Los Angelesl September 17th, 2025

# AWS SUMMIT

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SVS203

# Building Production-Ready Agentic AI architectures with AWS Serverless

**Anton Aleksandrov**
Principal Solutions Architect, Serverless
AWS

**Debasis Rath**
Senior Solutions Architect, Serverless
AWS

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Enterprises are doubling down on agents

33%

of enterprise software apps will include agentic AI by 2028, up from less than 1% in 2024.

Gartner, "Top strategic Technology Trends for 2025," October 2024.

15%

of day-to-day work decisions will be made autonomously through agentic AI by 2028.

Gartner, "Top Strategic Technology Trends: agentic AI – The evolution of Experience" February 2025

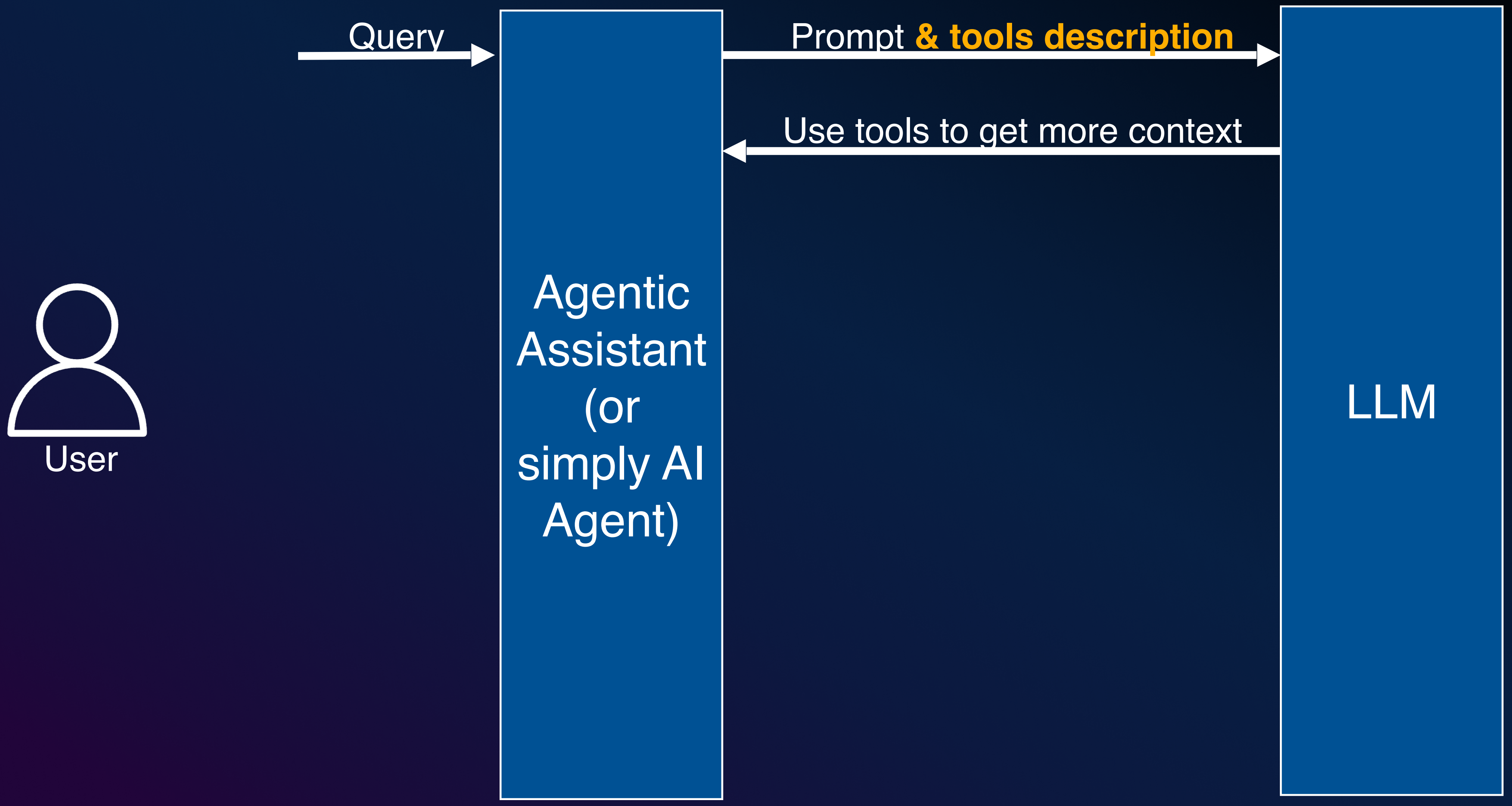
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# What are AI Agents?

Autonomous software systems that leverage AI to reason, plan, and complete tasks on behalf of humans or systems

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

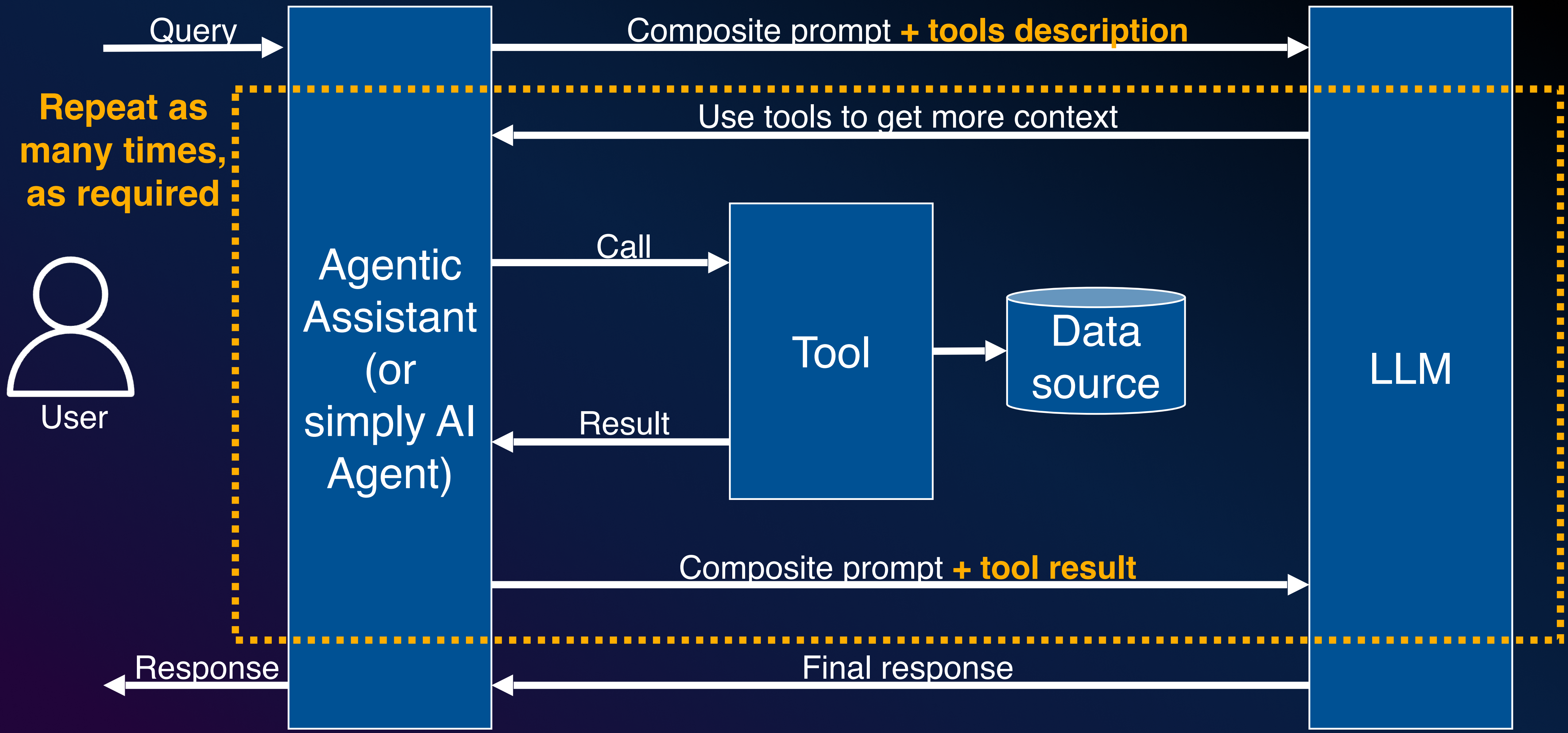
# A simple agentic workflow



Query
Prompt **& tools description**
Use tools to get more context
User
Agentic Assistant (or simply AI Agent)
LLM

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# A simple agentic workflow



Query

Composite prompt **+ tools description**

**Repeat as many times, as required**

Use tools to get more context

User

Agentic Assistant (or simply AI Agent)

Call

Tool

Data source

Result

LLM

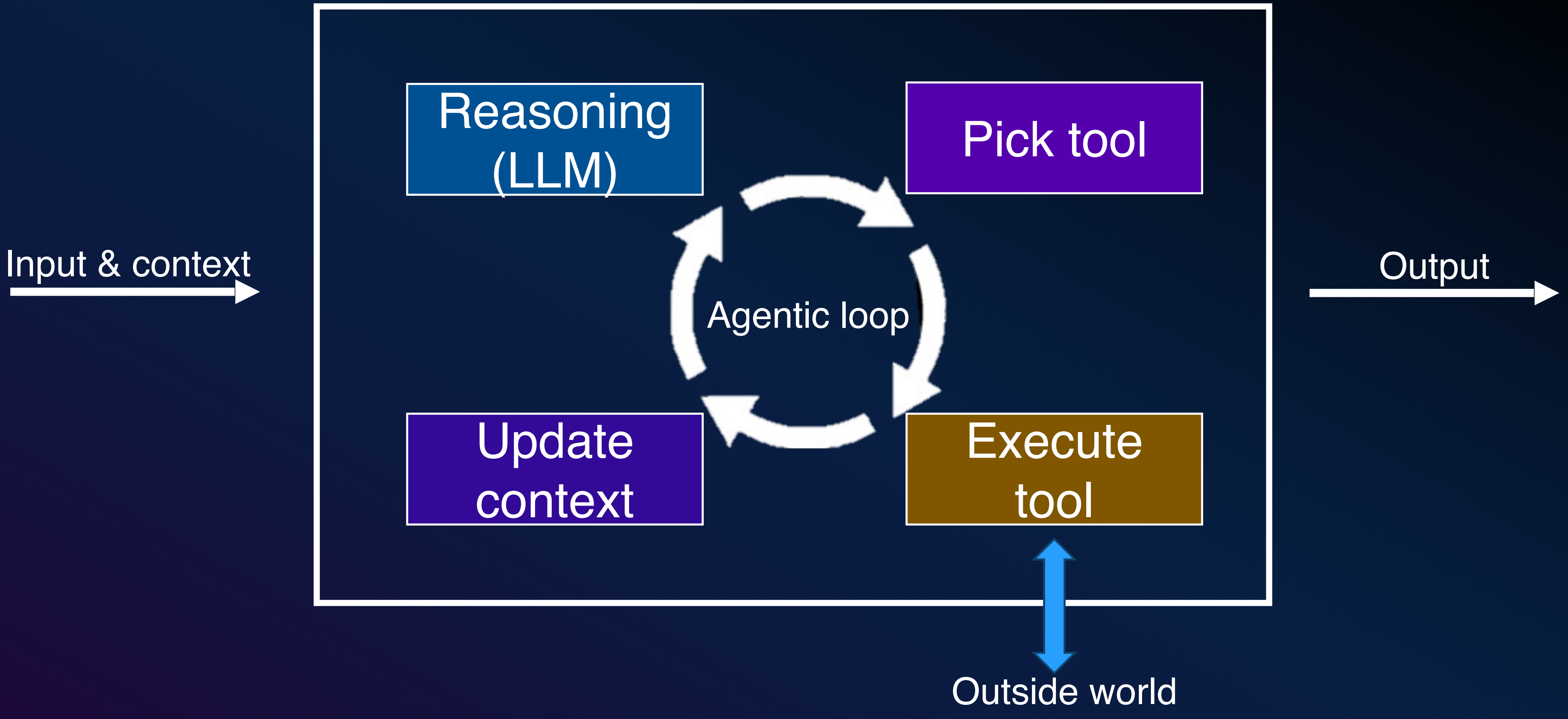
Composite prompt **+ tool result**

Response

Final response

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agentic AI system in a nutshell



Input & context

Reasoning (LLM)

Pick tool

Agentic loop

Update context

Execute tool

Output

Outside world

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Let’s build an agentic chatbot

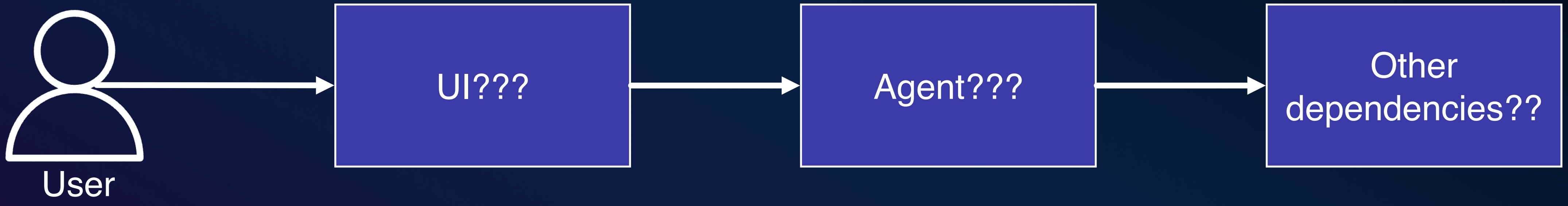
I want to build a Travel Agent that helps my employees to book business travel while enforcing corporate travel policies.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 1: Agentic Chatbot

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 1: Agentic Chatbot



User

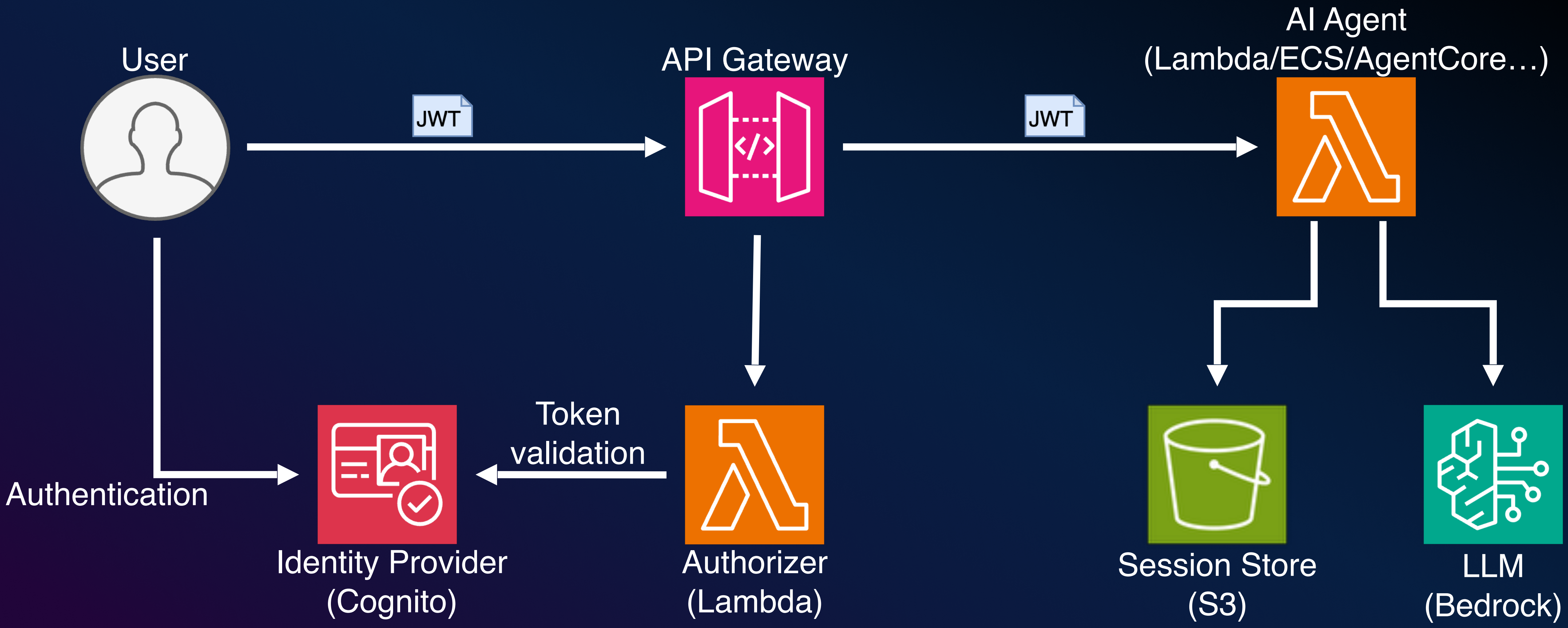
UI???

Agent???

Other dependencies??

