WSO2CON 2025

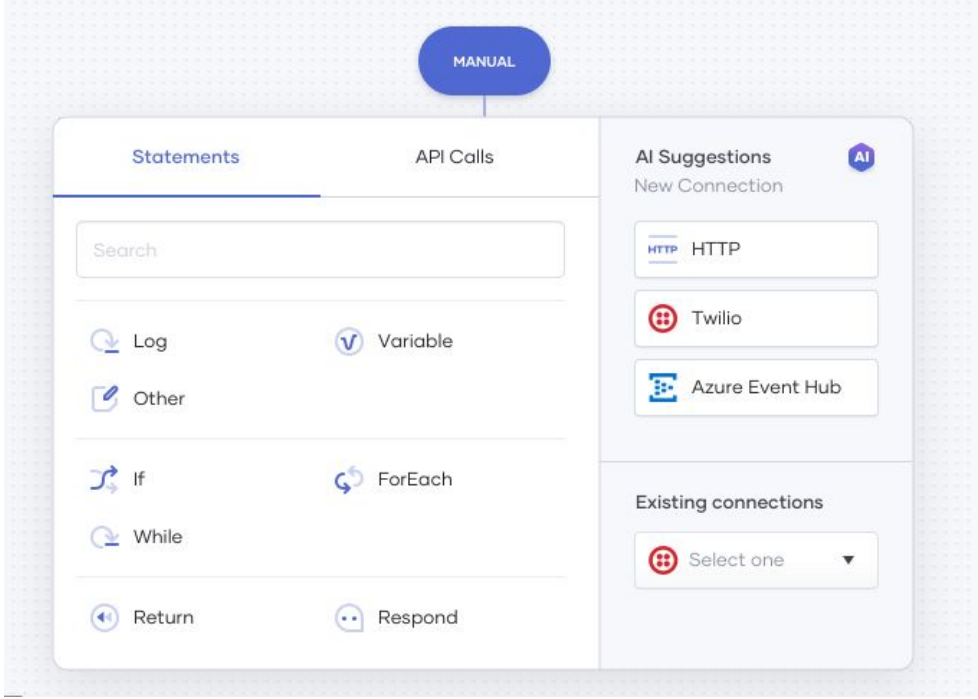
From PoC to Production: AI at Enterprise Scale



Name: Malith Jayasinghe, Ph.D.
Designation: VP - Research & AI
Company: WSO2

AI Development: Then

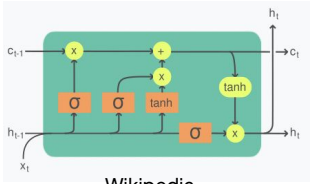
- Use case
  - **Suggest next component in low-code editor**
- Prediction target
  - Connector, Connector operation, Statements, etc
- Technology
  - This low-code platform was using Ballerina underneath



AI Development: Then

Data Set Preparation	Pre-processing	Data Labeling	Model Training	Testing & Iteration
Ballerina Programs	Generate **AST**s, clean data, and map to a visual-level abstraction	**Current Context:** Sequence of elements up to the current node **Next Element:** The next component	Train models (e.g. LSTM)	Multiple refinement cycles Expand the dataset Re-train models

```
main.bal > ...
1   import ballerina/http;
2
3   listener http:Listener httpDefaultListener = http:getDefaultListener();
4
    Run | Debug | Try It | Visualize
5   service / on httpDefaultListener {
6
        Visualize
7       resource function get weather/[string cityName]() returns WeatherInfo|error {
8           // Note: This is mock data. In a real implementation,
9           // you would integrate with a weather API
10          if cityName == "" {
11              return error("City name cannot be empty");
12          }
13
14          // Mock weather data
15          return {
16              cityName: cityName,
17              temperature: 25.5,
18              humidity: 65.0
19          };
20      }
21  }
```

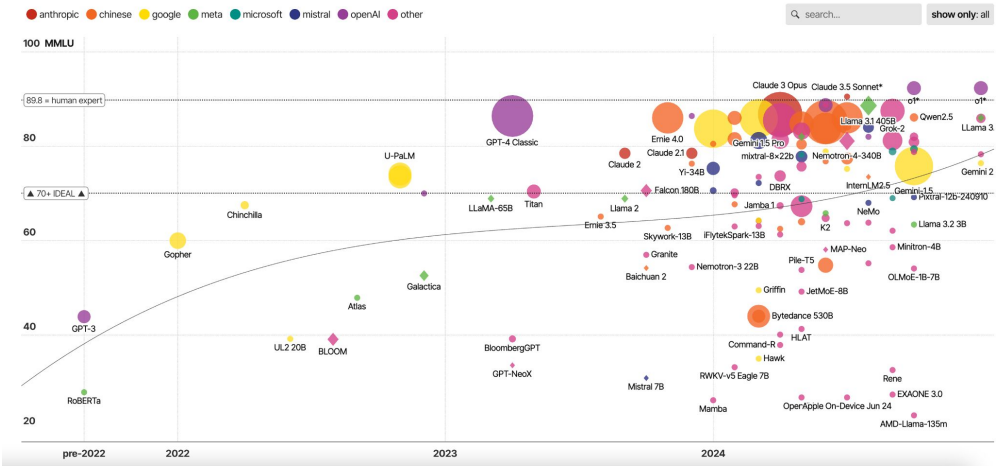


Wikipedia

Long Timelines, High Costs, and Scaling Challenges

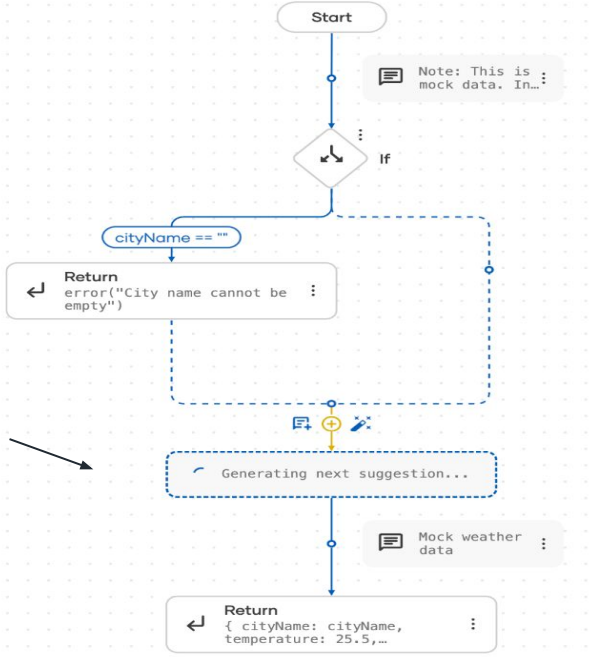
AI Development: Today

- Acceleration through Pre-trained models:
  - Trained on diverse, massive datasets
  - Reduced effort: Data collection, annotation model training for **many applications**
  - Cost-effective: Lower resource and time investment

