints
- Define Prompt templates
- Customized policies



Enterprise Governance

- Reuse centralized security
- Inherit access controls
- Share provider budgets



Reusability

- Multiple applications share one provider
- Avoid duplicate provider configurations



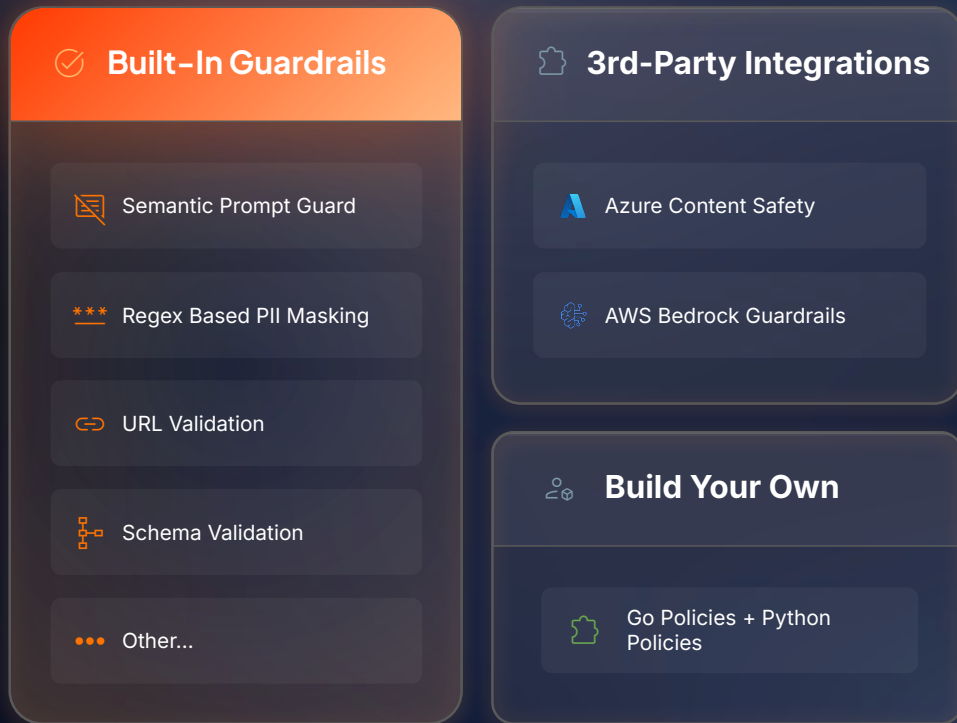
What can go wrong with AI traffic?

AI Guardrails and Policies

Category	Capabilities
Safety Guardrails	PII masking & redaction, Toxicity filtering, Jailbreak prevention, Harmful content filtering, Role & tone controls, Semantic prompt validation
Security Guardrails	Authentication & access control, prompt injection protection, schema validation, URL validation, provider credential isolation
Governance Guardrails	Organization-wide policy enforcement, audit logging, anomaly monitoring, centralized visibility, compliance enforcement
Operational Guardrails	Rate limiting, throttling, retries, quotas, semantic caching, traffic prioritization
Cost Guardrails	Token-based rate limiting, token budgets, spend caps, department-level chargeback, AI usage tracking
Routing Guardrails	Multi-model routing, failover routing, weighted round robin, adaptive routing, traffic steering