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 1: Agentic Chatbot

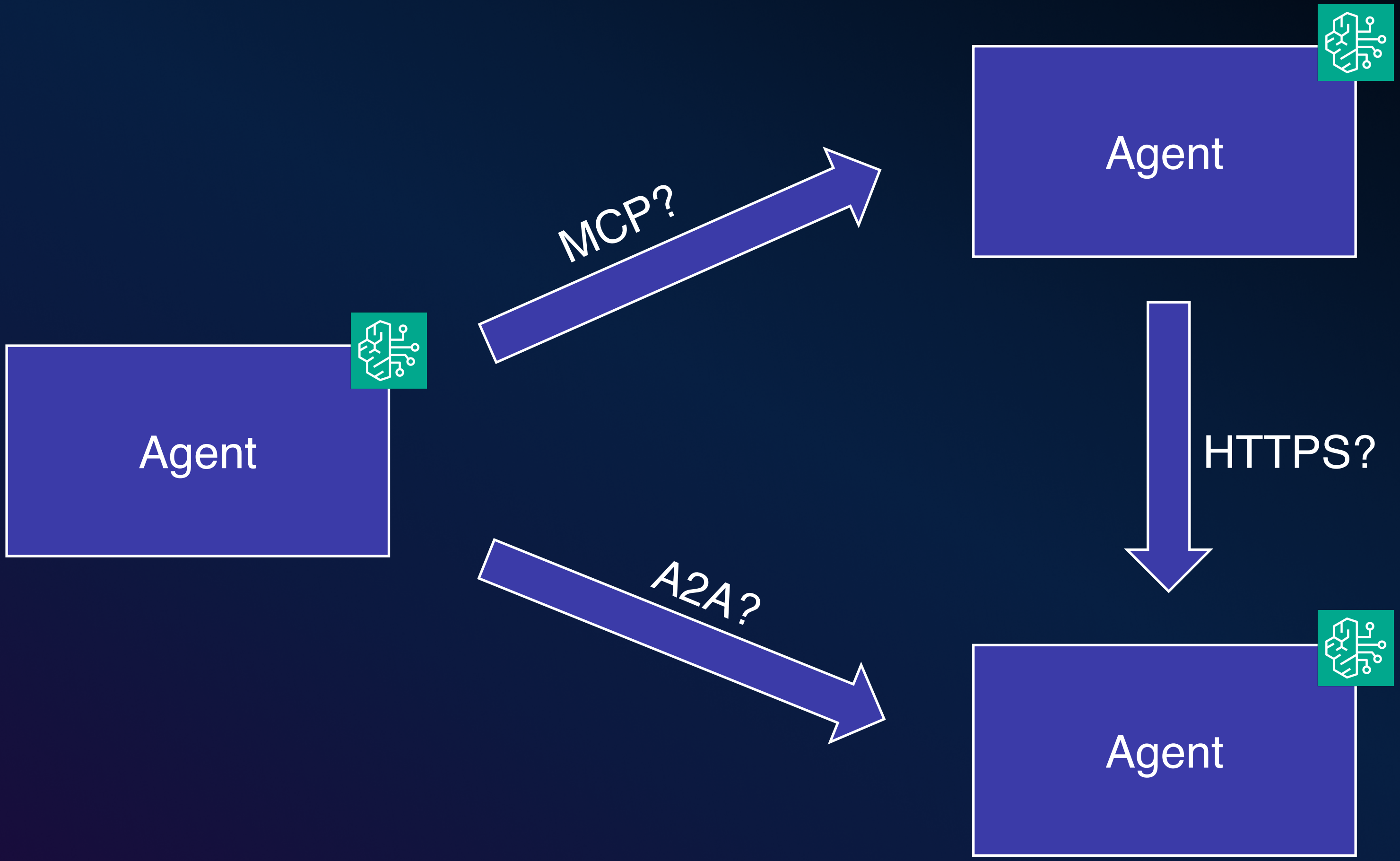


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 2:
# Multi-agent systems

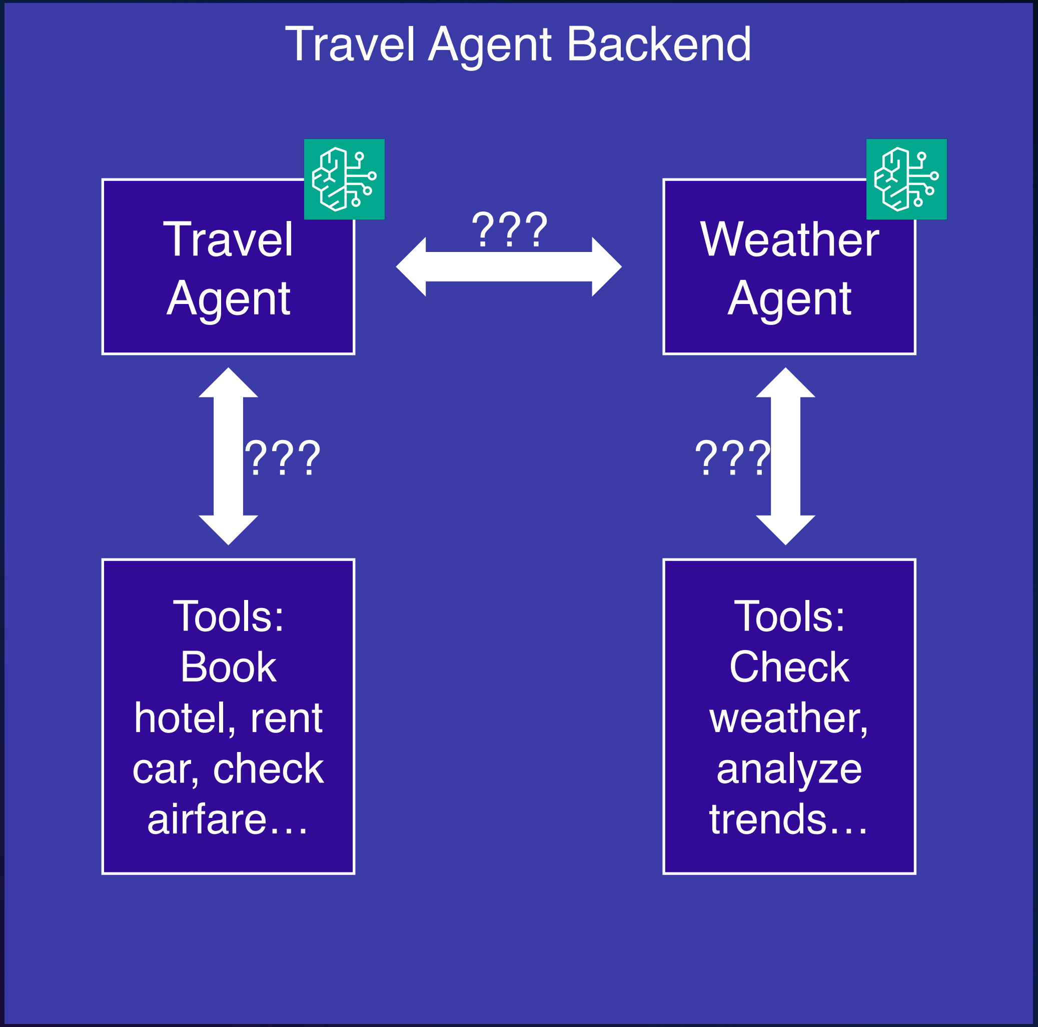
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 2: Multi-agent system



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 2: Multi-agent system



Travel Agent Backend

Travel Agent

???

Weather Agent

???

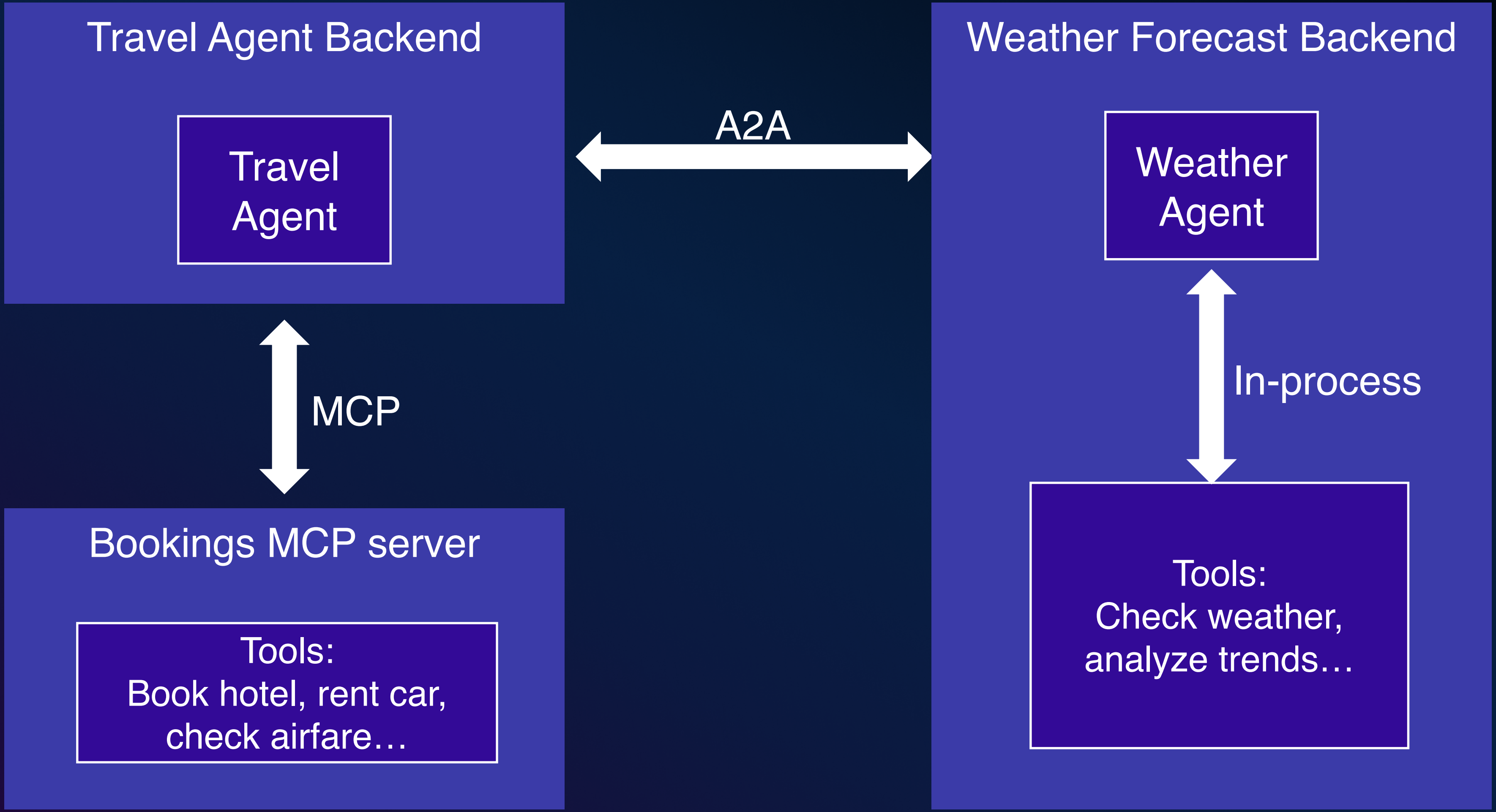
???

Tools: Book hotel, rent car, check airfare…

Tools: Check weather, analyze trends…

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 2: Multi-agent system



Travel Agent Backend

Travel Agent

A2A

Weather Forecast Backend

Weather Agent

MCP

In-process

Bookings MCP server

Tools:
Book hotel, rent car,
check airfare…

Tools:
Check weather,
analyze trends…

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

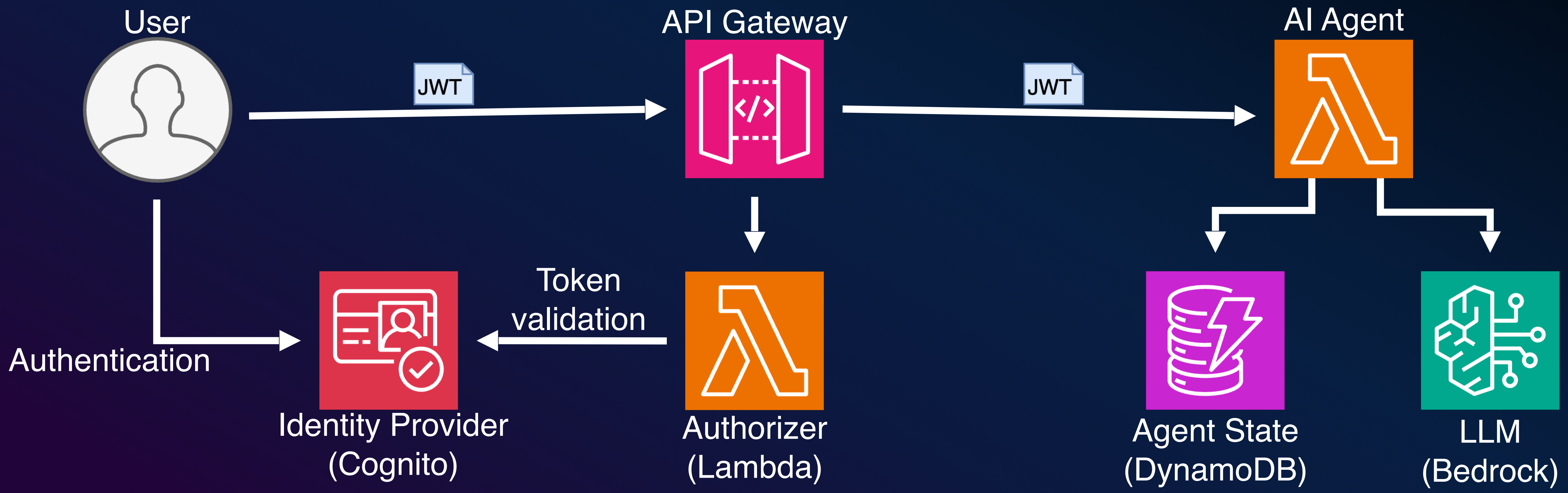
# Use-case 3:
# Security, Operations, Governance

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Use-case 3: Security, Operations, Governance

IAM
Cert. Manager
KMS
CloudWatch
X-Ray
GuardDuty
Inspector
WAF
CloudTrail
VPC

---



User

JWT

API Gateway

JWT

AI Agent

Authentication

Token validation

Identity Provider (Cognito)

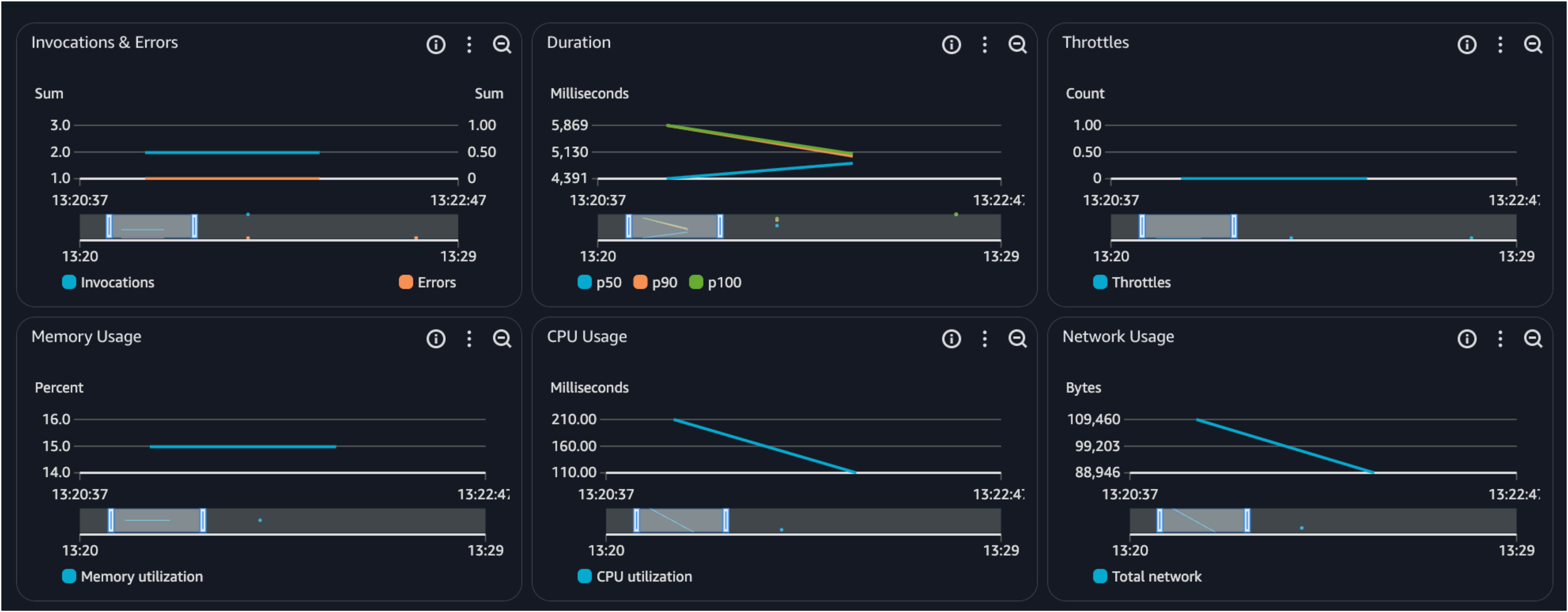
Authorizer (Lambda)

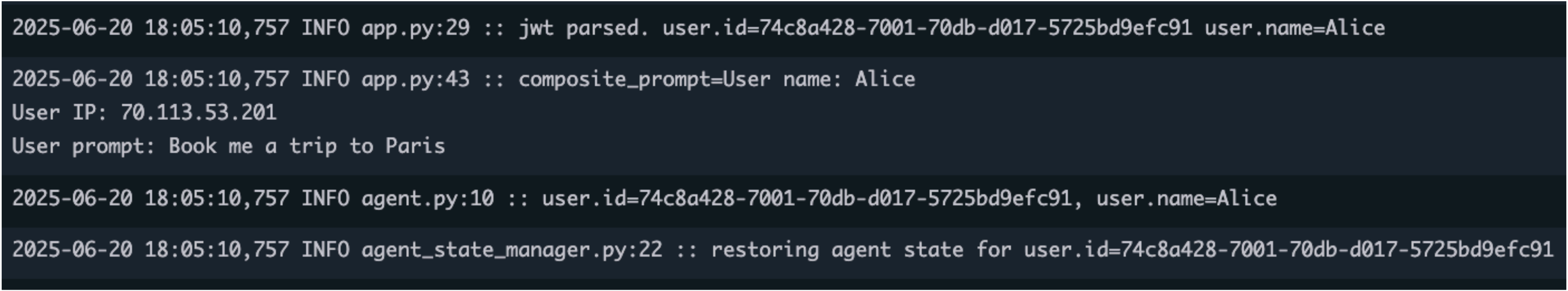
Agent State (DynamoDB)

LLM (Bedrock)