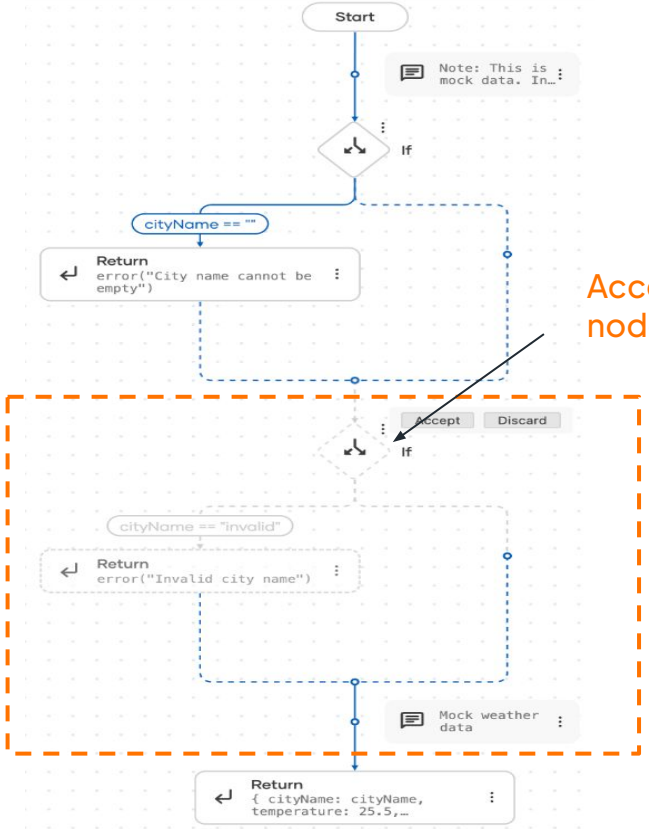


https://informationisbeautiful.net/

AI Development: Today



Predicting next node

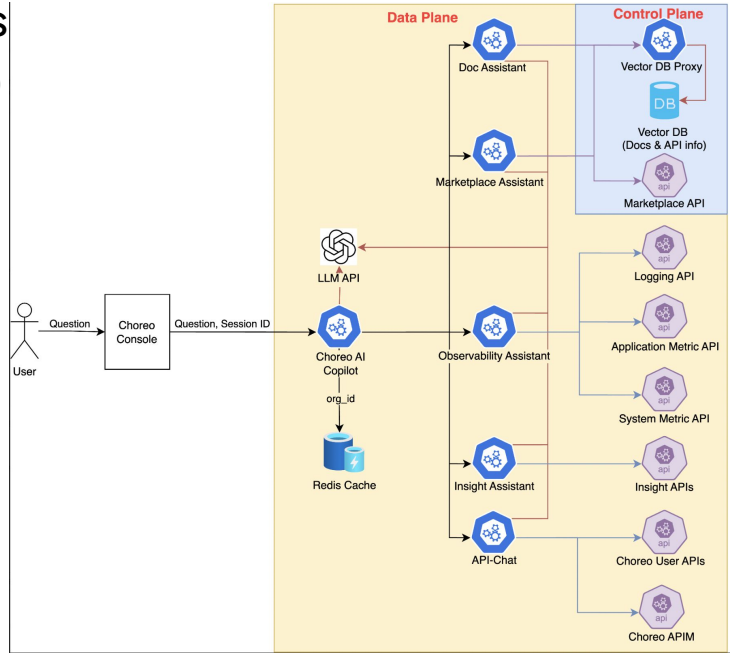


Accept / Discard the node suggestion

- Same use case implemented by invoking LLM APIs
- Program information provided as context
- No pre-processing, labeling, or training needed

AI Development: Today

- Building modern AI applications involves
  - Developing the AI component (e.g., Agents)
  - Integrating it with the larger application
- Building the AI component itself is solving an integration problem
- The focus has shifted from training models to integrating systems:
  - Choreo copilot architecture
- These integrations have different characteristics
  - Execution is not deterministic



The AI Production Gap – Scaling Beyond Prototyping

- Prototyping is easy (can be made to work locally), but deploying AI at scale is challenging
- 74% of AI projects fail to move beyond PoCs (IBM Investor Day 2025).
- The complexity lies in
  - Integration, Evaluation, Security, Governance, Monitoring
- These are primarily software engineering challenges
- Addressing these challenges requires a new set of capabilities/abstractions

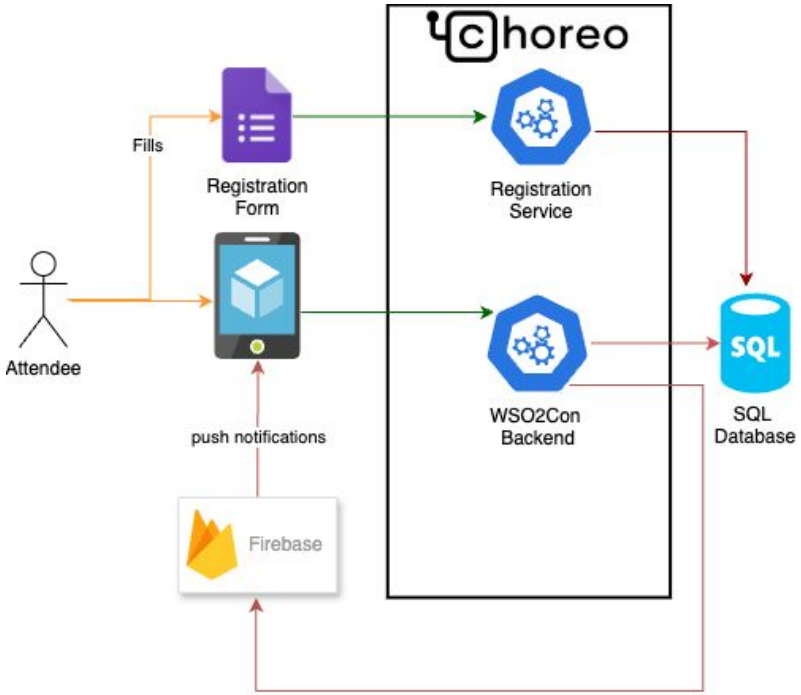


Taking AI to Production – Where to Start?