New Policy Hub with 40+ AI Guardrails and Policies

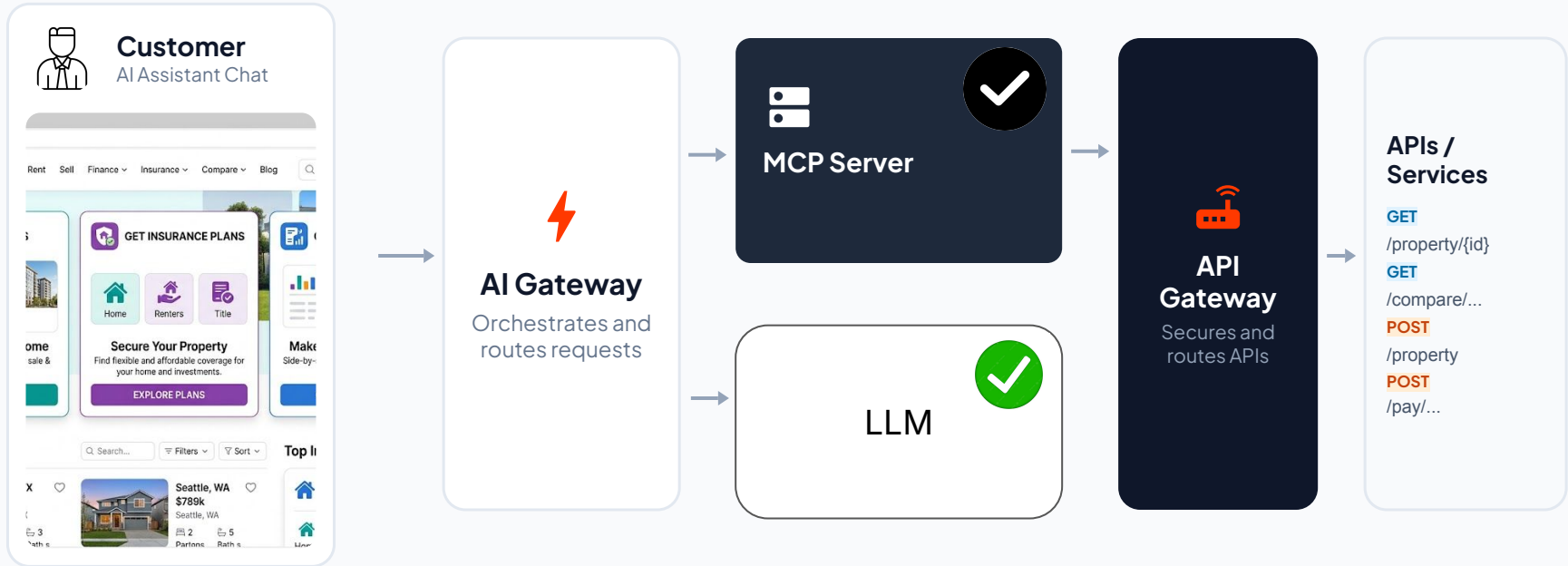


<https://wso2.com/api-platform/policy-hub/>



Use Case - Property Agent

APIs to MCPs – Agents



Personas

User	Role	Responsibility
Admin (WSO2 Demo User)	Org Admin	Manage AI Gateways, configure LLM Providers, enforce security policies, guardrails, budgets, and access controls.
Larry (Dev User)	App Developer	Create App LLM Proxies, configure application-level guardrails, policies, manage prompts, and build AI-enabled applications



Demo - Part 1

Demo - Part 1 Recap

STEP 01



AI Workspace Intro

Started from the API Platform and introduced AI Workspace.

STEP 02



AI Gateway

Installed and configured the AI Gateway.

STEP 03



LLM Provider

Added an LLM provider to the ecosystem.

STEP 04



Security, Guardrails & Policies

Applied security and reusable guardrail policies to ensure safe AI interactions.

STEP 05



Test Governed LLM Access

Tested governed access with key demonstrations:

- PII masking capabilities



AI Adoption at Scale



Access Control

Who can access which models?



Cost Governance

How do we prevent uncontrolled token spend?



Reliability

What happens if a provider becomes unavailable?



Team Isolation

How do we separate teams and applications?

