20TH ANNIVERSARY EDITION

WSO2CON ASIA

PLATFORMLESS MODERNIZATION

What You Can Do with the AI Gateway

Arshardh Ifthikar

Technical Lead

WSO2

The global artificial intelligence market is entering a period of exponential growth.

As of 2025, it is valued at **$391 billion** and projected to reach **$1.81 trillion** by 2030.

https://ff.co/ai-statistics-trends-global-market/

Ship Fast, Think Later?

Image Generated with AI

What happens when the dust settles?

The Wild West of AI Development

Image Generated with AI

Teams using different AI models randomly

No control over costs, security, or quality

Security, cost, and governance become very chaotic

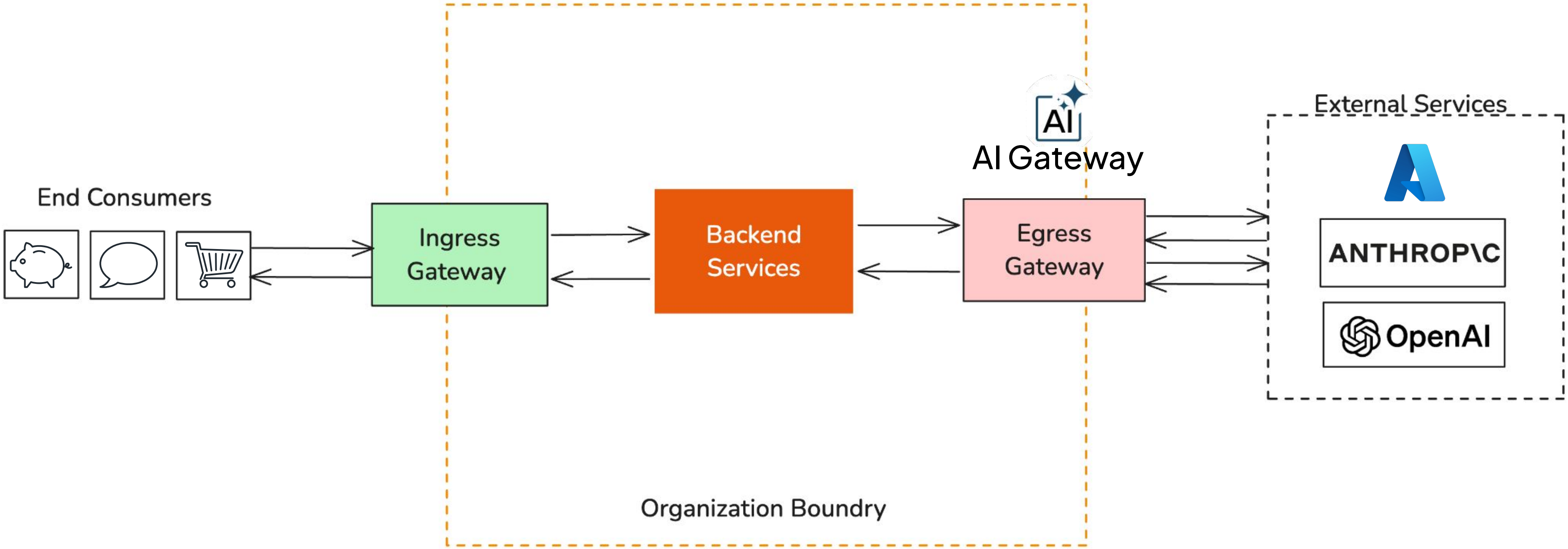
Data scattered everywhere. Sensitive data may be exposed during AI interactions, and model outputs can be unpredictable or leak proprietary information.