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Observability

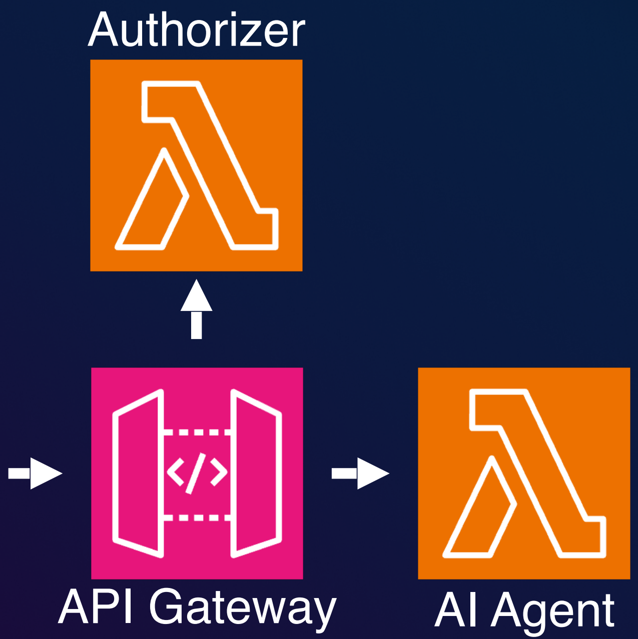


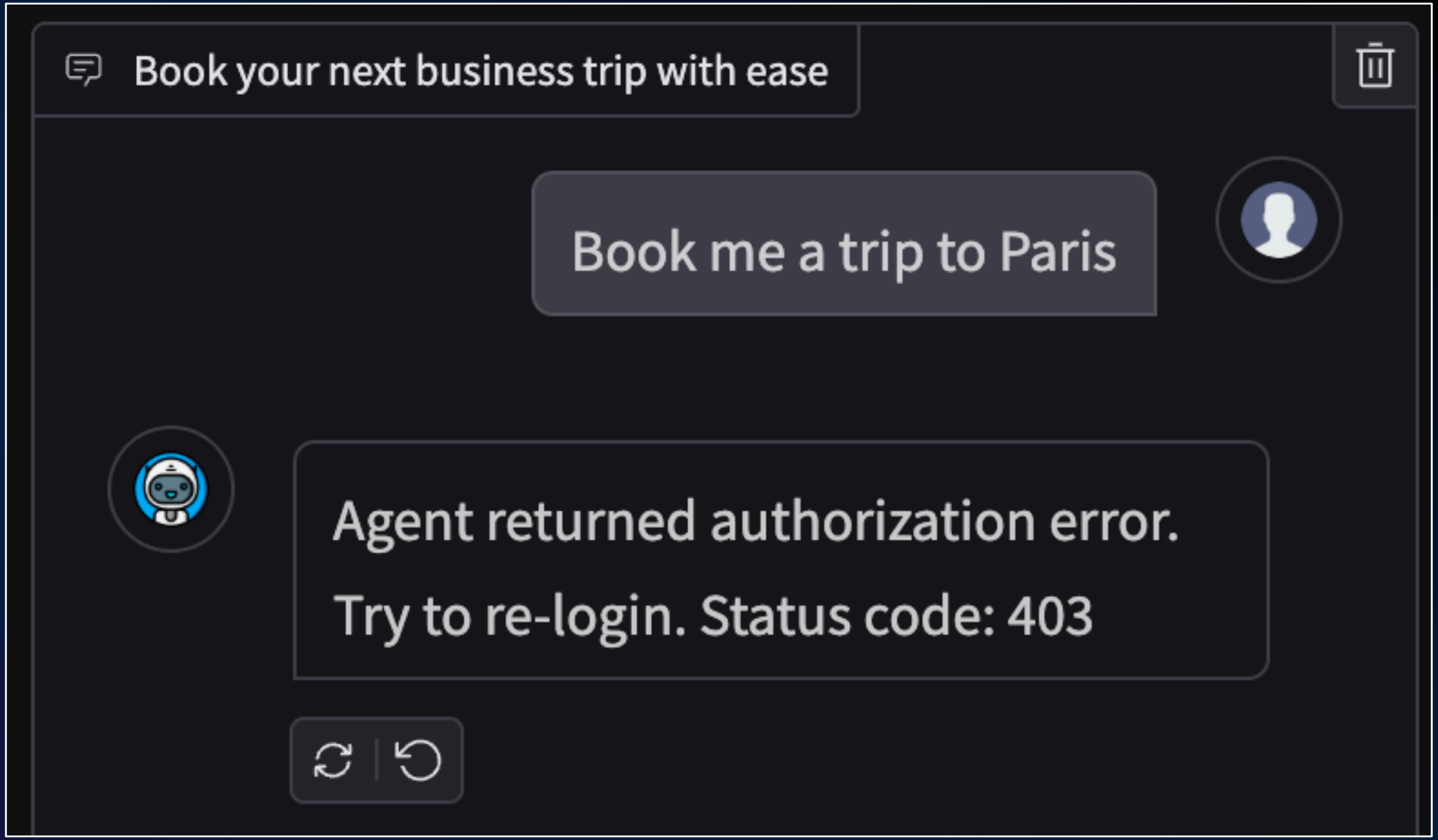


```
2025-06-20 18:05:10,757 INFO app.py:29 :: jwt parsed. user.id=74c8a428-7001-70db-d017-5725bd9efc91 user.name=Alice

2025-06-20 18:05:10,757 INFO app.py:43 :: composite_prompt=User name: Alice
User IP: 70.113.53.201
User prompt: Book me a trip to Paris

2025-06-20 18:05:10,757 INFO agent.py:10 :: user.id=74c8a428-7001-70db-d017-5725bd9efc91, user.name=Alice

2025-06-20 18:05:10,757 INFO agent_state_manager.py:22 :: restoring agent state for user.id=74c8a428-7001-70db-d017-5725bd9efc91
```

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Authorization



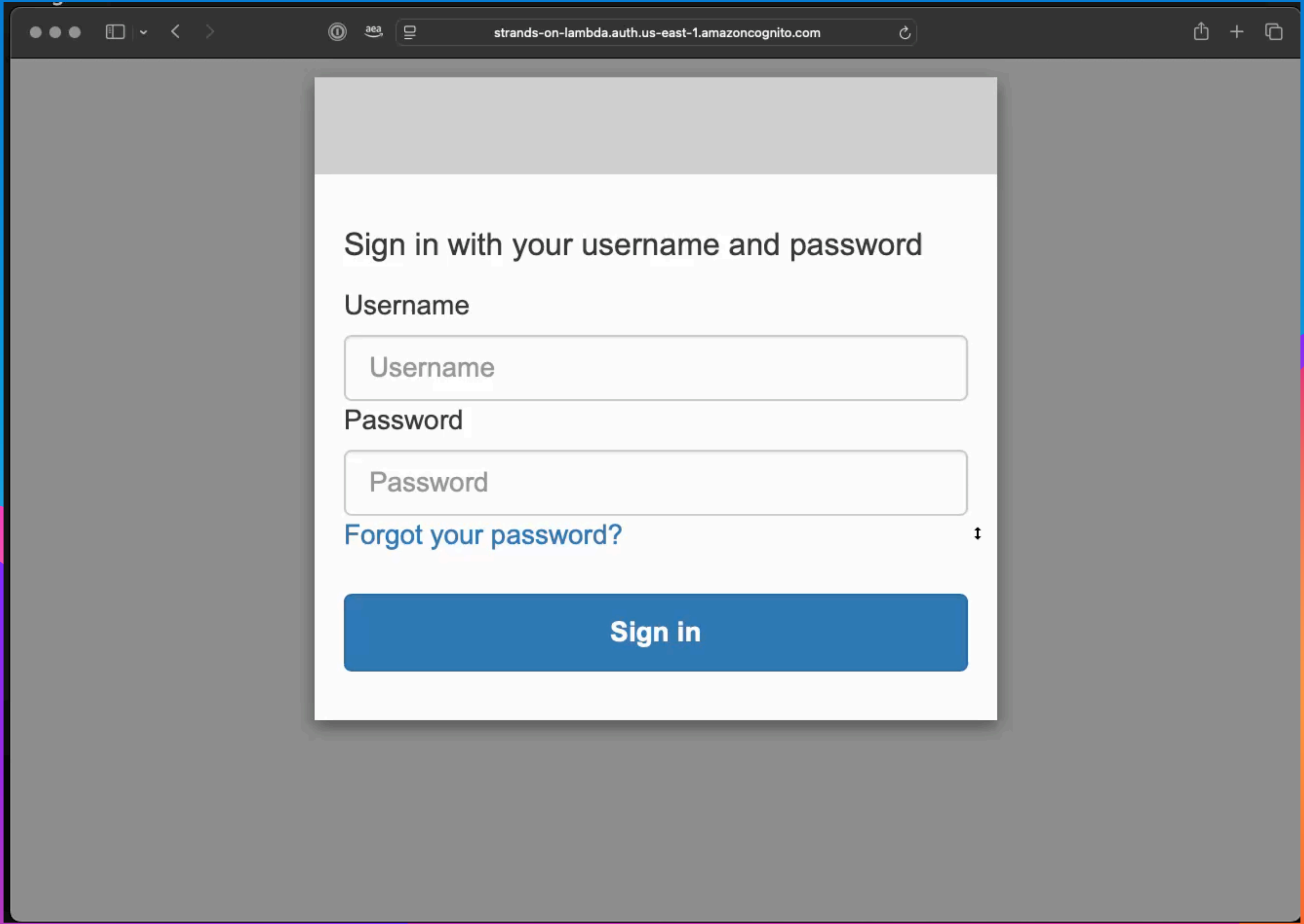

Book your next business trip with ease

Book me a trip to Paris

Agent returned authorization error.
Try to re-login. Status code: 403

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Demo time



strands-on-lambda.auth.us-east-1.amazoncognito.com

Sign in with your username and password

Username

Username

Password

Password

Forgot your password?

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Conclusion

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Conclusion

1

**Agents are applications calling LLMs in loops. Your existing knowledge is your accelerator.**

2

**Use the right tool for the job. Decouple when needed, use proper protocols.**

3

**AWS services provide built-in enterprise-grade capabilities for building AI agents.**

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Continue your learning journey

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!

Please complete the session survey in the mobile app

**Anton Aleksandrov**
in antonal80
aal80.github.io/whoami

**Debasis Rath**
in rathd

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SVS203

# Building Production-Ready Agentic AI architectures with AWS Serverless

**Anton Aleksandrov**
Principal Solutions Architect, Serverless
AWS

**Debasis Rath**
Senior Solutions Architect, Serverless
AWS

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Enterprises are doubling down on agents

33%

of enterprise software apps will include agentic AI by 2028, up from less than 1% in 2024.

Gartner, “Top strategic Technology Trends for 2025,” October 2024.

15%

of day-to-day work decisions will be made autonomously through agentic AI by 2028.

Gartner, “Top Strategic Technology Trends: agentic AI – The evolution of Experience” February 2025

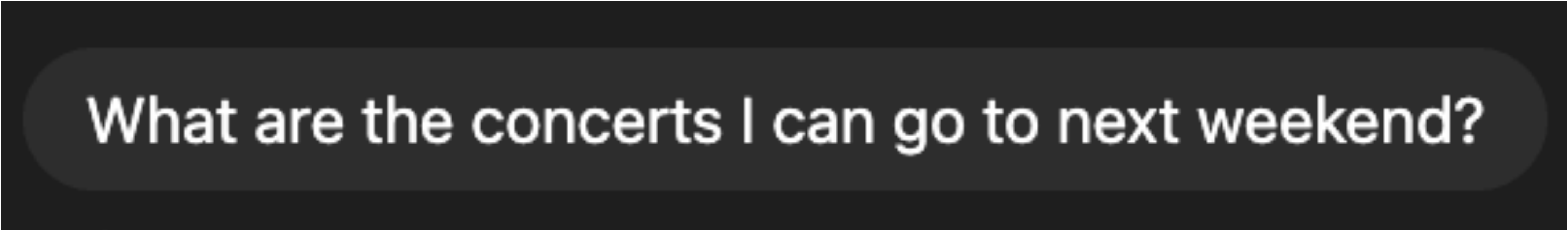
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# What are AI Agents?

Autonomous software systems that leverage AI to reason, plan, and complete tasks on behalf of humans or systems

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

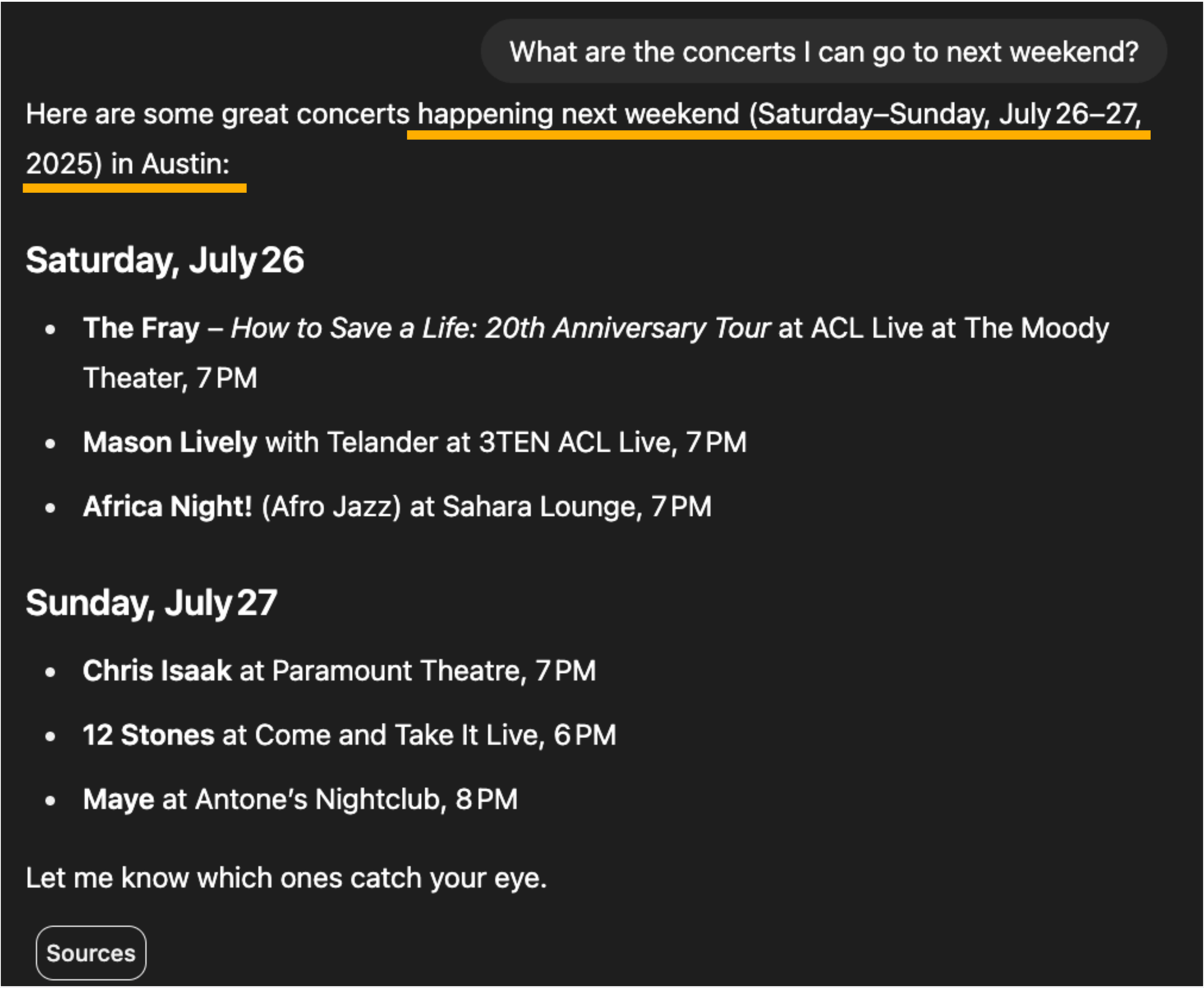
# Adding real-time context dynamically



What are the concerts I can go to next weekend?

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Adding real-time context dynamically



What are the concerts I can go to next weekend?