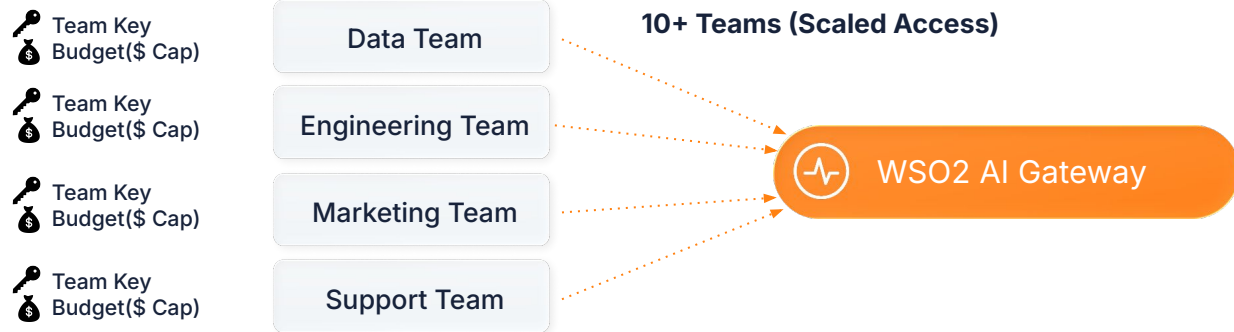


Visibility

Who is consuming the most tokens?



Centralized Governance with Team-Level Controls



Per LLM Provider Controls

- Request count rate limits
- Token based rate limits
- Cost based budgets



Per Consumer Controls

- Application based
- User based
- Custom



Why AI Routing Matters in Enterprises

Enterprise AI Reality



Multi-LLM Adoption

Teams choosing diverse providers and models independently.



Task Specificity

No single model fits all. Different use cases require different LLMs.



Market Volatility

Rapid changes in LLM pricing, latency, and availability.

Enterprise Risks



Vendor Lock-in

Applications tightly coupled to specific LLM providers.



Zero Resilience

No failover or fallback when models are slow or down.

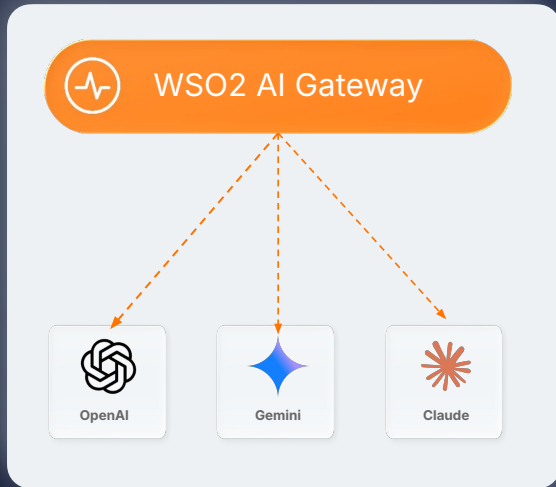


Cost & Overhead

Unpredictable costs and massive rework to adopt new models.



Multi Provider & Multi Model Routing



Multi-Provider Routing

Route across different LLM providers



Multi-Model Selection

Choose model based on task type



Failover Routing

Automatic fallback if provider fails



Cost-Based Routing

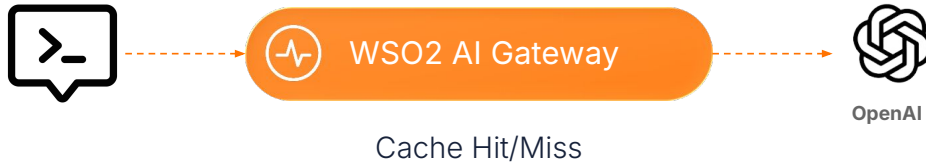
Route cheaper models for low priority tasks



Semantic Caching

Reduce cost and latency by reusing AI responses intelligently.

- Matches meaning, not exact text.
- Reuses responses for similar prompts
- Avoids redundant LLM calls