- Selecting tools to build AI Apps
  - Low-code vs. Pro-code development
  - Integration developers vs. AI developers
- **Integration tools with AI app building capabilities** can significantly speed up development
  - Why? As pointed out above, **modern AI app building has shifted from training models to integrating systems**
- Low-code tools are becoming increasingly popular
  - Particularly useful for building:
    - AI agents
    - Retrieval-Augmented Generation (RAG) systems, etc.

Buildings AI Apps

- WSO2 conference mobile app, 2024
  - Registration flow: Attendees submit the registration form, and their data is recorded
  - Conference dashboard: View session details, speaker information, and more
- WSO2 conference mobile app, 2025
  - New features: WSO2Con Assistant, Session Advisor, Attendee Connections, Expert Finder
  - **Discussed in the AI Lab: Install the latest version and log out and back in (if needed)**

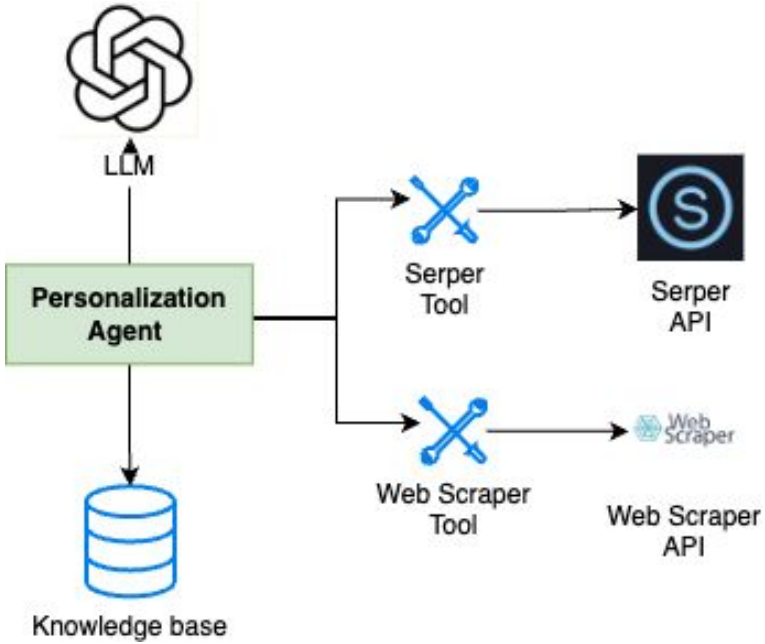


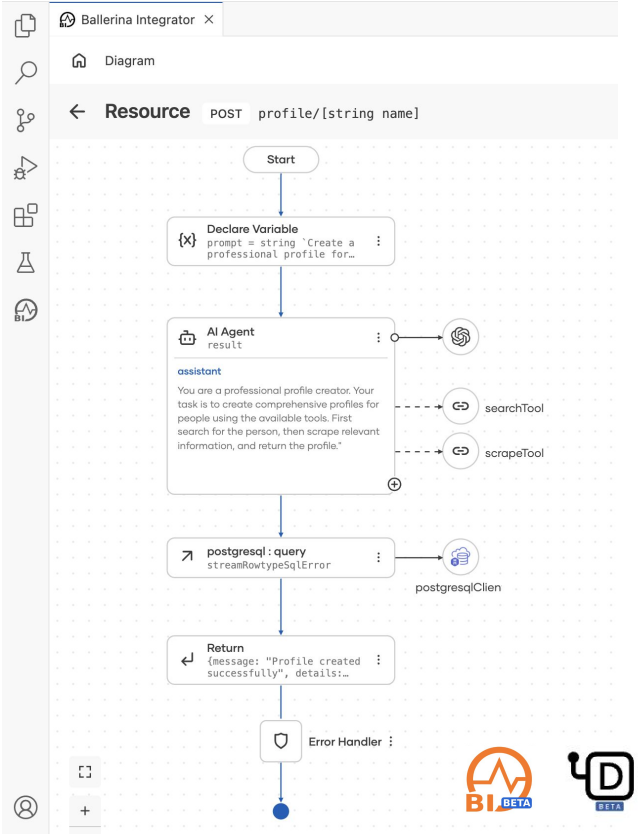
WSO2 conference mobile app, 2024

Buildings AI Apps

Pattern	Requirements
Application that uses Gen-AI model to generate content (Generates text, images, voice, etc.)	Models → LLMs, Diffusion Models, Speech Models
Application with a RAG (Grounds AI responses with knowledge for better accuracy)	Data Sources → Databases, documents, real-time feeds Models → LLMs, Embeddings models Retrieval System → Vector database (Pinecone, Chroma, etc)
An application that uses an agent The agent is the application	Determine Required Tools → APIs, functions, external systems, databases Tools → Use existing tools or build new? Models → LLMs for reasoning and planning

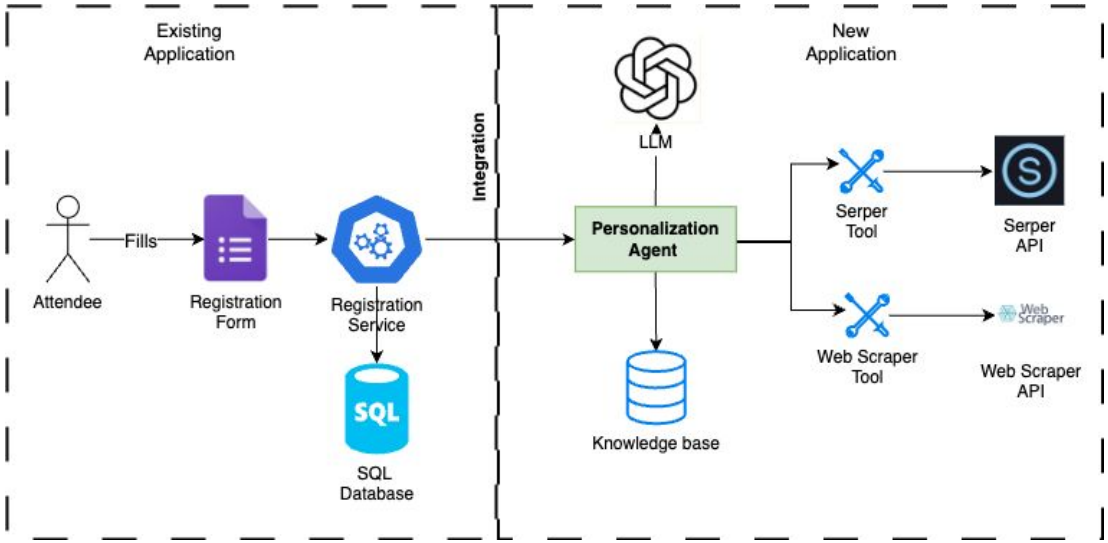
Buildings AI Apps: Personalization Agent