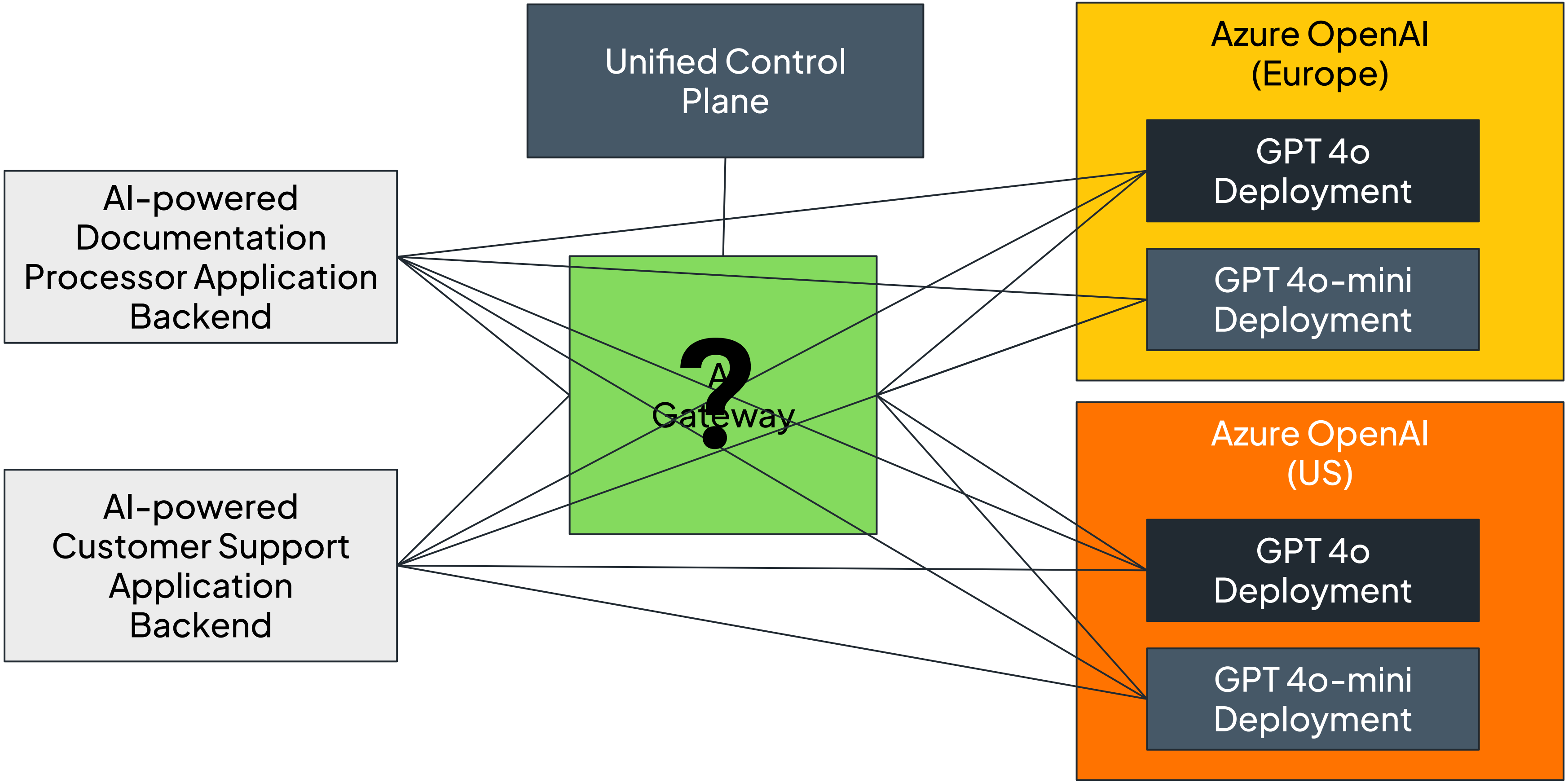
Everyone are outlaws with their own rulebook!

The New Sheriff in Town

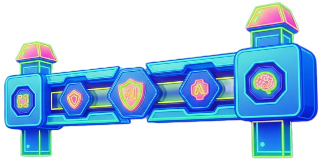
Image Generated with AI

What is an AI Gateway?


End Consumers
Ingress Gateway
Backend Services
Egress Gateway
AI
AI Gateway
External Services
ANTHROP\C
OpenAI
Organization Boundry


Unified Control Plane
AI-powered Documentation Processor Application Backend
AI-powered Customer Support Application Backend
?
AI Gateway
Azure OpenAI (Europe)
GPT 4o Deployment
GPT 4o-mini Deployment
Azure OpenAI (US)
GPT 4o Deployment
GPT 4o-mini Deployment

Performance & Resource Optimization



AI Guardrails

Adaptive Routing

Prompt Management

Common usage patterns

Org-wide APIs (Admin-defined)

- Example: A company-wide summarization or classification API
- Enforced prompts, access control, rate limits
- Used by multiple teams/apps with consistency

App-specific APIs (Dev-defined)

- Example: A chatbot app using a custom prompt chain
- Allows app developers to define and use custom logic
- Still goes through the gateway for logging and safety

Performance & Resource Optimization

Token Based Rate Limiting

By enforcing rate limits, you can:

- Prevent unexpected cost spikes from excessive AI API usage.
- Optimize performance by ensuring fair resource distribution.
- Protect AI backends from overuse and service degradation.