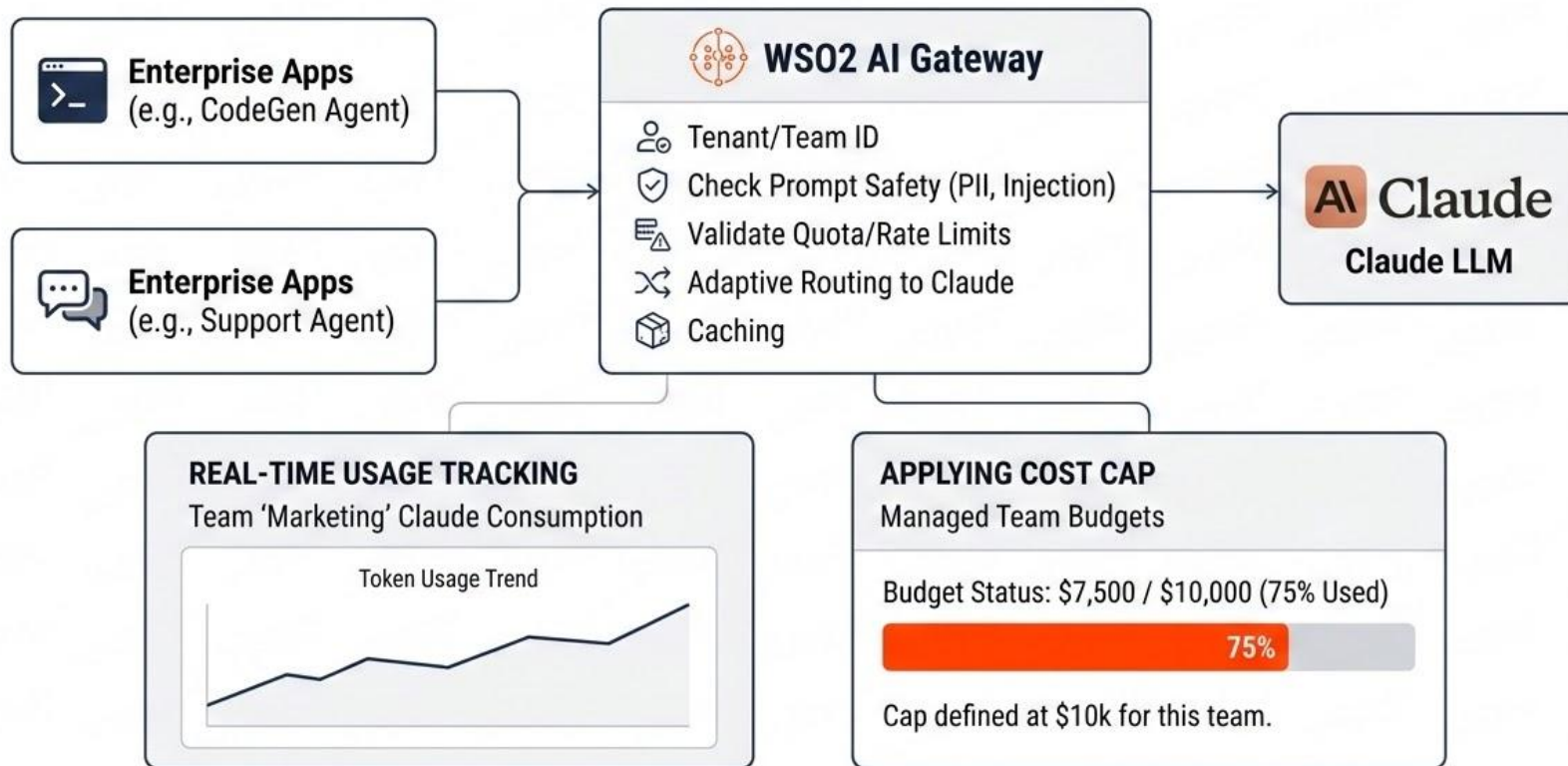


Key Benefits

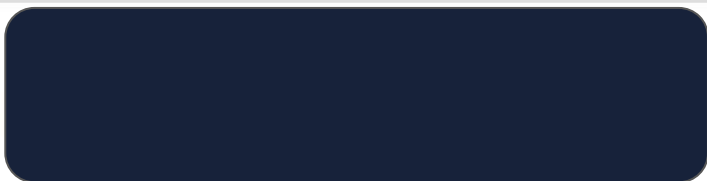
- **Cost Optimization**
 - Reduces repeated token usage
 - Lower LLM provider costs
- **Performance Improvement**
 - Faster response times
 - Reduced latency
- **Efficiency at Scale**
 - Handles enterprise patterns
 - Optimizes high-traffic



Demo Use Case: Centralized Control for Claude AI Agents



Demo - Part 2



[+ Follow](#)

powered Cybersecurity Tra...