Here are some great concerts happening next weekend (Saturday–Sunday, July 26–27, 2025) in Austin:

### Saturday, July 26

- **The Fray** – *How to Save a Life: 20th Anniversary Tour* at ACL Live at The Moody Theater, 7 PM
- **Mason Lively** with Telander at 3TEN ACL Live, 7 PM
- **Africa Night!** (Afro Jazz) at Sahara Lounge, 7 PM

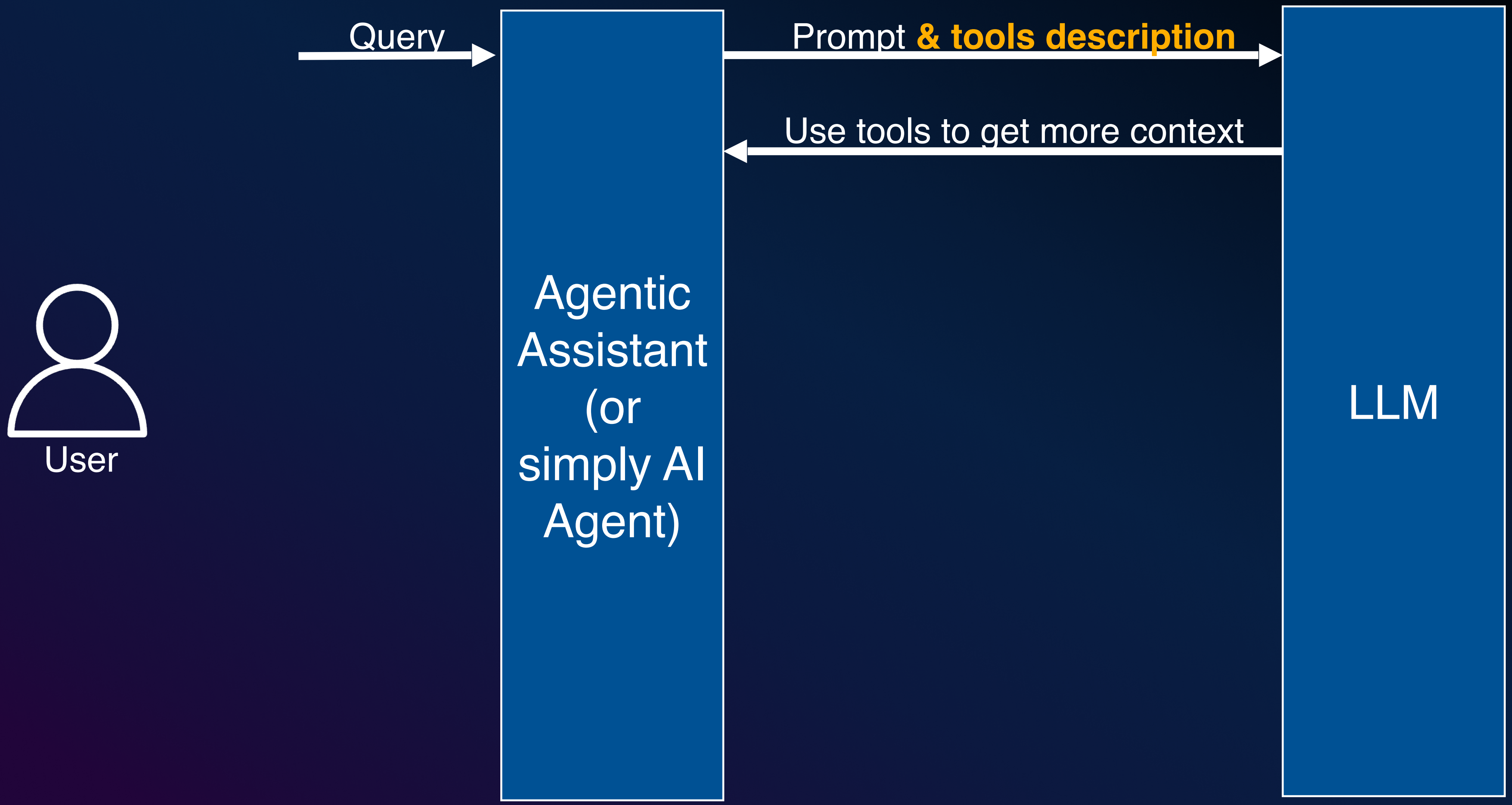
### Sunday, July 27

- **Chris Isaak** at Paramount Theatre, 7 PM
- **12 Stones** at Come and Take It Live, 6 PM
- **Maye** at Antone's Nightclub, 8 PM

Let me know which ones catch your eye.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# A simple agentic workflow



Query

Prompt **& tools description**

Use tools to get more context

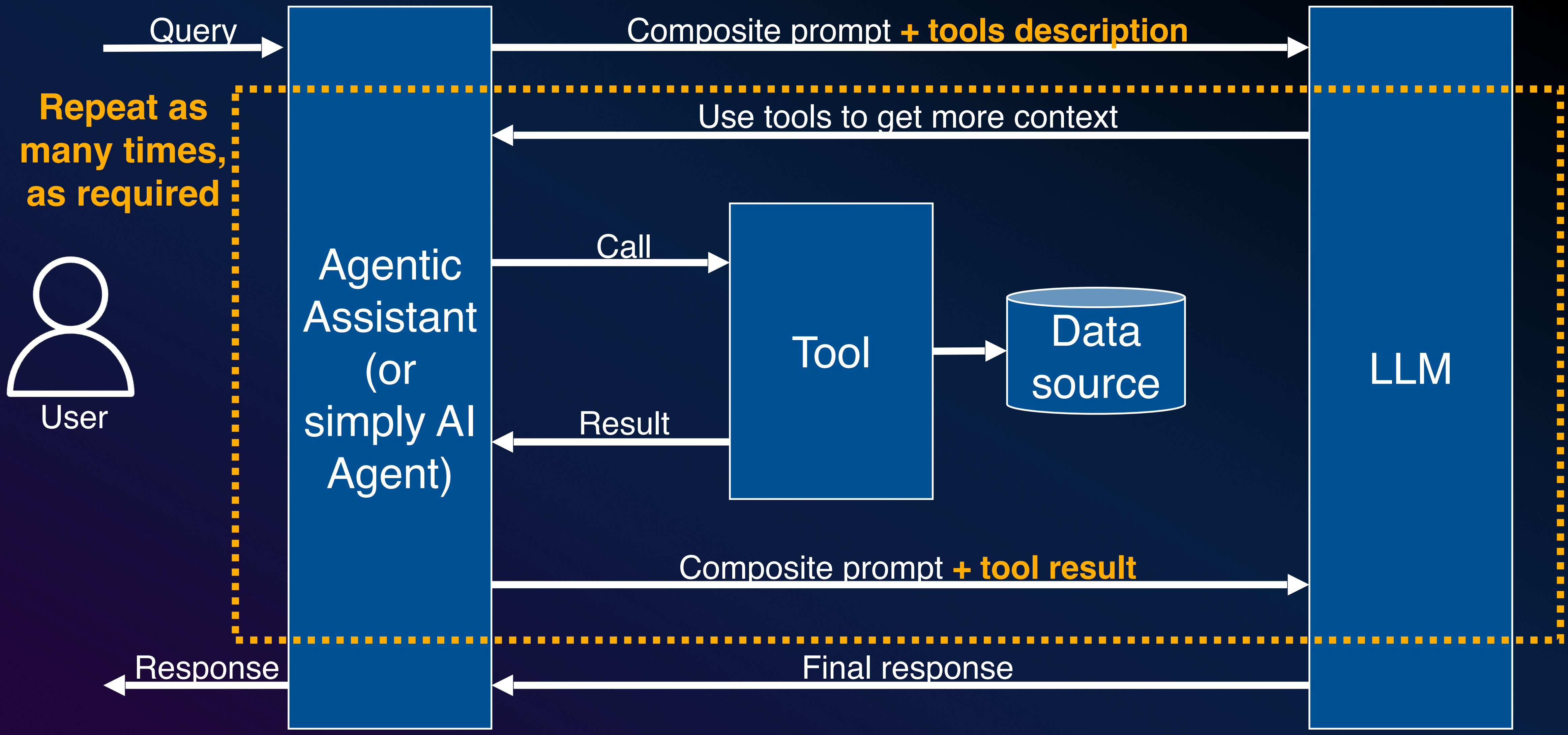
User

Agentic Assistant (or simply AI Agent)

LLM

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

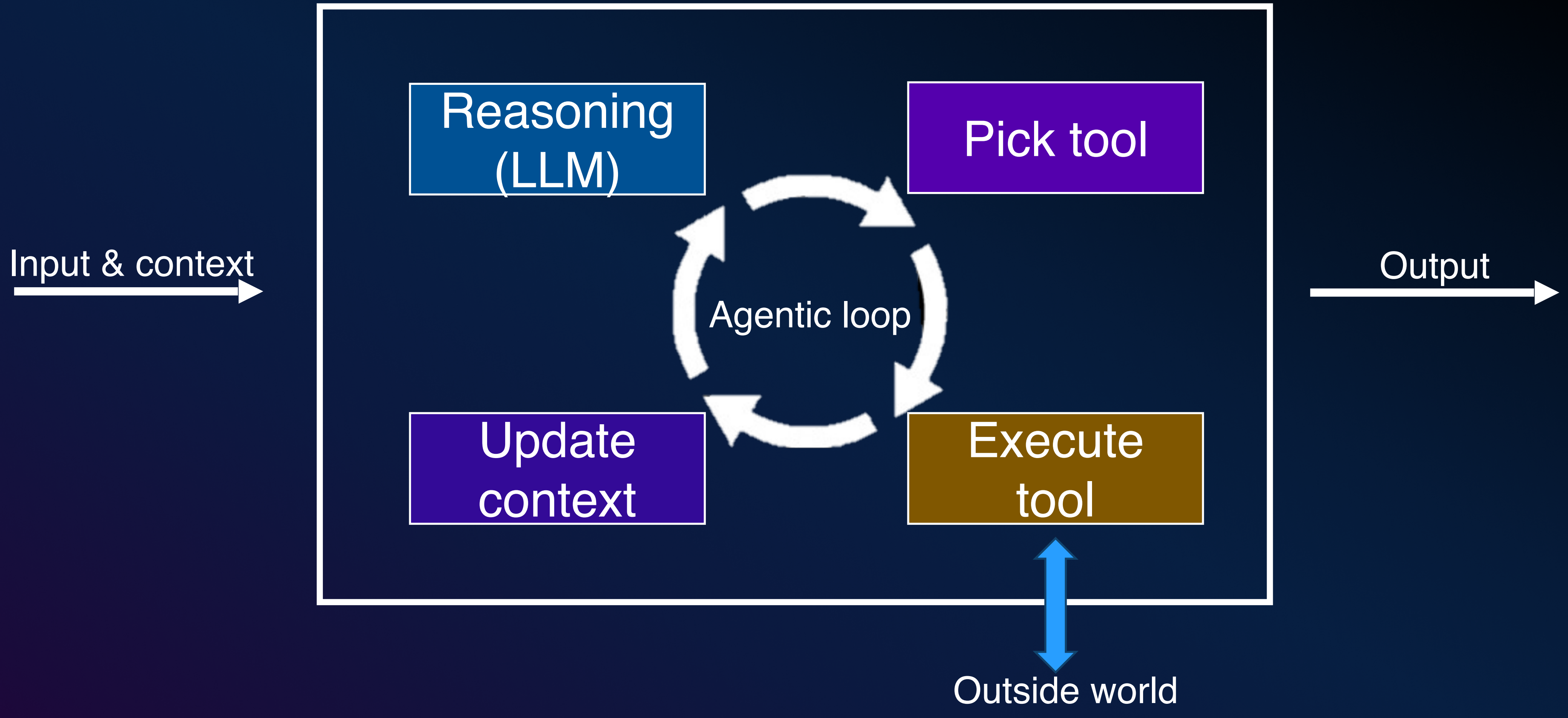
# A simple agentic workflow



Query
Composite prompt + **tools description**
**Repeat as many times, as required**
Use tools to get more context
User
Agentic Assistant (or simply AI Agent)
Call
Tool
Data source
Result
LLM
Composite prompt + **tool result**
Response
Final response

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agentic AI system in a nutshell



Input & context

Reasoning (LLM)

Pick tool

Agentic loop

Execute tool

Update context

Output

Outside world

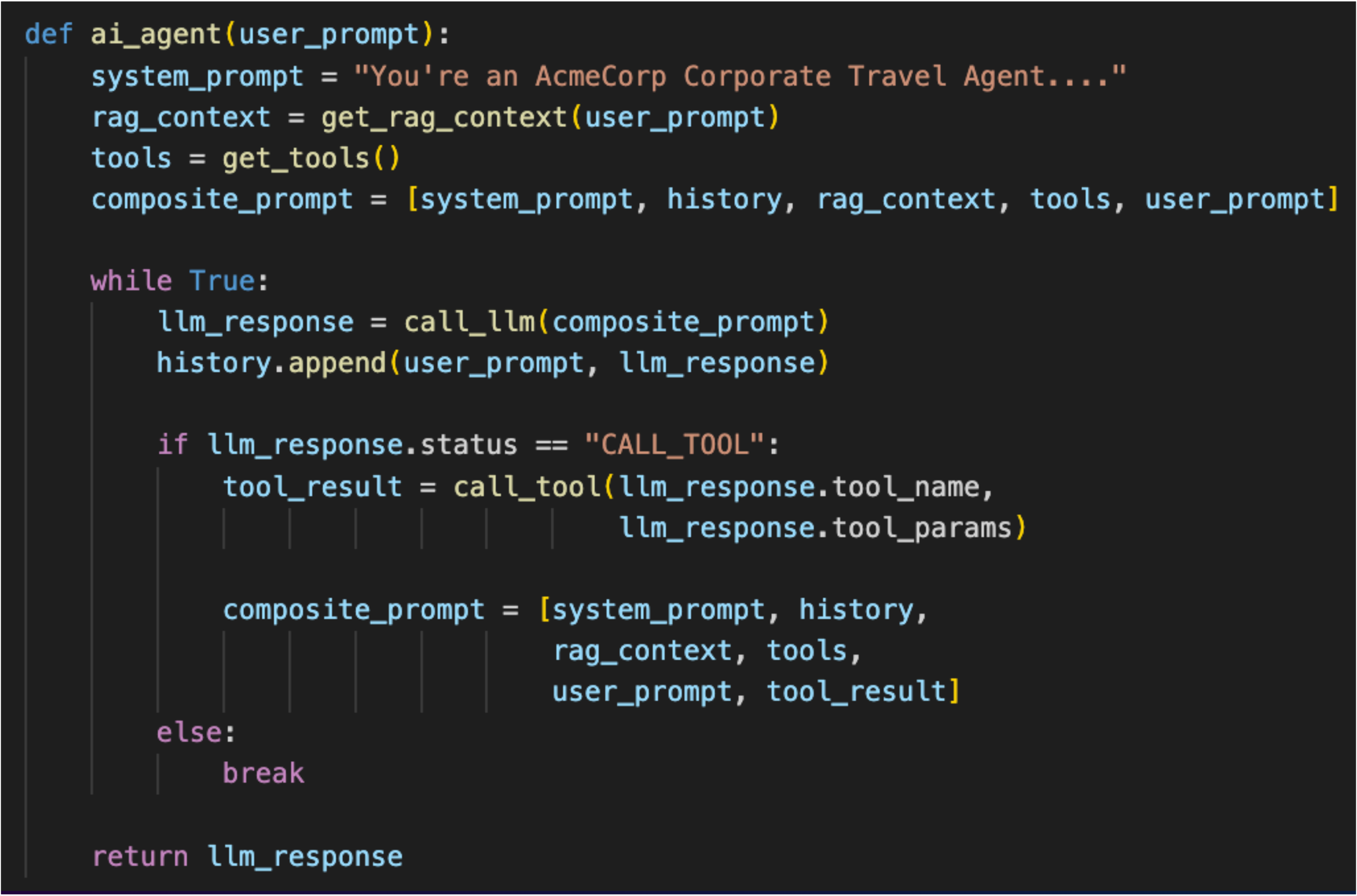
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Let’s build an agentic chatbot

**I want to build a Travel Agent that helps my employees to book business travel while enforcing corporate travel policies.**

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Implementing a simple AI Agent workflow

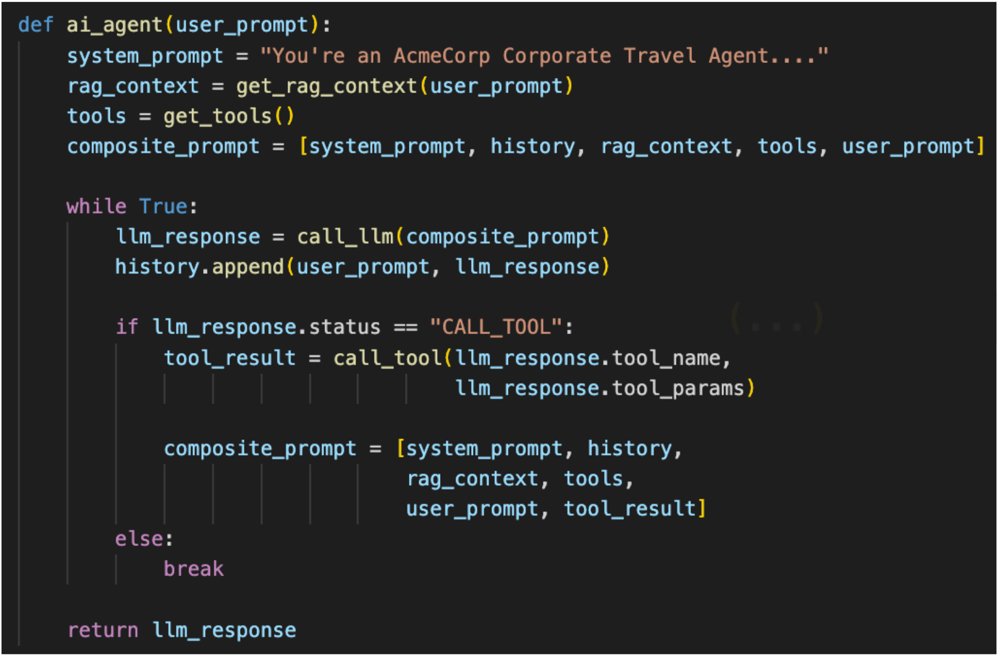


```
def ai_agent(user_prompt):
    system_prompt = "You're an AcmeCorp Corporate Travel Agent...."
    rag_context = get_rag_context(user_prompt)
    tools = get_tools()
    composite_prompt = [system_prompt, history, rag_context, tools, user_prompt]

    while True:
        llm_response = call_llm(composite_prompt)
        history.append(user_prompt, llm_response)

        if llm_response.status == "CALL_TOOL":
            tool_result = call_tool(llm_response.tool_name,
                                    llm_response.tool_params)

            composite_prompt = [system_prompt, history,
                                rag_context, tools,
                                user_prompt, tool_result]
        else:
            break

    return llm_response
```

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Implementing a simple AI Agent workflow

# BOILERPLATE



```python
def ai_agent(user_prompt):
    system_prompt = "You're an AcmeCorp Corporate Travel Agent...."
    rag_context = get_rag_context(user_prompt)
    tools = get_tools()
    composite_prompt = [system_prompt, history, rag_context, tools, user_prompt]

    while True:
        llm_response = call_llm(composite_prompt)
        history.append(user_prompt, llm_response)

        if llm_response.status == "CALL_TOOL":
            tool_result = call_tool(llm_response.tool_name,
                                    llm_response.tool_params)

            composite_prompt = [system_prompt, history,
                                rag_context, tools,
                                user_prompt, tool_result]
        else:
            break

    return llm_response
```

# CODE

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Strands Agents SDK - Built for builders

## Who value flexibility, speed, and simplicity

### Model & deployment choice

- ✓ Model choice
- ✓ Custom model providers
- ✓ Deploy anywhere

### Highly flexible

- ✓ Safeguard with guardrails
- ✓ Native observability
- ✓ Monitoring
- ✓ Evaluation

### Broad selection of tools

- ✓ MCP integration
- ✓ Custom tools
- ✓ Coordinate multiple agents
- ✓ Multi-modal
- ✓ Fetch web data
- ✓ Read and write files
- ✓ Interpret code