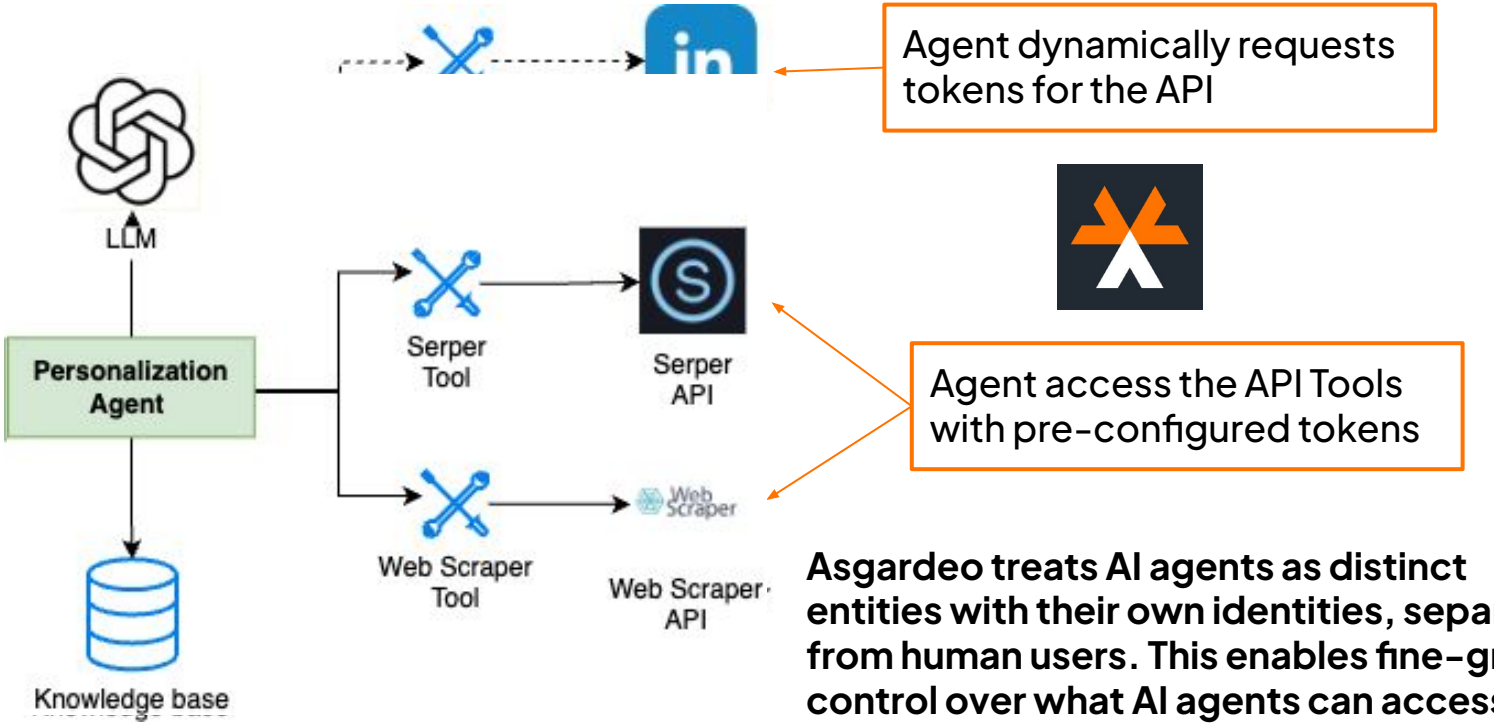
Search the internet and scrape information about the attendees and create a personalized profile for each attendee (agent code)

Integrated application – Agent in an Application



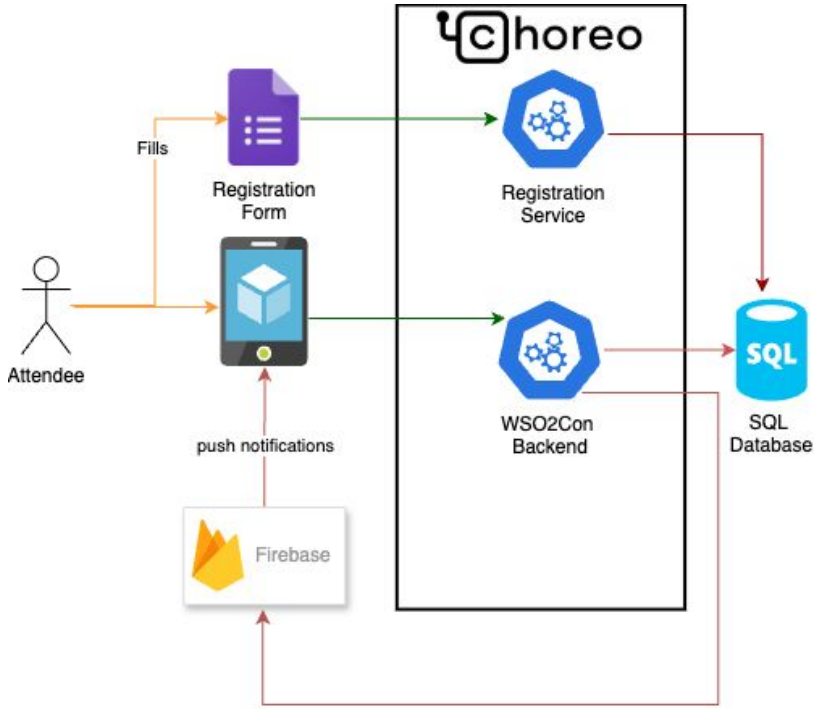
Personalization agent was integrated to the existing enterprise application for attendee registration