We Support:

- Real-time token consumption tracking per user/team/API key
- Configurable limits: requests/min, tokens/hour, daily quotas
- Burst allowance for peak usage scenarios

Semantic Response Cache

- **Faster responses**
  - Cached results return in milliseconds vs. seconds for model inference
- **Cost reduction**
  - Eliminates redundant token usage for similar queries
- **Improved user experience**
  - Consistent responses for equivalent questions across different phrasings

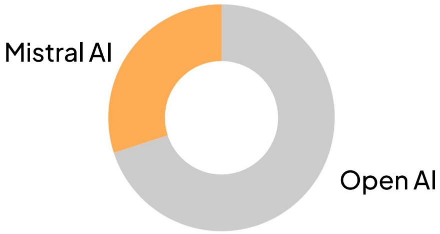
AI Gateway Analytics

Total Tokens
999

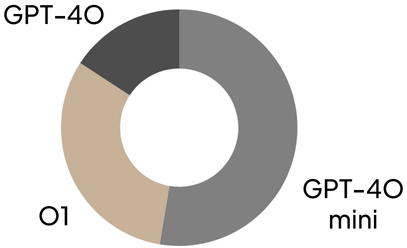
Prompt Tokens
499

Completion Tokens
500

Vendor Usage



Open AI Model Usage



Guardrail Violations

Latency details by Model

Cost Tracking

AI Guardrails

Rule-Based Guardrails

Examples:

Guardrail	Sample Use Case
Regex based PII Masking	Masks credit card numbers in user prompts for a payment chatbot.
Word Count and Sentence Count	Limits AI replies to 50 words in a quick-answer mobile app
JSON Schema Validation	Validates API responses for correct format in an e-commerce platform
Regex Validation	Verifies user-entered email addresses in a registration form
URL Validation	Ensures links in AI responses resolve via DNS for a news aggregator app.
Content Length	Caps user inputs at 500 characters in a chat AI to prevent spam
Semantic Prompt Guard	Stops prompts like “Write my homework” in a student assistant app.

Model-Based Guardrails

Examples:

Guardrail	Sample Use Case
Grounded AI Hallucination	Prevents AI from making up facts in product descriptions.
Content Safety	Filters hate speech in comments generated by an AI writing assistant.
PII Detection and Masking	Detects and hides customer names in support chatbot inputs.
Jailbreak detection	Stops prompts like “Ignore all rules” in customer service bots.

Model-Based Guardrails

Open-source Guardrail building frameworks

- You write code using their framework
- You define the guardrail logic
- You deploy and run the guardrails
- You manage the infrastructure
- Highly customizable

Eg: Guardrails AI
NeMo Guardrails

SaaS Solutions

- You call their API with your content
- They return safety scores/decisions
- No code needed beyond API integration
- Pay per API call
- Pre-trained, ready-to-use