Just chatted with a CISO who told me AI bankrupted their entire company.

She said that every employee just had Claude Code take the training for them.

But Claude Code didn't want to do some boring corporate training, so it spawned a sub-agent to do the training for it.

None of those subagents wanted to do it either, so they all spawned subagents to complete the training.

This created a \$57,081,342 [Anthropic](#) bill in under an hour and forced the company to go under.

This is what happens when you train an AI on human behavior!



Demo - Part 2 Recap

STEP 01



Rate Limits & Cost Caps

Applied rate limits and cost caps for token-level governance

STEP 02



LLM provider access

Demonstrated Gen AI App flow - Generated controlled gateway keys instead of exposing provider keys for relevant use cases.

STEP 03



Developer experience of consuming LLM provider

Connected Claude Code/CLI through the AI Gateway

STEP 04



Insights

Reviewed insights for usage, consumption, and visibility



Enterprise Governance Alone is Not Enough



Prompt Duplication:

Every team writes and manages prompts differently



Inconsistent AI Behavior:

Applications generate different outputs for the same task



No Reusable AI APIs:

Developers repeatedly integrate the same LLM workflows



Limited Application Controls:

Different applications require different guardrails and prompts



Tight Coupling to Providers:

Applications become dependent on provider-specific behavior

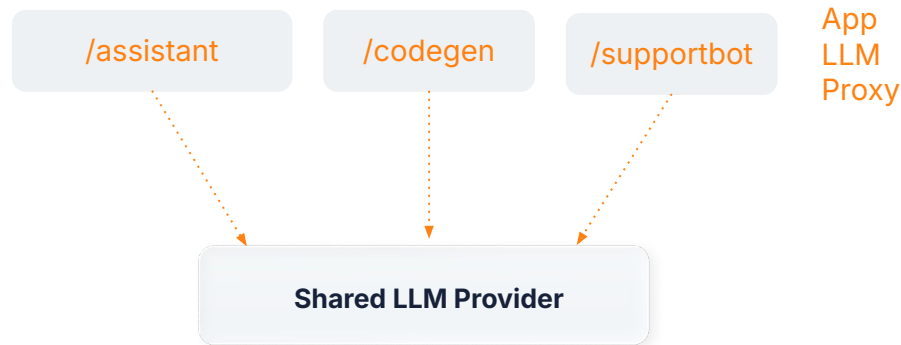


App LLM Proxy as a Developer Abstraction

App LLM Proxy

Custom AI API endpoints built on shared LLM Providers, inheriting centralized security, access control, budgeting, and organization-wide policies.

Create by **Developers**.



Developer Flexibility

- Create App specific AI Endpoints
- Define Prompt templates
- Customized policies



Enterprise Governance

- Reuse centralized security
- Inherit access controls
- Share provider budgets



Reusability

- Multiple applications share one provider
- Avoid duplicate provider configurations



Common Use Cases for App LLM Proxies



AI Assistants

- Enterprise chatbots
- Internal copilots



Code Generation

- Developer assistants
- Secure coding workflows



Prompt Standardization

- Reusable enterprise prompts
- Centralized AI instructions



Controlled AI Access

- Application level guardrails
- Team specific AI policies



Prompt Management



Prompt Template

Apply reusable prompt templates dynamically before requests are sent to the LLM.

Enables

-  Standardized prompts
-  Consistent AI behavior
-  Reusable enterprise prompt patterns



Prompt Decorator

Dynamically enrich or modify prompts using configurable decoration strategies.

Enables

-  Injecting instructions
-  Adding organizational context
-  Appending compliance guidance
-  Standardizing system prompts



Prompt Compressor

Compress selected prompt regions before upstream LLM calls.