Agent IAM: Managing Access Control for AI Agents

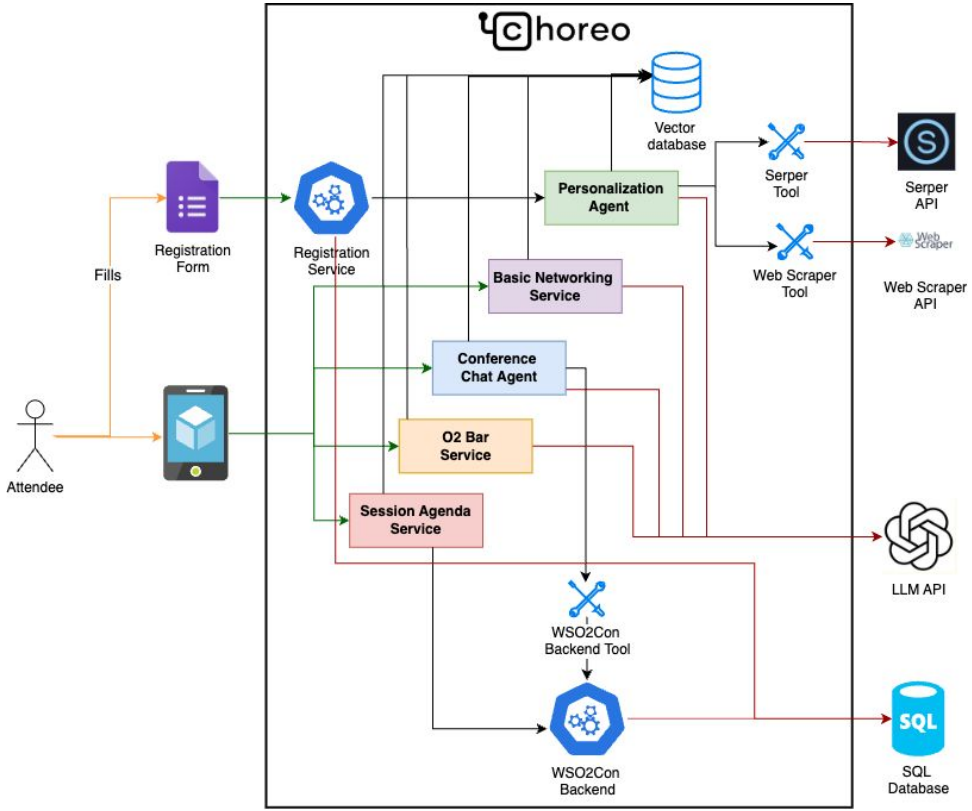


Asgardeo treats AI agents as distinct entities with their own identities, separate from human users. This enables fine-grained control over what AI agents can access and how they operate