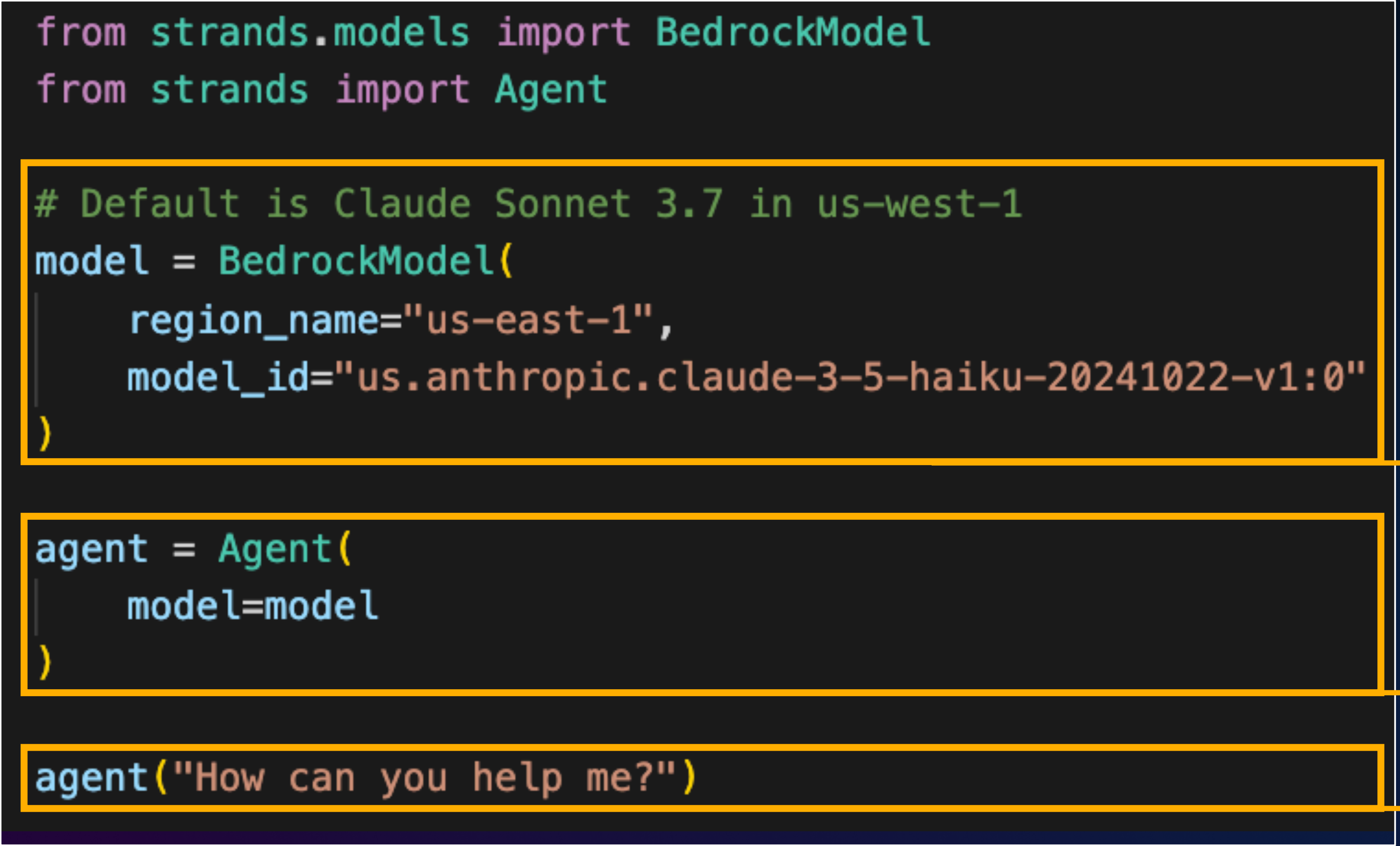
### Integrations

- ✓ Use AWS services
- ✓ LiteLLM
- ✓ Mem0
- ✓ RAGAS
- ✓ Tavily
- ✓ Langfuse

https://strandsagents.com/latest/

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Creating your first agent



```python
from strands.models import BedrockModel
from strands import Agent

# Default is Claude Sonnet 3.7 in us-west-1
model = BedrockModel(
    region_name="us-east-1",
    model_id="us.anthropic.claude-3-5-haiku-20241022-v1:0"
)

agent = Agent(
    model=model
)

agent("How can you help me?")
```

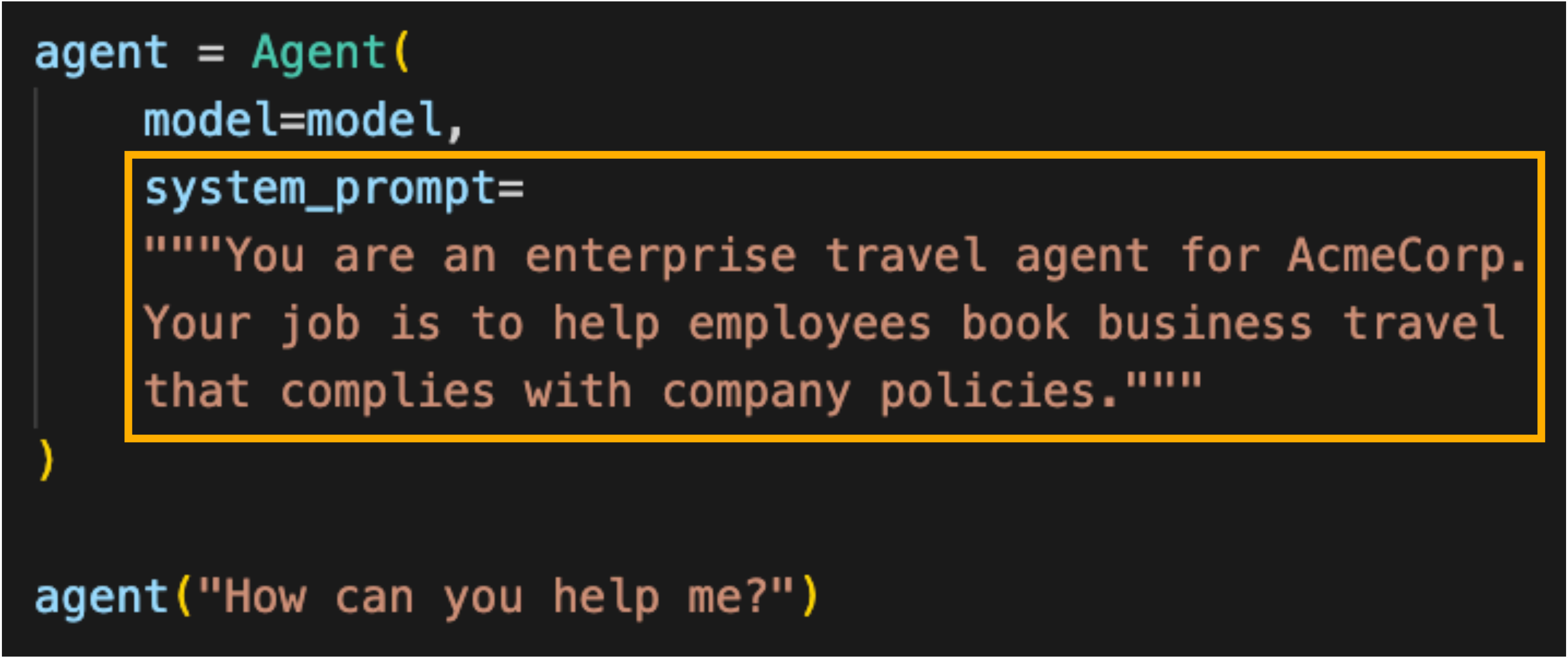
Configure model

Create agent

Prompt

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

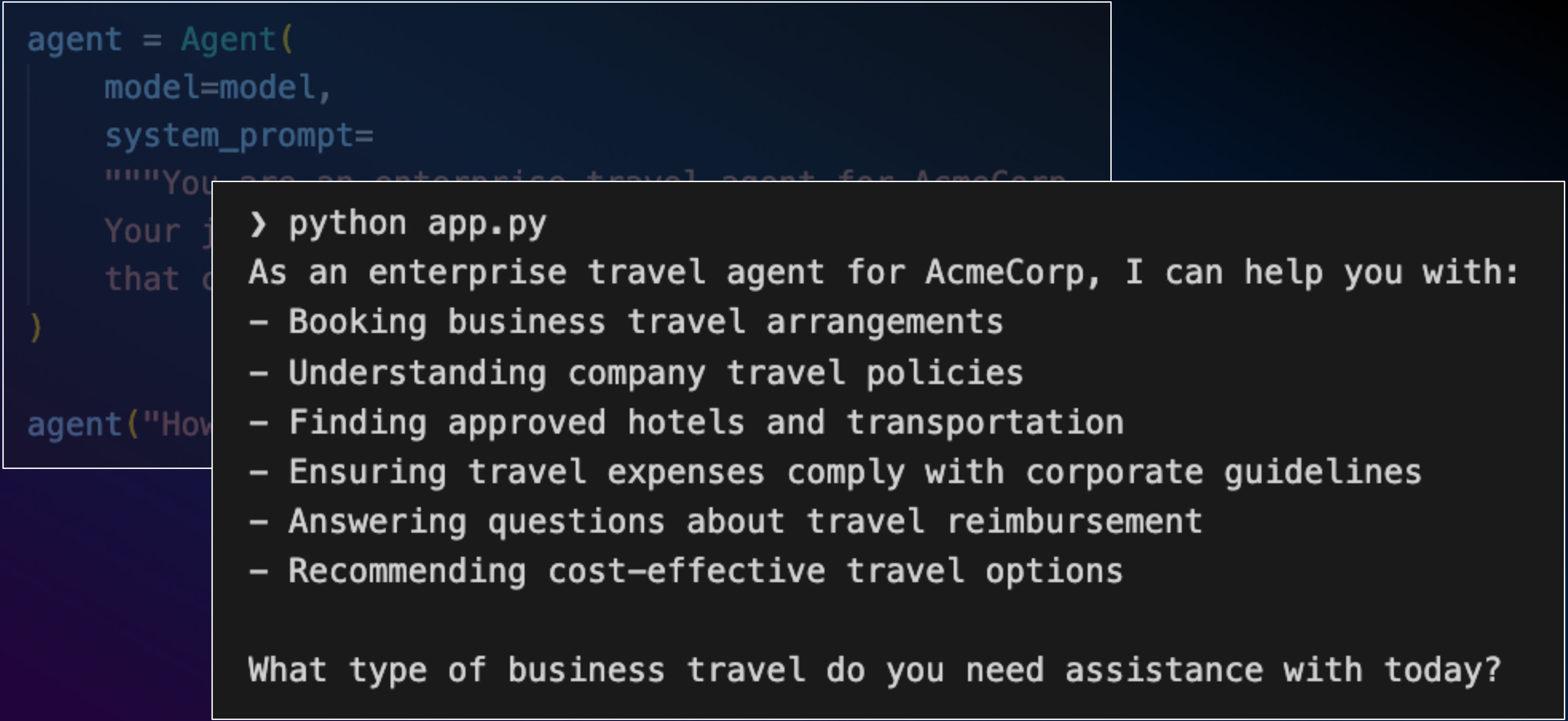
# Adding system prompt



```
agent = Agent(
    model=model,
    system_prompt=
    """You are an enterprise travel agent for AcmeCorp.
    Your job is to help employees book business travel
    that complies with company policies."""
)

agent("How can you help me?")
```

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Adding system prompt

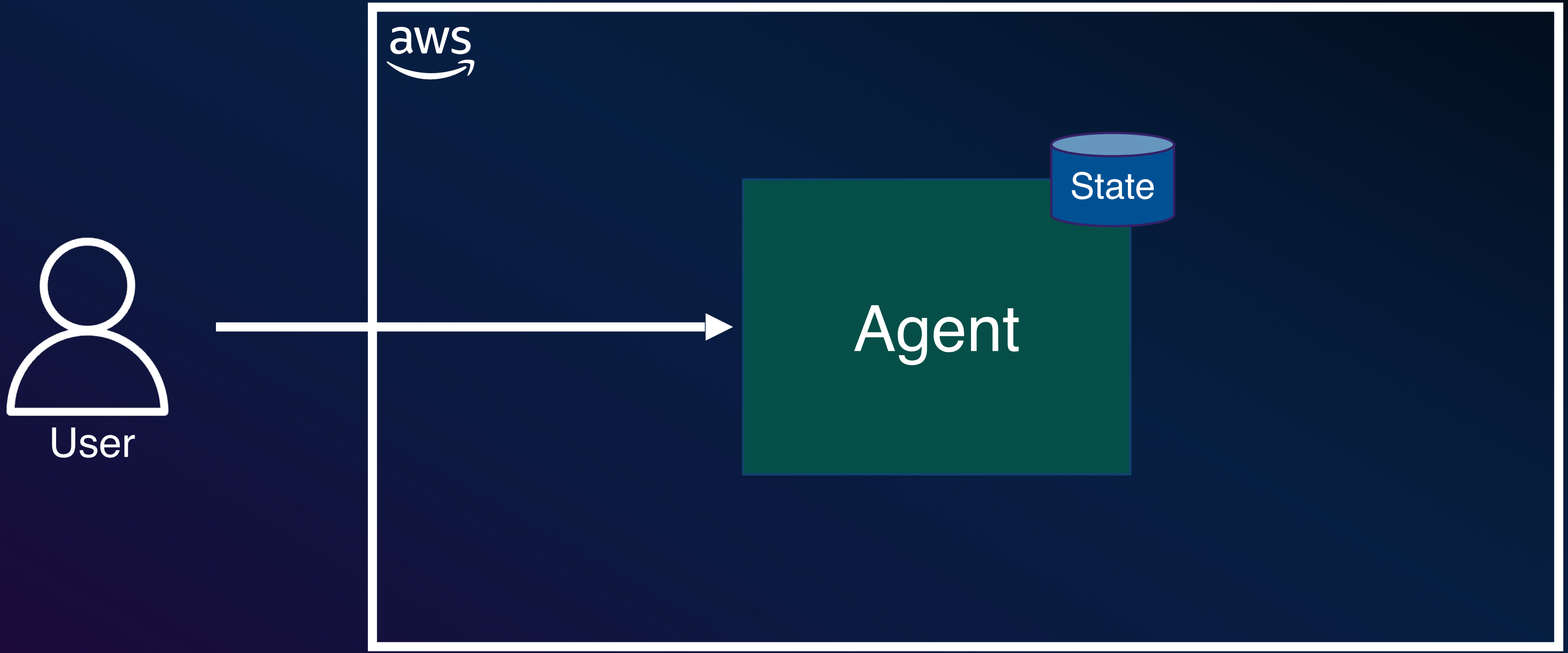


```python
agent = Agent(
    model=model,
    system_prompt=
    """You are an enterprise travel agent for AcmeCorp
    Your j
    that c
)


agent("How
```

```
❯ python app.py
As an enterprise travel agent for AcmeCorp, I can help you with:
- Booking business travel arrangements
- Understanding company travel policies
- Finding approved hotels and transportation
- Ensuring travel expenses comply with corporate guidelines
- Answering questions about travel reimbursement
- Recommending cost-effective travel options

What type of business travel do you need assistance with today?
```

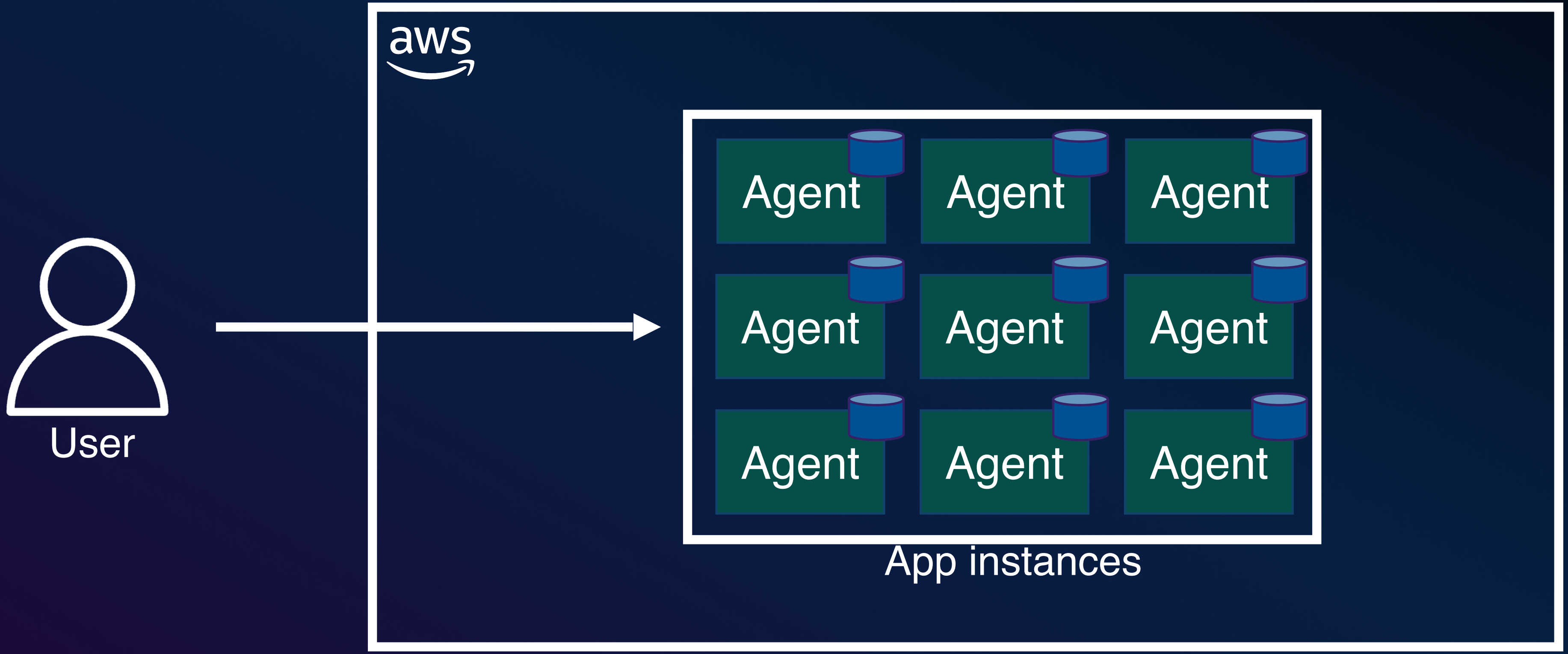
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# The problem with stateful apps