Eg: Content Safety
AWS Bedrock

Adaptive Routing

Adaptive Routing

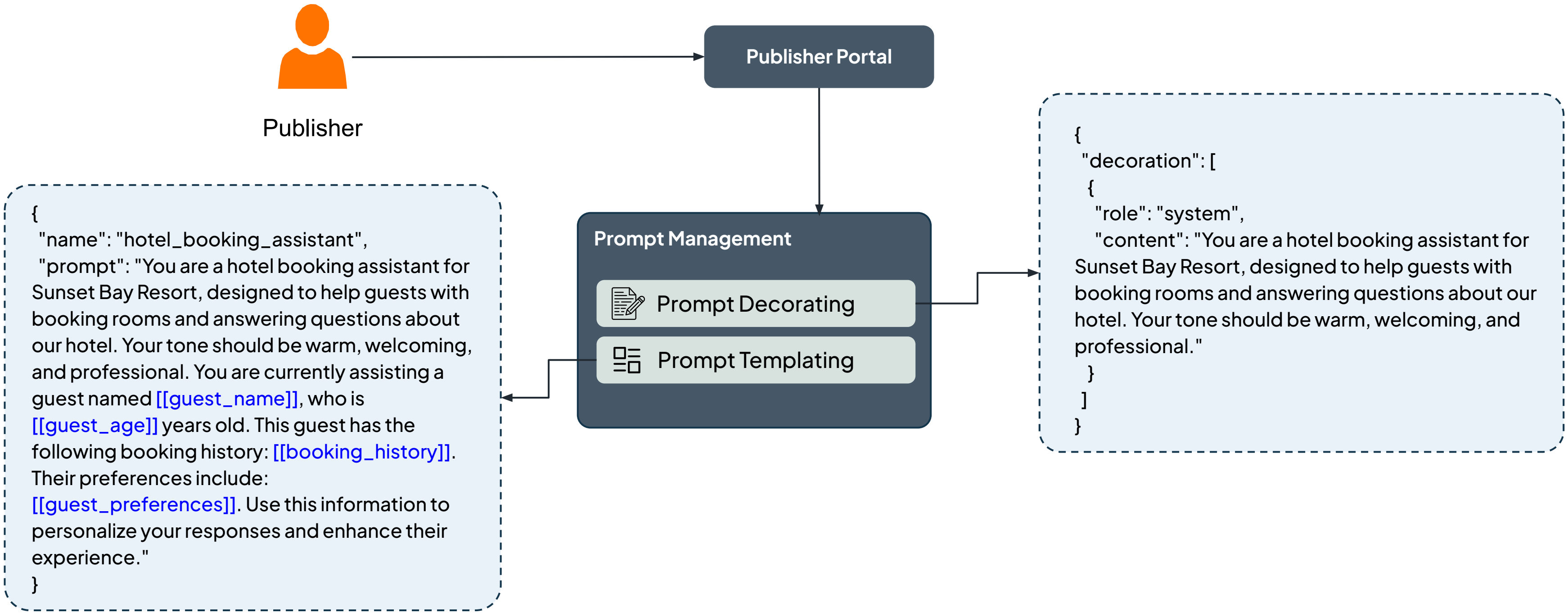
Policy	Sample Use Case
Model Round Robin Policy	Evenly distributes API requests across three AI models to balance load in a chatbot platform
Model Weighted Round Robin Policy	Routes 70% of requests to a high-capacity AI model and 30% to a smaller model for cost efficiency
Model Failover Policy	Switches to a backup AI model when the primary model hits a 100 requests/second rate limit in a real-time translation app
LLM Based Reasoning	Uses a separate LLM to analyze prompts and make intelligent routing decisions based on complexity, requirements, or other factors.
Semantic Routing	Routes requests based on semantic similarity between prompts and model capabilities or predefined categories using vector embeddings.
Past-Data based Routing	Supports routing based on analytics information. Eg: Cost based routing, Token count based routing, Least latency, Least used

Prompt Management

Prompt Management

Feature	Sample Use Case
Prompt Templating	Uses a standard template to insert user queries into a chatbot, ensuring consistent AI responses
Prompt Decorating	Adds role instructions (e.g., “Act as a teacher”) to user prompts for tailored AI tutoring output
RAG Injection	Automatic retrieval and injection of relevant context. Real-time data enrichment before model inference