Buildings AI apps



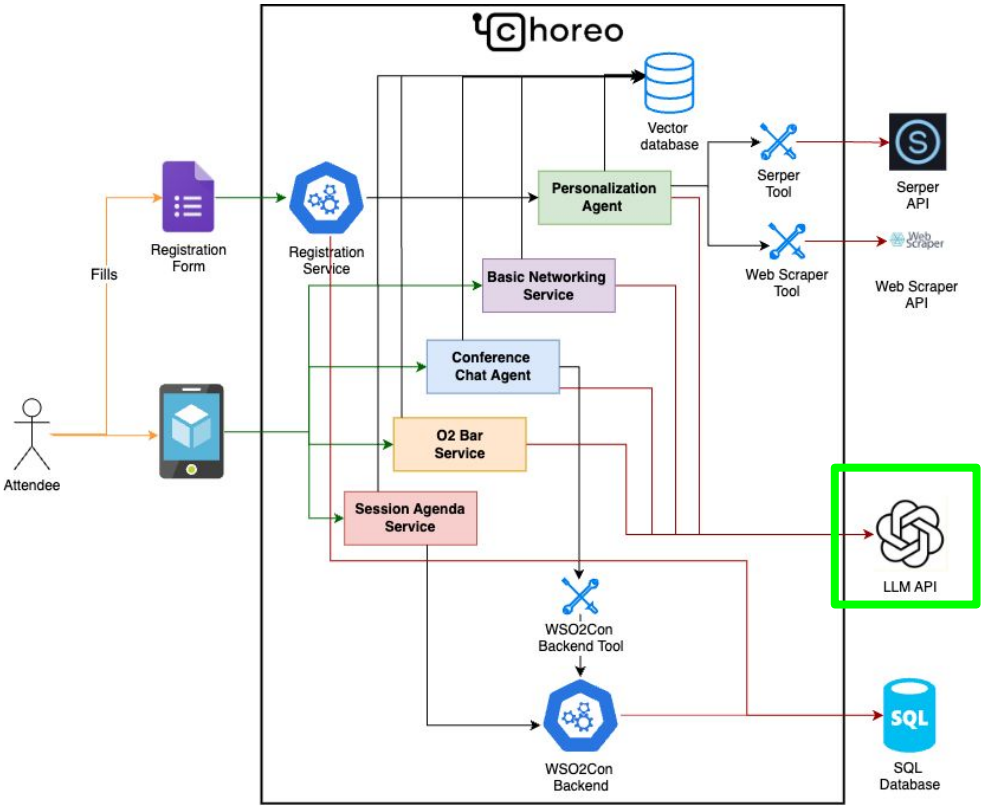
WSO2 Conference mobile app, 2024



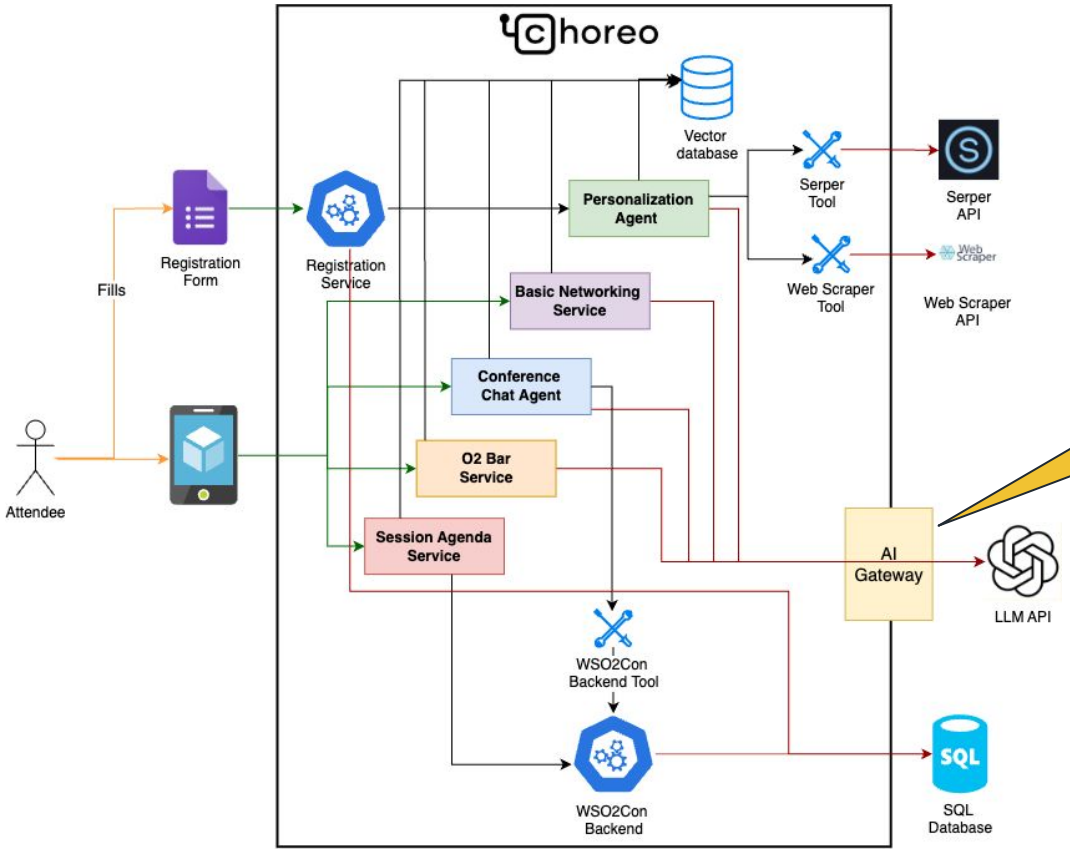
WSO2 Conference mobile app, 2025

AI API Management

- AI Apps access AI APIs (as a service)
- You need to secure, manage, monitor, and govern these AI APIs (to ensure reliability, efficiency, and compliance)
- APIM management solutions need to be introduced with additional capabilities to support AI API management (e.g., Gen-AI usage, token-based rate limiting, Guardrails, Model Routing, Failover load balancing, etc.).
- We have introduced these capabilities to WSO2 API Manager
- AI Gateway (core part of AI API Management solutions)



Deploy, Monitor and Manage LLM traffic

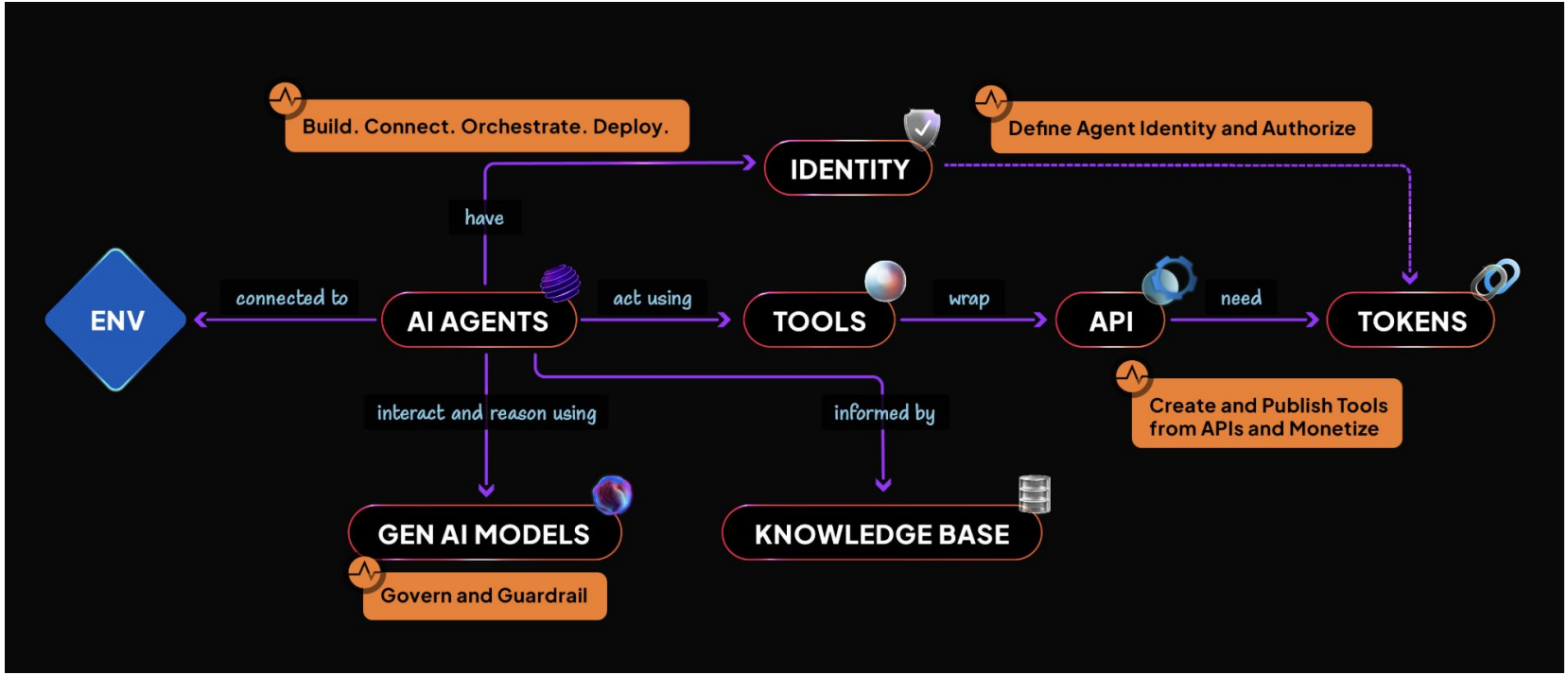


- Sanitize PII
- Stay in budget
- Get insights into GenAI usage
- Increase availability with load balancing and fallbacks