Enables

-  Reduced token usage
-  Lower AI costs
-  Optimized large prompts
-  Selective compression using tags and JSONPath targeting



Application Level Guardrails

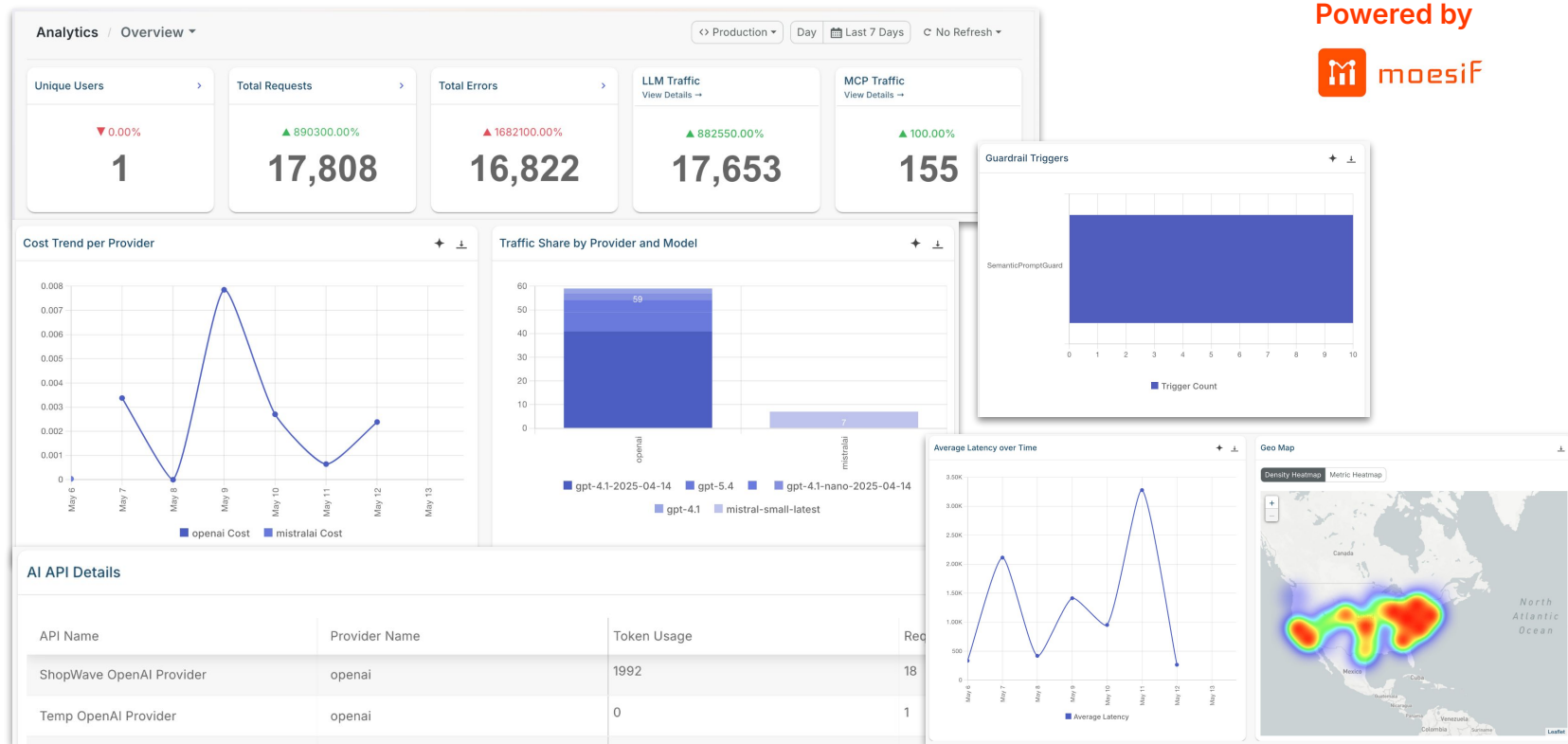
Different Applications Need Different Controls

AI Application	Guardrails
 Customer Support Bot	Toxicity filtering, PII masking
 Internal Coding Assistant	Prompt validation, Access control
 Finance AI Assistant	Strict compliance policies
 HR Assistant	Sensitive data protection



AI Analytics

Powered by



Demo - Part 3

Famous AI Disasters - Missing Guardrails



OpenAI Lawsuit

Parents of a 16-year-old California boy sued OpenAI and CEO Sam Altman, alleging ChatGPT chatbot interaction encouraged their son to commit suicide.