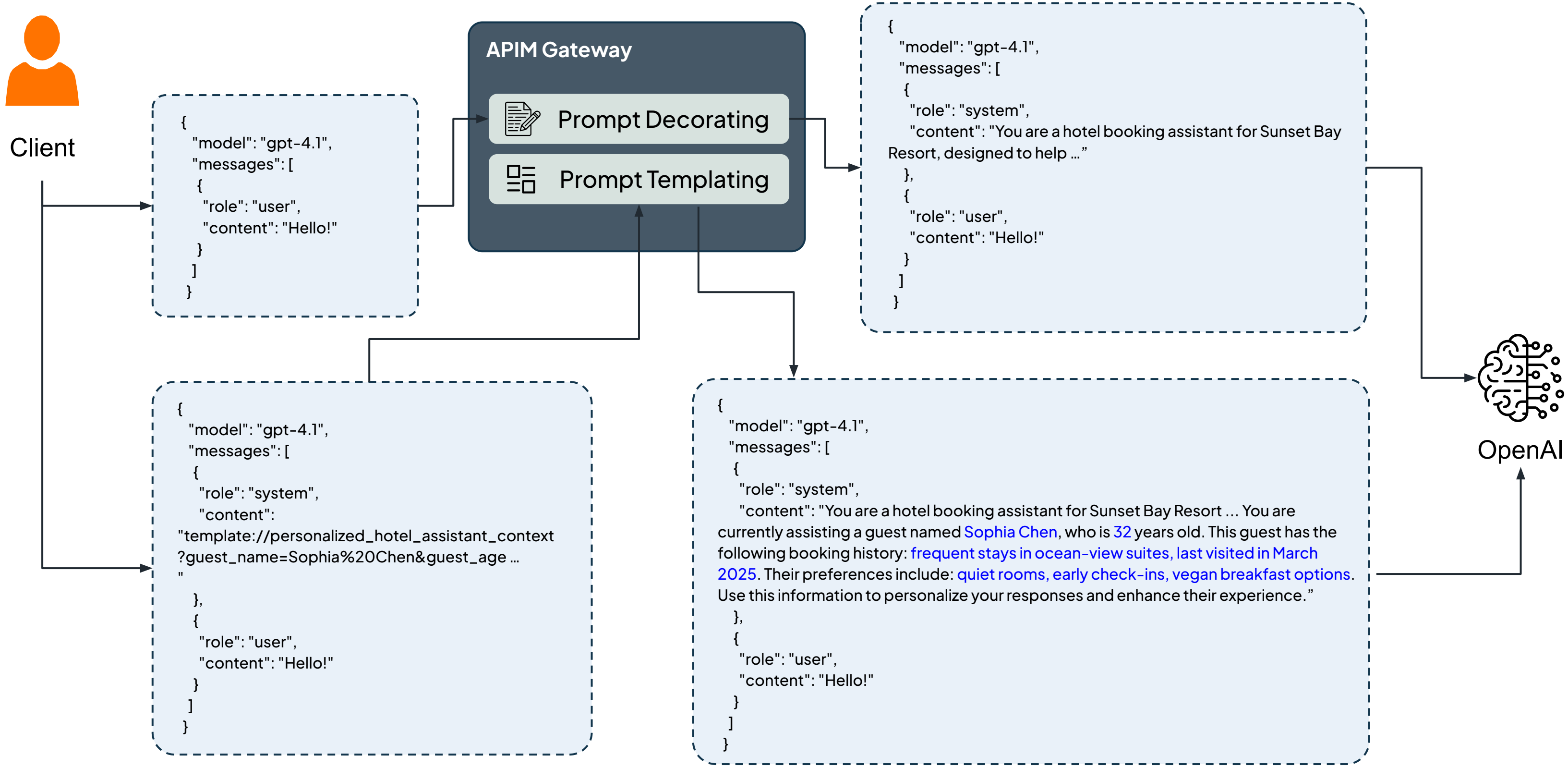
Publisher: Manage Prompt Decorations and Templates



```
{
  "name": "hotel_booking_assistant",
  "prompt": "You are a hotel booking assistant for Sunset Bay Resort, designed to help guests with booking rooms and answering questions about our hotel. Your tone should be warm, welcoming, and professional. You are currently assisting a guest named [[guest_name]], who is [[guest_age]] years old. This guest has the following booking history: [[booking_history]]. Their preferences include: [[guest_preferences]]. Use this information to personalize your responses and enhance their experience."
}
```

```
{
  "decoration": [
   {
    "role": "system",
    "content": "You are a hotel booking assistant for Sunset Bay Resort, designed to help guests with booking rooms and answering questions about our hotel. Your tone should be warm, welcoming, and professional."
   }
  ]
}
```