Conclusion

- In the past people spent years building models **before they could even build a prototype**
- Today, with pre-trained models, agents, RAGS **building prototypes is much faster**
- The real challenge comes when **taking these prototypes to production**
- This requires integrating systems, securing them, governing them (setting up proper guardrails), and monitoring them
- To do this, you need new set of capabilities and abstractions
- This is what we are introducing into WSO2 products
- We've made significant progress, but there's still more to do
- We are **expanding our existing capabilities** and introducing new ones, including **evaluation capabilities and agent observability**

Conclusion


Build. Connect. Orchestrate. Deploy.
IDENTITY
Define Agent Identity and Authorize
have
ENV
connected to
AI AGENTS
act using
TOOLS
wrap
API
need
TOKENS
Create and Publish Tools from APIs and Monetize
interact and reason using
informed by
GEN AI MODELS
KNOWLEDGE BASE
Govern and Guardrail

Question Time!

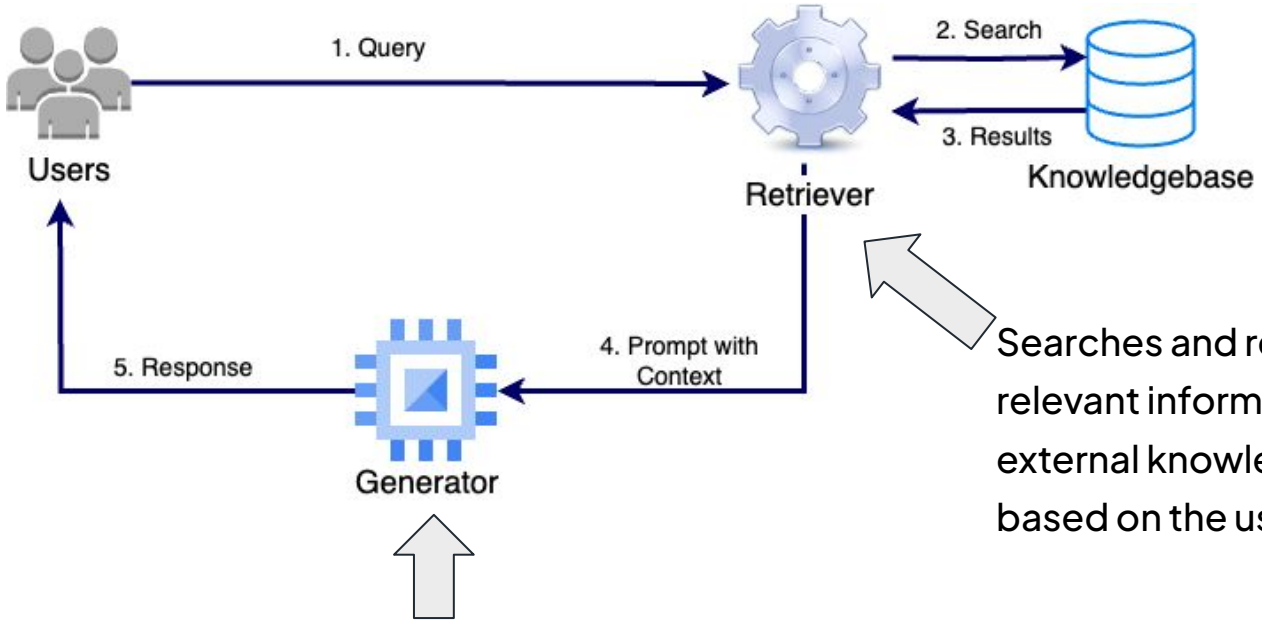



Thank you!
WSO2CON2025

Evaluating AI Apps

- Why evaluation is important
  - Ensures accuracy: Verifies performance and effectiveness
  - Maintains consistency: Guarantees uniform results across scenarios
  - Reduces bias: Ensures fairness and balanced outcomes
  - Optimizes speed: Enhances performance efficiency
- Key evaluation steps
  - Define metrics: Choose relevant performance metrics (e.g., accuracy, precision)
  - Prepare dataset: Use data reflective of your use case for reliable results
  - Interpret results: Analyze to identify strengths and weaknesses
  - Improve quality: Fine-tune the model, retrain (e.g. use diverse, representative datasets), and optimize speed, etc.

Evaluating RAGs



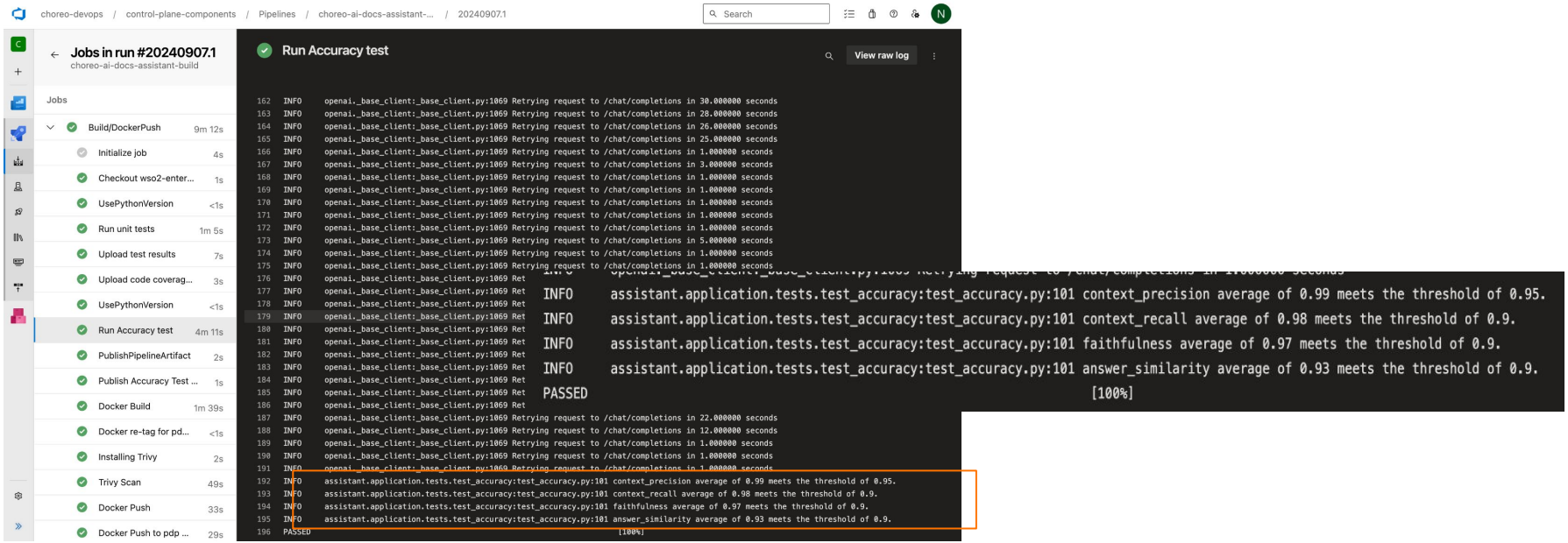
Searches and retrieves relevant information from an external knowledge base based on the user's query