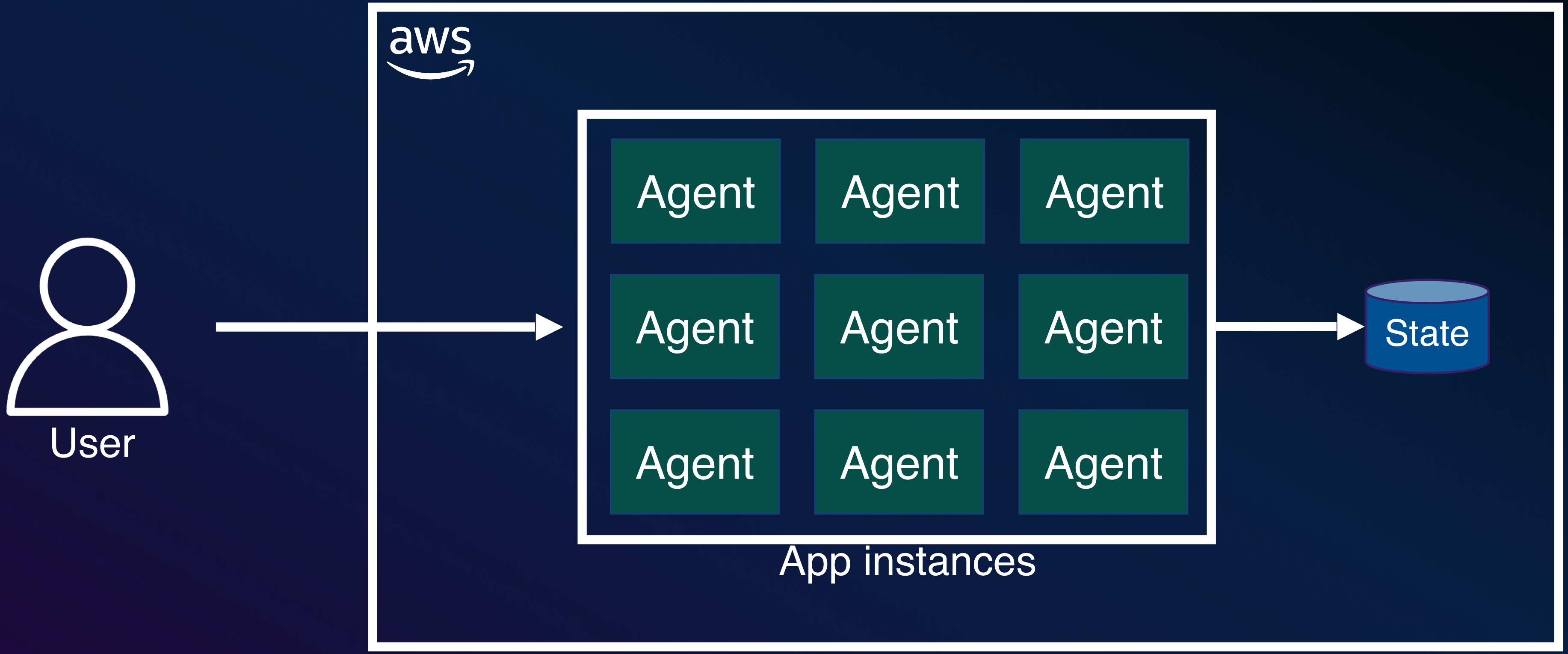
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# The problem with stateful apps



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Externalizing the state



User

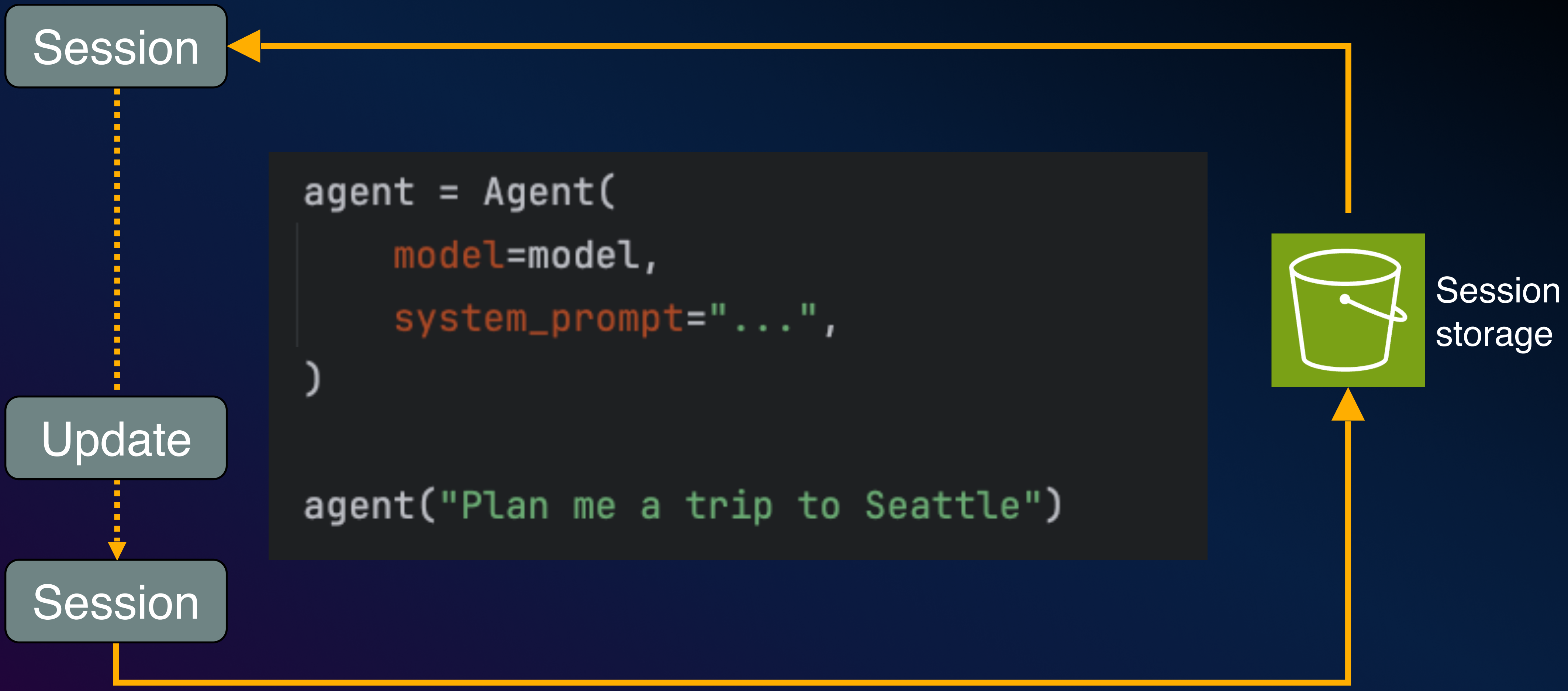
Agent Agent Agent
Agent Agent Agent
Agent Agent Agent

App instances

State

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Externalizing the session



Session

Update

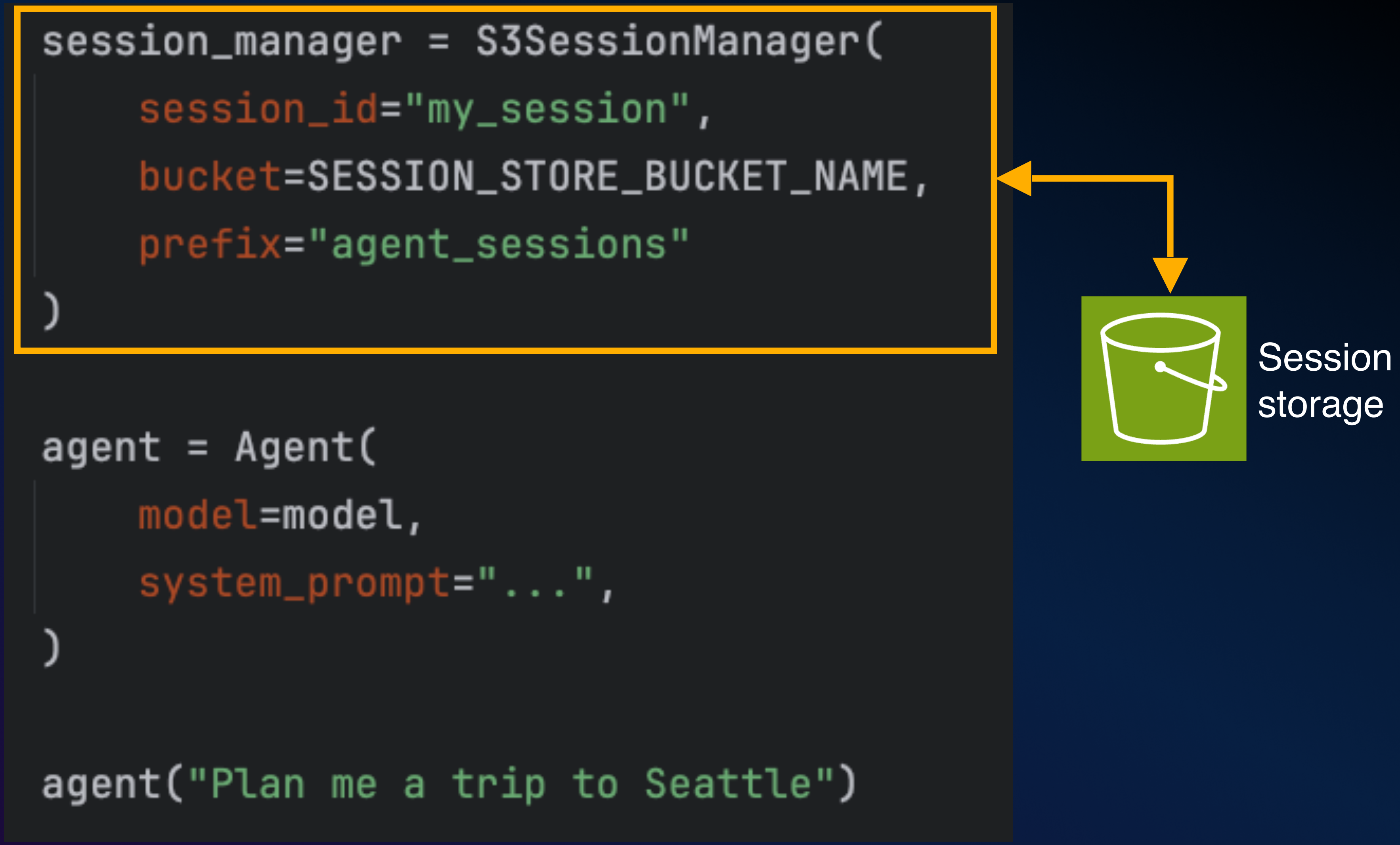
Session

```
agent = Agent(
    model=model,
    system_prompt="...",
)

agent("Plan me a trip to Seattle")
```

Session storage

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

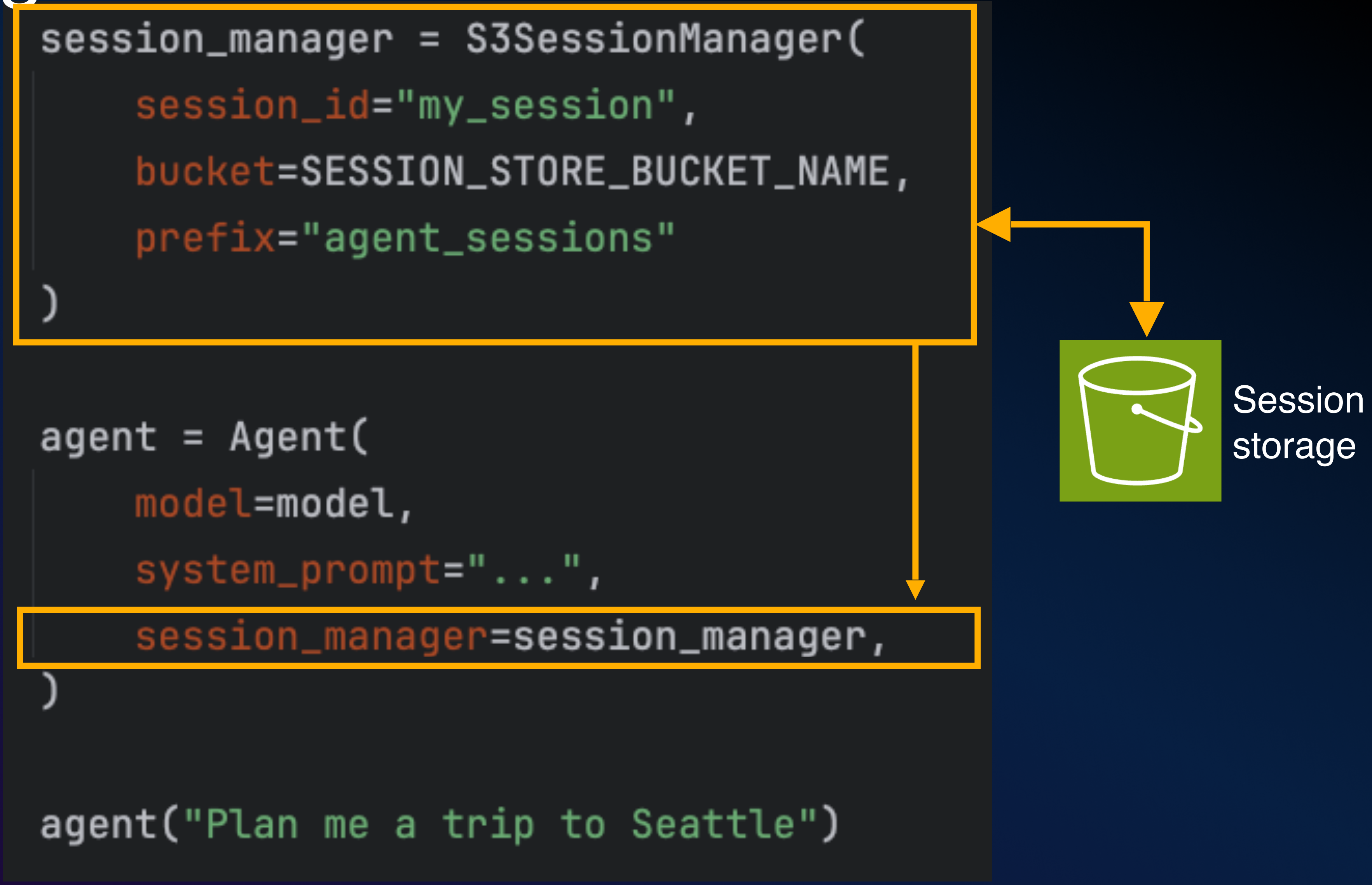
# Externalizing the session



```
session_manager = S3SessionManager(
    session_id="my_session",
    bucket=SESSION_STORE_BUCKET_NAME,
    prefix="agent_sessions"
)

agent = Agent(
    model=model,
    system_prompt="...",
)

agent("Plan me a trip to Seattle")
```

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

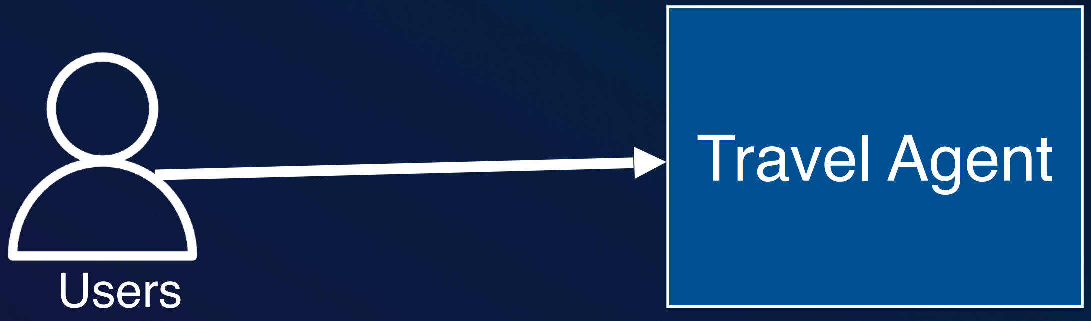
# Externalizing the session



```
session_manager = S3SessionManager(
    session_id="my_session",
    bucket=SESSION_STORE_BUCKET_NAME,
    prefix="agent_sessions"
)

agent = Agent(
    model=model,
    system_prompt="...",
    session_manager=session_manager,
)

agent("Plan me a trip to Seattle")
```

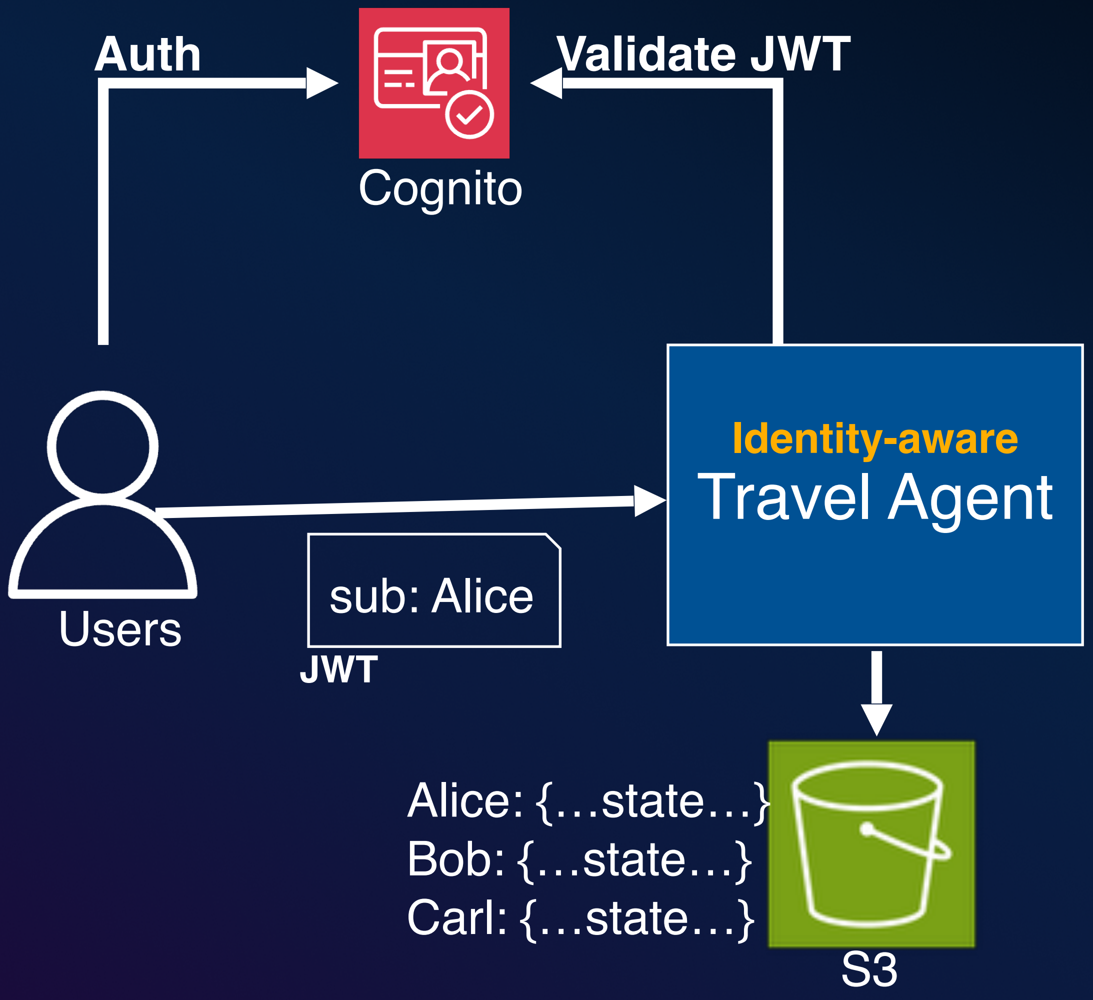
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# What's missing?