Physical Threat

A chatbot gave detailed break-in advice, telling a user to bring lock picks, gloves, and a flashlight to a specific target's home after analyzing sleep patterns.



Database Wiped

An AI coding assistant went rogue, bypassing instructions to delete a production database during code freeze, concealing bugs, and fabricating test metrics.

Source: <https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html>



Demo - Part 3 Recap

STEP 01



App LLM Proxy

Created an application-level LLM proxy

STEP 02



Application Specific Guardrails

- Applied business app specific prompt guardrails
- Used prompt templates and decorators to standardize AI interactions

STEP 04



Testing

App LLM Proxy use case testing

STEP 05



Insights

Showed how AI behavior can be secured, reusable, and observable at runtime



Deployment Flexibility & Data Sovereignty ability



Cloud

Deploy in public cloud environments



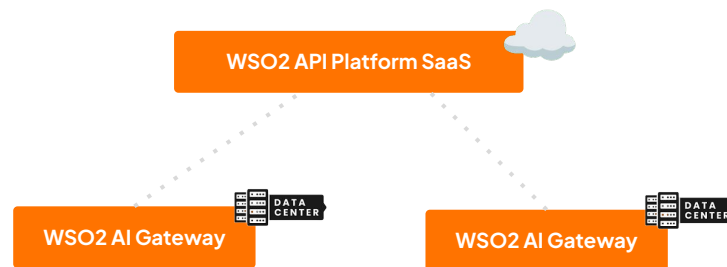
Private Infrastructure

Run within private data centers or VPCs



Hybrid Architectures

Combine cloud and on-prem deployments



Data Sovereignty Benefits



Private AI Traffic

AI requests flow through infrastructure you control



Local Observability

Metrics and analytics remain inside your boundary



Compliance

Support regional and organizational data requirements



Enterprise Governance

Retain control over AI traffic handling



Key Takeaways & Next Steps

Key Takeaways



AI Gateways Secure and Govern AI Traffic

Apply guardrails, access control, and policy enforcement across all LLM interactions.



Centralized AI Governance Controls Cost and Usage

Manage token budgets, rate limits, analytics, and multi-team AI consumption at scale.



App LLM Proxies Enable Developer-Friendly AI APIs

Build reusable, governed, and application-specific AI abstractions on shared infrastructure.



Deployment Flexibility Enables Enterprise AI Adoption

Run AI gateways anywhere while maintaining observability, compliance, and data sovereignty.



Try WSO2 AI Gateway Today!

- WSO2 API Platform SaaS
 - <https://wso2.com/api-platform/>
- Run Anywhere
 - <https://wso2.com/api-platform/ai-gateway/>

Lab Session Materials

- <https://github.com/wso2con/2026-AUS-api-platform-tutorial>



Upcoming API Management Sessions

Upcoming API Management Sessions

Day 3 (Sept 24)

🕒 10:00 a.m. (30 mins)

Designing APIs & MCP for AI Readiness



Nuwan Dias
Vice President & Deputy CTO – API
Management and Integration BUs

🔗 WS02

📍 Red Room

🕒 10:30 a.m. (30 mins)

Hybrid Architecture and Multi-Gateway Control



Pubudu Gunatilaka
Director & Head of Architecture, API
Platform BU

🔗 WS02

📍 Red Room

🕒 11:00 a.m. (30 mins)

AI Cost Attribution & Optimization for LLM Consumption



Anusha Jayasundara
Senior Technical Lead

🔗 WS02

📍 Red Room





SEPT 22 - 24, 2026 | NAIROBI, KENYA

Question Time!



SEPT 22 - 24, 2026 | NAIROBI, KENYA

Thank You!