It processes both the retrieved data and the user's query to generate an informed and relevant output

Evaluating RAGs

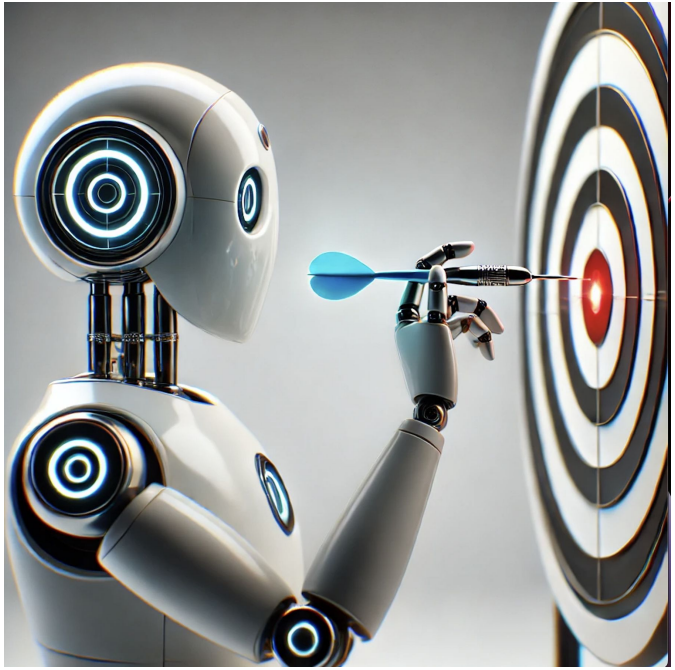
- Metrics:
  - Various metrics are used to evaluate RAG applications
  - RAGAS Tool: A comprehensive framework for RAG evaluation, though not all metrics apply to every use case
  - LLM as judge
- Evaluation Needs
  - Question: User's query input to the RAG system
  - Context: Documents retrieved from the knowledge base by the retriever
  - Answer: Response generated by the RAG based on the retrieved context
  - Ground Truth (GT): Human-annotated correct answer to the user's query

Evaluating RAGs

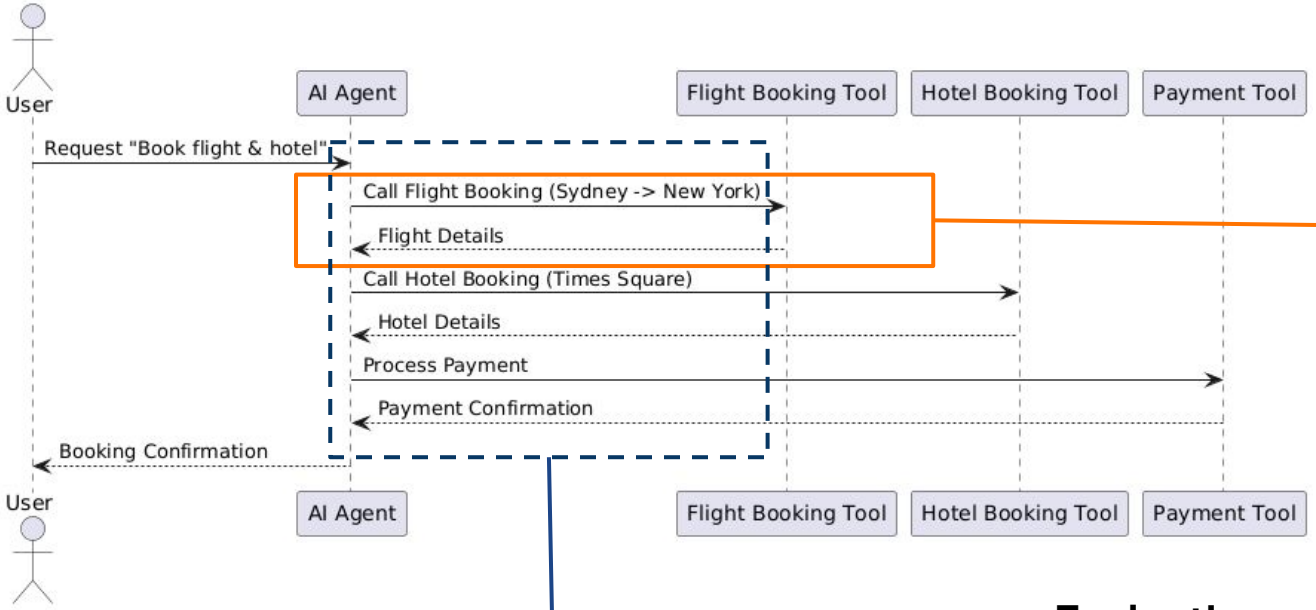


Evaluating Agents

- Final Response
  - Evaluate the agent’s final output for correctness, relevance, and completeness.
- Single Step
  - Evaluate each individual step, including the input to the tool, the tool's output, and whether the correct tool and parameters were selected.
- Trajectory
  - Evaluate whether the agent followed the expected sequence of actions (e.g., tool calls) to arrive at the correct final answer



Agent Evaluation



Single Step Evaluation

Expected Tool call:

tool_name: "Flight Booking Tool"
parameters:
from: "Sydney"
to: "Newyork"
travel_date: "17/03/2025"
return_date: "24/03/2025"

Trajectory Evaluation

Expected Trajectory:
Agent -> Flight Booking Tool -> Agent -> Hotel Booking Tool -> Agent -> Payment Tool

Evaluation capabilities will be available in WSO2 products soon!

Question Time!