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Authentication and authorization



**Auth**

Cognito

**Validate JWT**

Users

sub: Alice

**JWT**

**Identity-aware**

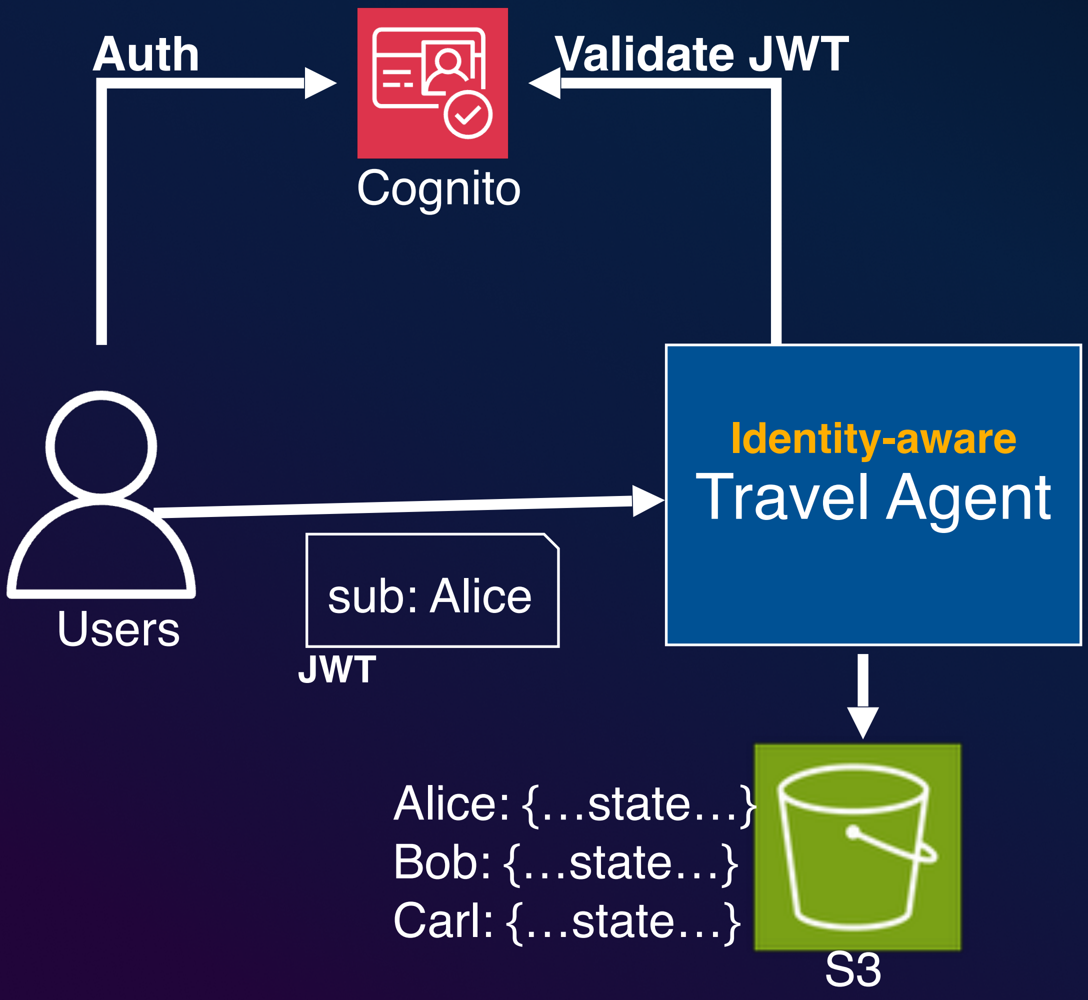
Travel Agent

Alice: {…state…}
Bob: {…state…}
Carl: {…state…}

S3

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Authentication and authorization



Auth

Cognito

Validate JWT

Users

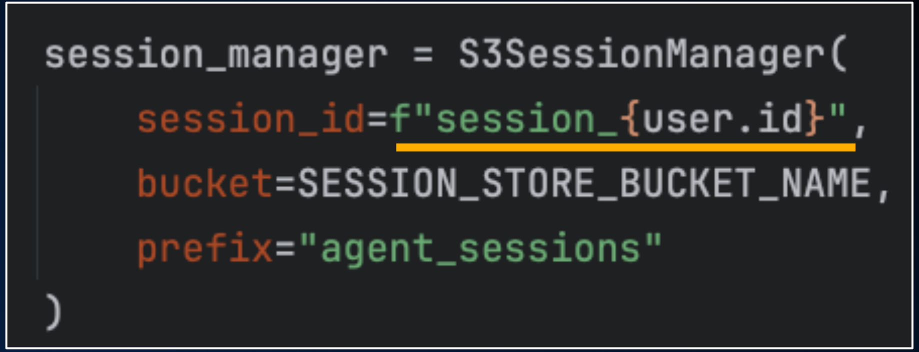
sub: Alice

JWT

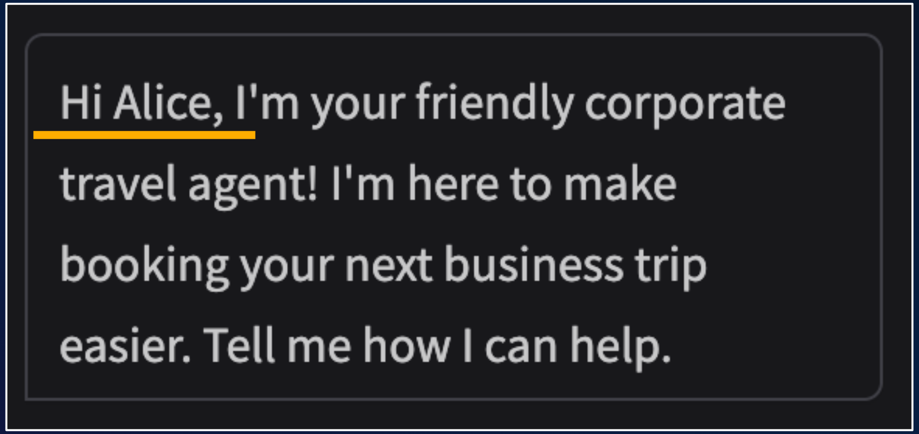
**Identity-aware**
Travel Agent

Alice: {…state…}
Bob: {…state…}
Carl: {…state…}

S3



```
session_manager = S3SessionManager(
    session_id=f"session_{user.id}",
    bucket=SESSION_STORE_BUCKET_NAME,
    prefix="agent_sessions"
)
```



Hi Alice, I'm your friendly corporate travel agent! I'm here to make booking your next business trip easier. Tell me how I can help.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Deploying to AWS

“Book me a business trip to NYC”

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

# Deploying to AWS

AWS Lambda

Amazon Elastic
Container Service (ECS)

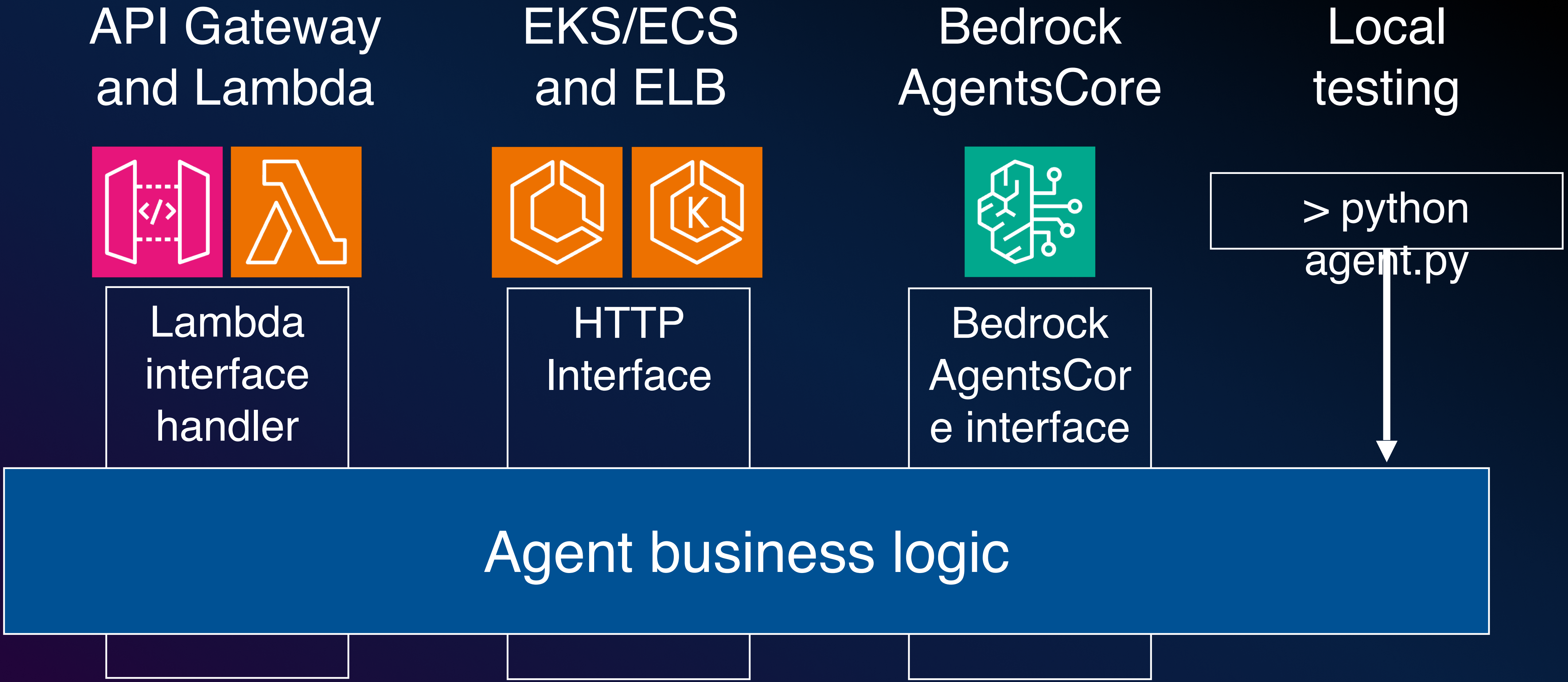
Amazon Elastic
Kubernetes Service (EKS)

Amazon Bedrock
AgentCore (preview)

**(or ANY other compute type, your Agent is just a Python app)**

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Building portable agents



API Gateway and Lambda

Lambda interface handler

EKS/ECS and ELB

HTTP Interface

Bedrock AgentsCore

Bedrock AgentsCore interface

Local testing

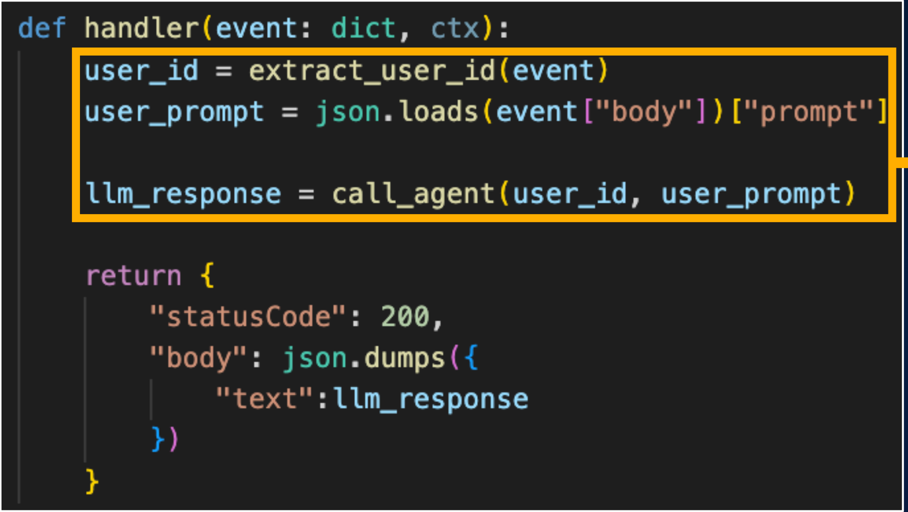
```
> python agent.py
```

Agent business logic

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

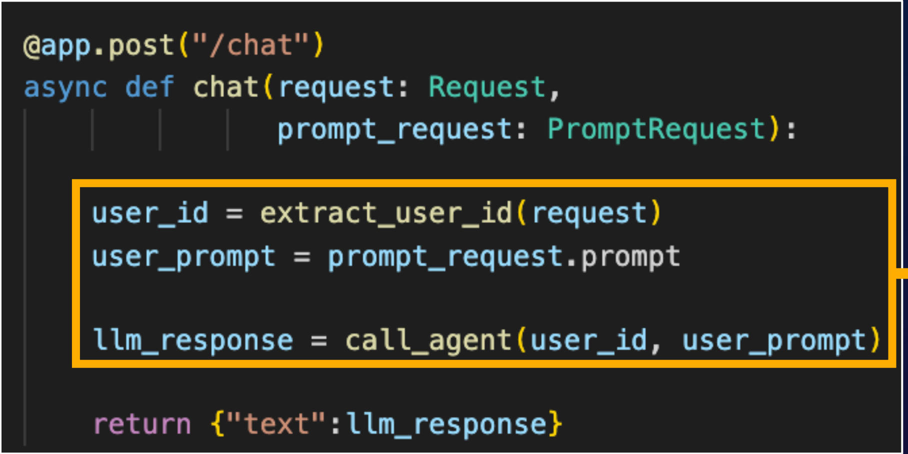
# Building portable agents

lambda_handler.py:



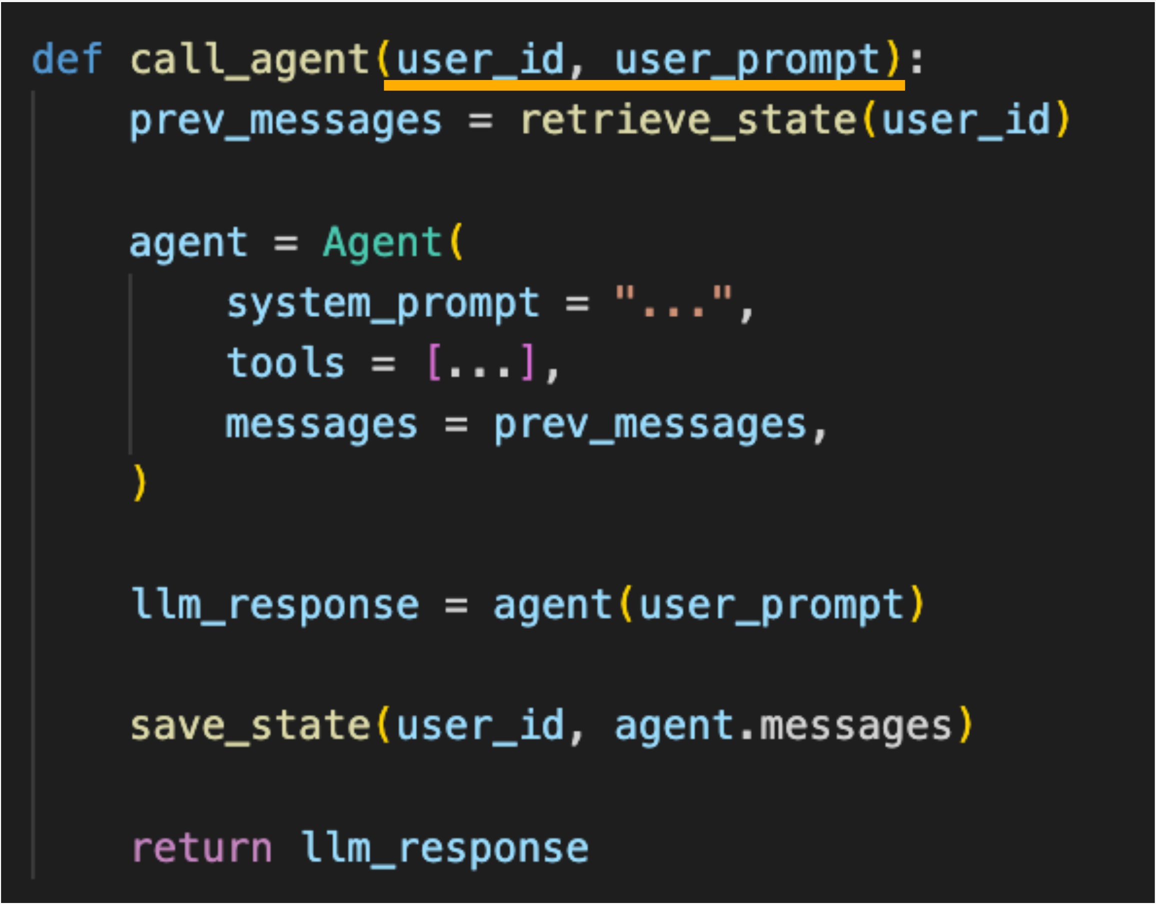
```
def handler(event: dict, ctx):
    user_id = extract_user_id(event)
    user_prompt = json.loads(event["body"])["prompt"]

    llm_response = call_agent(user_id, user_prompt)

    return {
        "statusCode": 200,
        "body": json.dumps({
            "text":llm_response
        })
    }
```

http_handler.py:



```
@app.post("/chat")
async def chat(request: Request,
               prompt_request: PromptRequest):

    user_id = extract_user_id(request)
    user_prompt = prompt_request.prompt

    llm_response = call_agent(user_id, user_prompt)

    return {"text":llm_response}
```