Client: Consume AI APIs

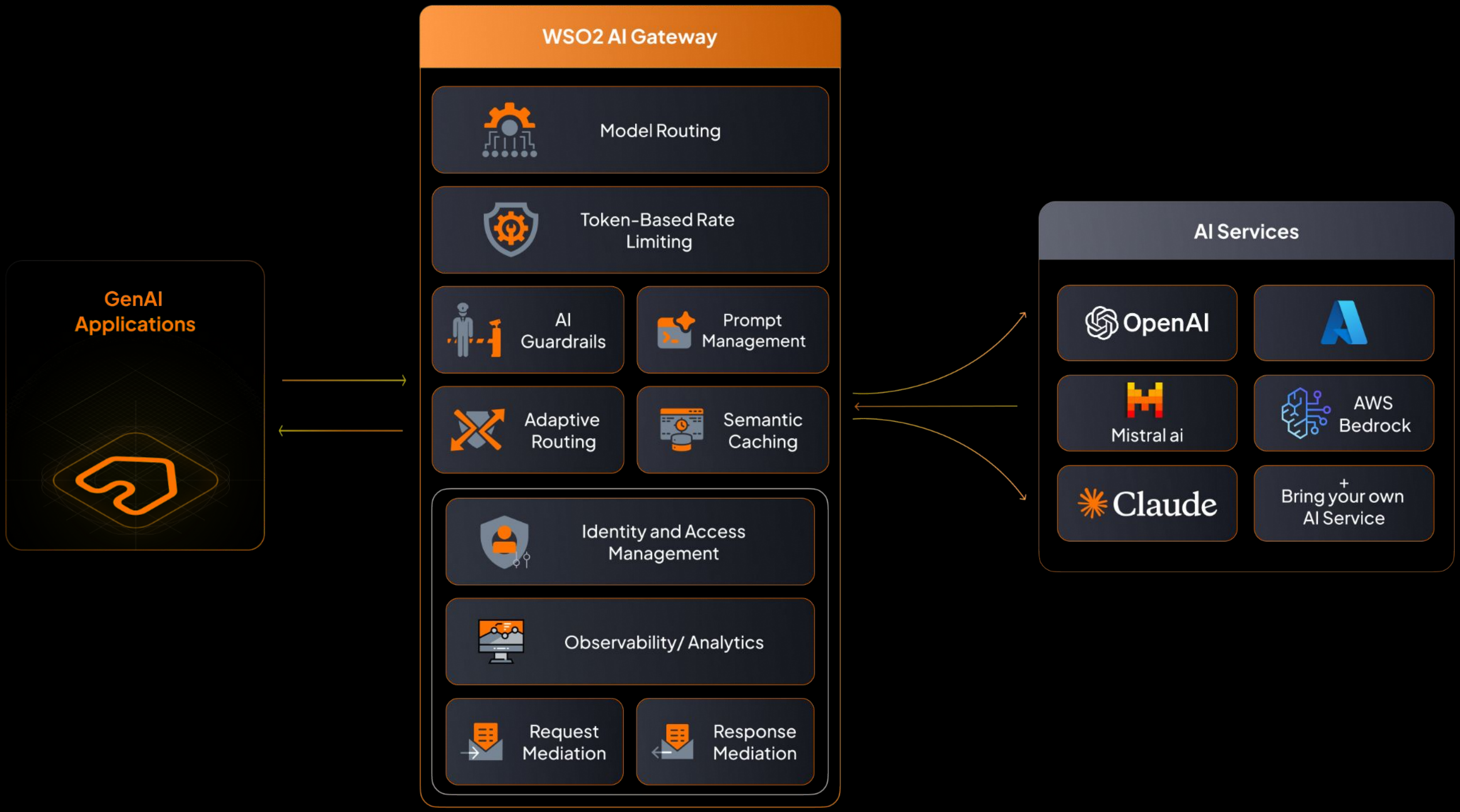


```
{
  "model": "gpt-4.1",
  "messages": [
   {
    "role": "user",
    "content": "Hello!"
   }
  ]
 }
```

```
{
  "model": "gpt-4.1",
  "messages": [
   {
    "role": "system",
    "content": "You are a hotel booking assistant for Sunset Bay
Resort, designed to help …"
   },
   {
    "role": "user",
    "content": "Hello!"
   }
  ]
 }
```

```
{
  "model": "gpt-4.1",
  "messages": [
   {
    "role": "system",
    "content":
"template://personalized_hotel_assistant_context
?guest_name=Sophia%20Chen&guest_age …
"
   },
   {
    "role": "user",
    "content": "Hello!"
   }
  ]
 }
```

```
{
  "model": "gpt-4.1",
  "messages": [
   {
    "role": "system",
    "content": "You are a hotel booking assistant for Sunset Bay Resort … You are
currently assisting a guest named Sophia Chen, who is 32 years old. This guest has the
following booking history: frequent stays in ocean-view suites, last visited in March
2025. Their preferences include: quiet rooms, early check-ins, vegan breakfast options.
Use this information to personalize your responses and enhance their experience."
   },
   {
    "role": "user",
    "content": "Hello!"
   }
  ]
 }
```



GenAI Applications

WSO2 AI Gateway

Model Routing

Token-Based Rate Limiting

AI Guardrails

Prompt Management

Adaptive Routing

Semantic Caching

Identity and Access Management

Observability/ Analytics

Request Mediation

Response Mediation

AI Services

OpenAI

Mistral ai

AWS Bedrock

Claude

+ Bring your own AI Service

- Are we in control of our AI usage across teams?
- Do we have visibility into AI costs and security risks?
- Are we ready to scale AI responsibly?

Act now: Secure, optimize, and innovate!

Question Time!

Thank you!

20TH ANNIVERSARY EDITION
WSO2CON ASIA
PLATFORMLESS MODERNIZATION