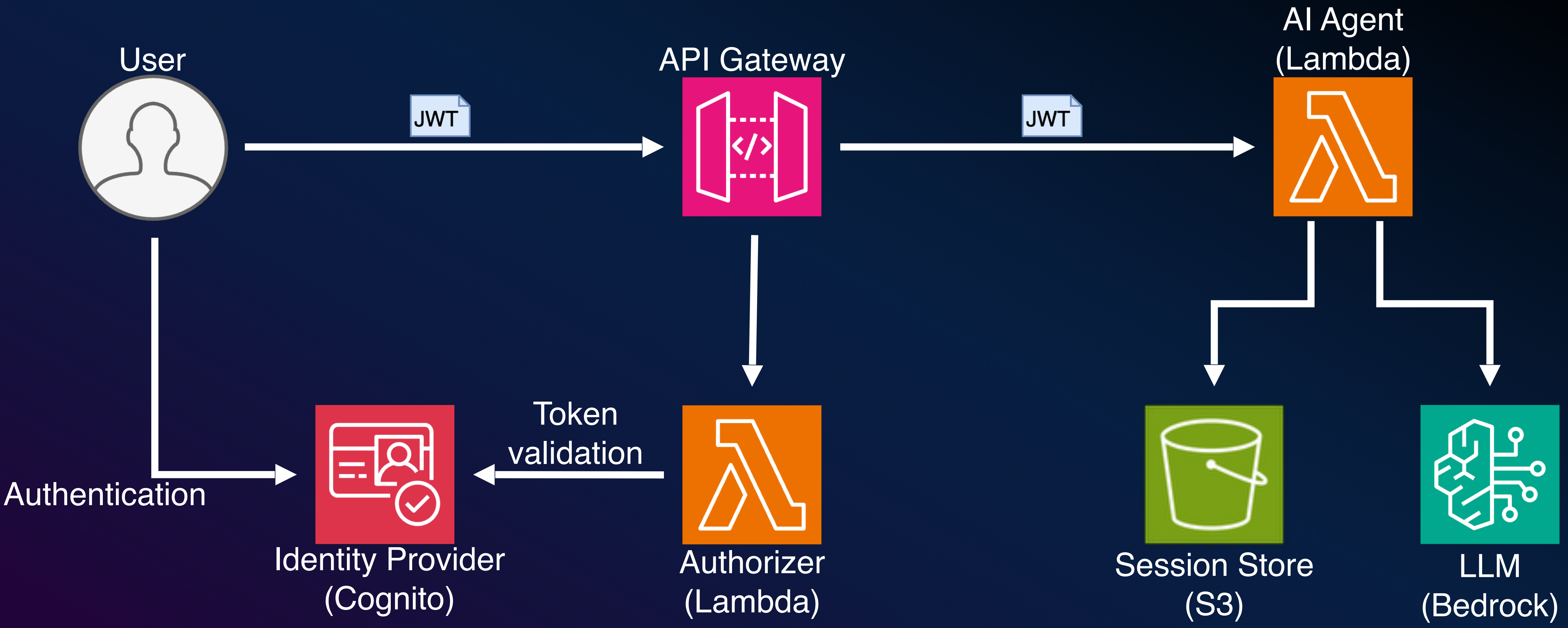
my_awesome_agent.py:



```
def call_agent(user_id, user_prompt):
    prev_messages = retrieve_state(user_id)

    agent = Agent(
        system_prompt = "...",
        tools = [...],
        messages = prev_messages,
    )

    llm_response = agent(user_prompt)

    save_state(user_id, agent.messages)

    return llm_response
```

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Deploying to AWS



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Deploying to AWS

But what about observability?
Security? Governance? CI/CD?

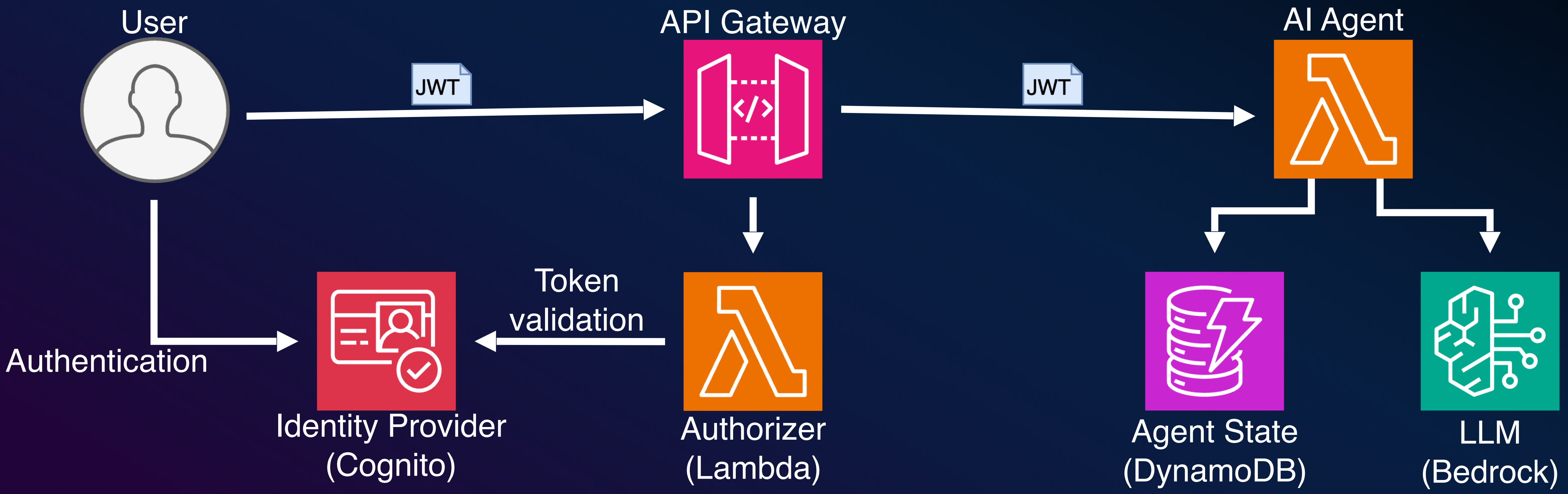
---

**Everything you know about building applications and APIs on AWS is still applicable!**

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Deploying to AWS

IAM Cert. Manager KMS CloudWatch X-Ray GuardDuty Inspector WAF CloudTrail VPC



User

JWT

API Gateway

JWT

AI Agent

Authentication

Token validation

Identity Provider (Cognito)

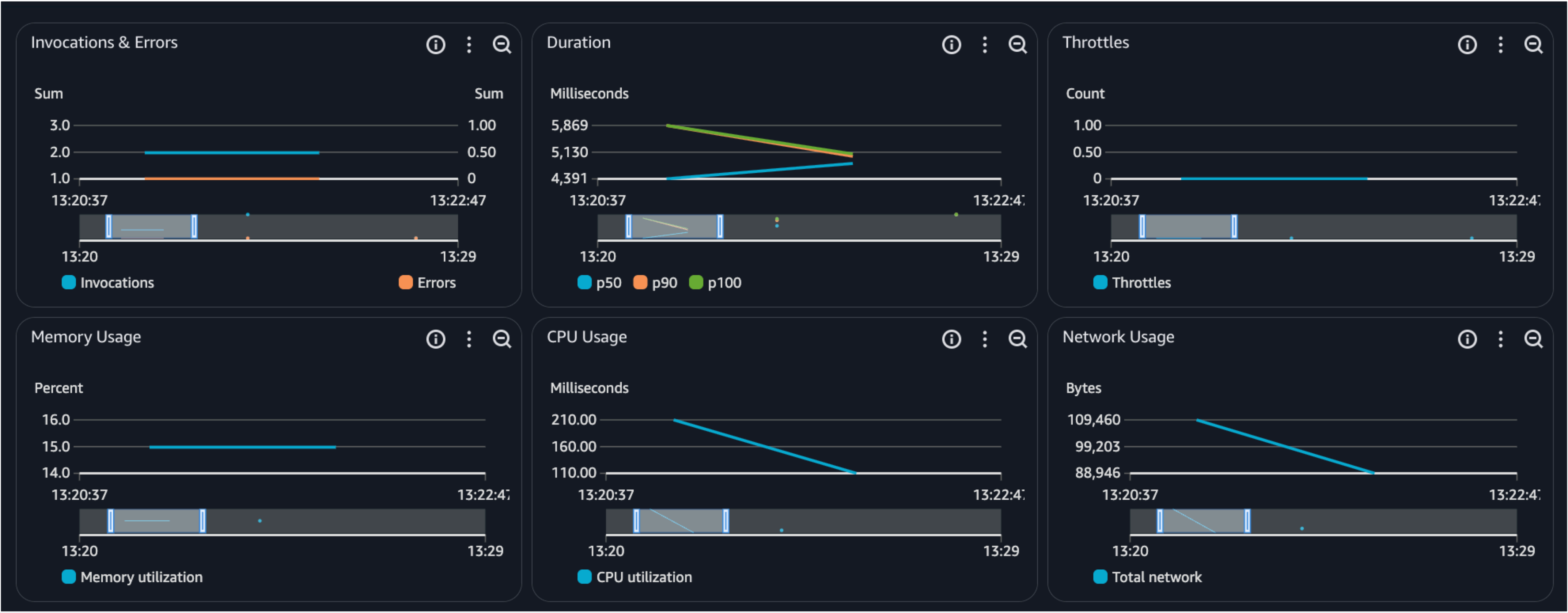
Authorizer (Lambda)

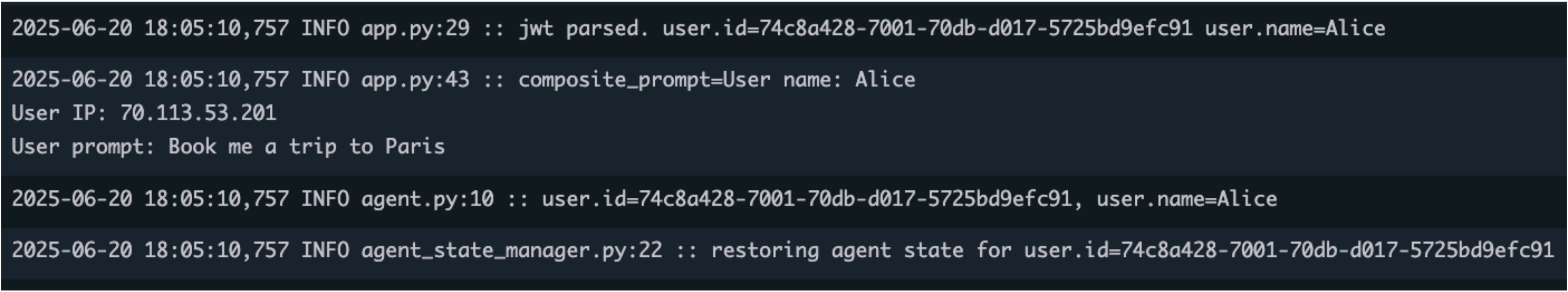
Agent State (DynamoDB)

LLM (Bedrock)

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Observability

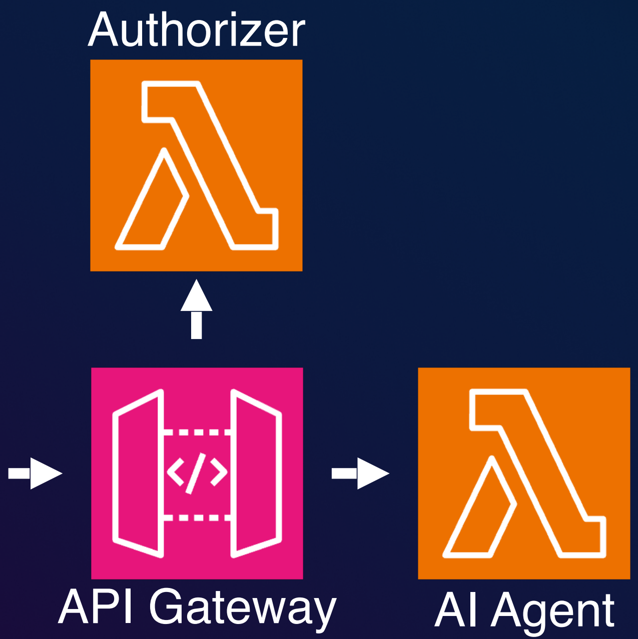


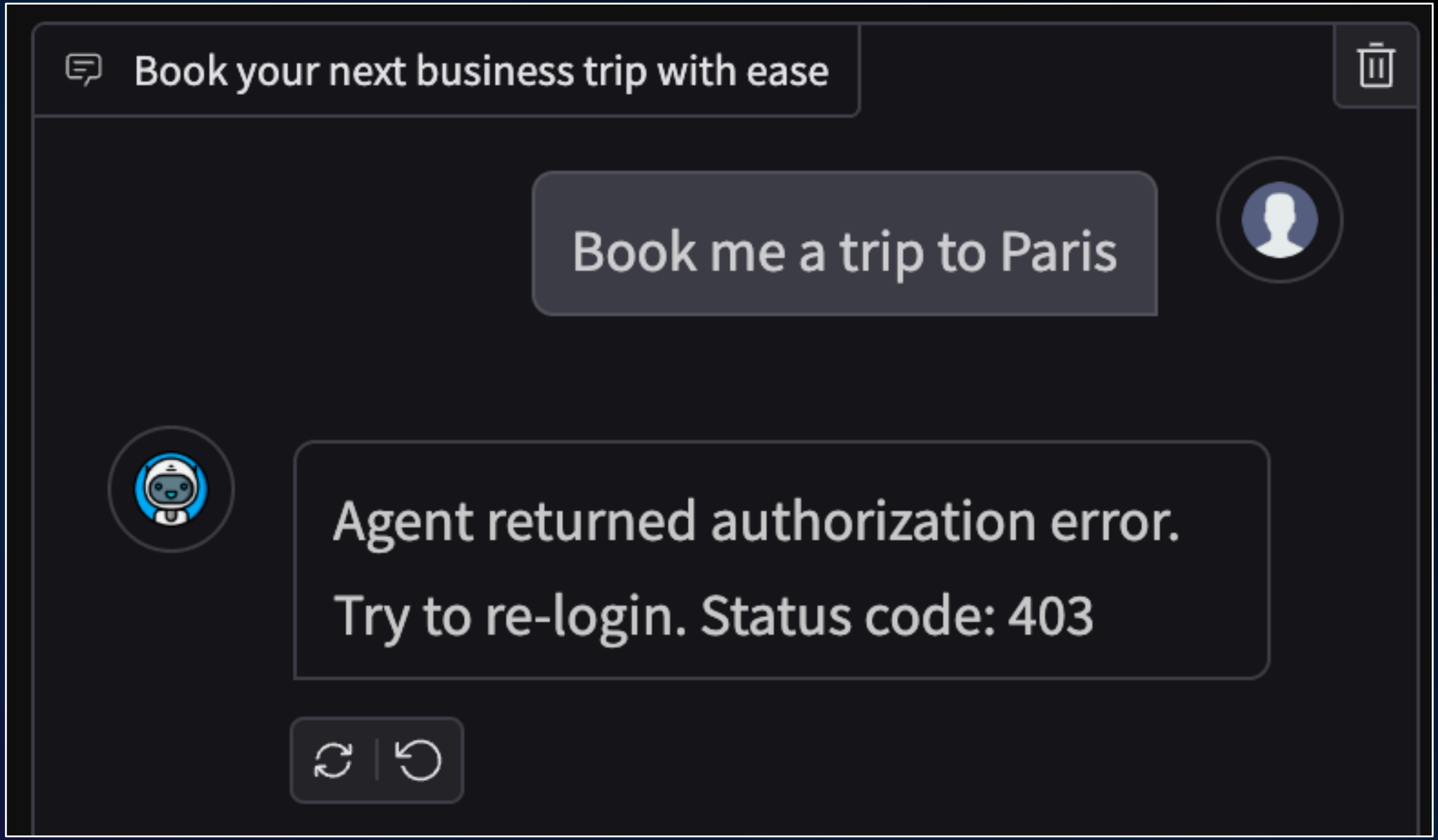


```
2025-06-20 18:05:10,757 INFO app.py:29 :: jwt parsed. user.id=74c8a428-7001-70db-d017-5725bd9efc91 user.name=Alice

2025-06-20 18:05:10,757 INFO app.py:43 :: composite_prompt=User name: Alice
User IP: 70.113.53.201
User prompt: Book me a trip to Paris

2025-06-20 18:05:10,757 INFO agent.py:10 :: user.id=74c8a428-7001-70db-d017-5725bd9efc91, user.name=Alice

2025-06-20 18:05:10,757 INFO agent_state_manager.py:22 :: restoring agent state for user.id=74c8a428-7001-70db-d017-5725bd9efc91
```

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Authorization





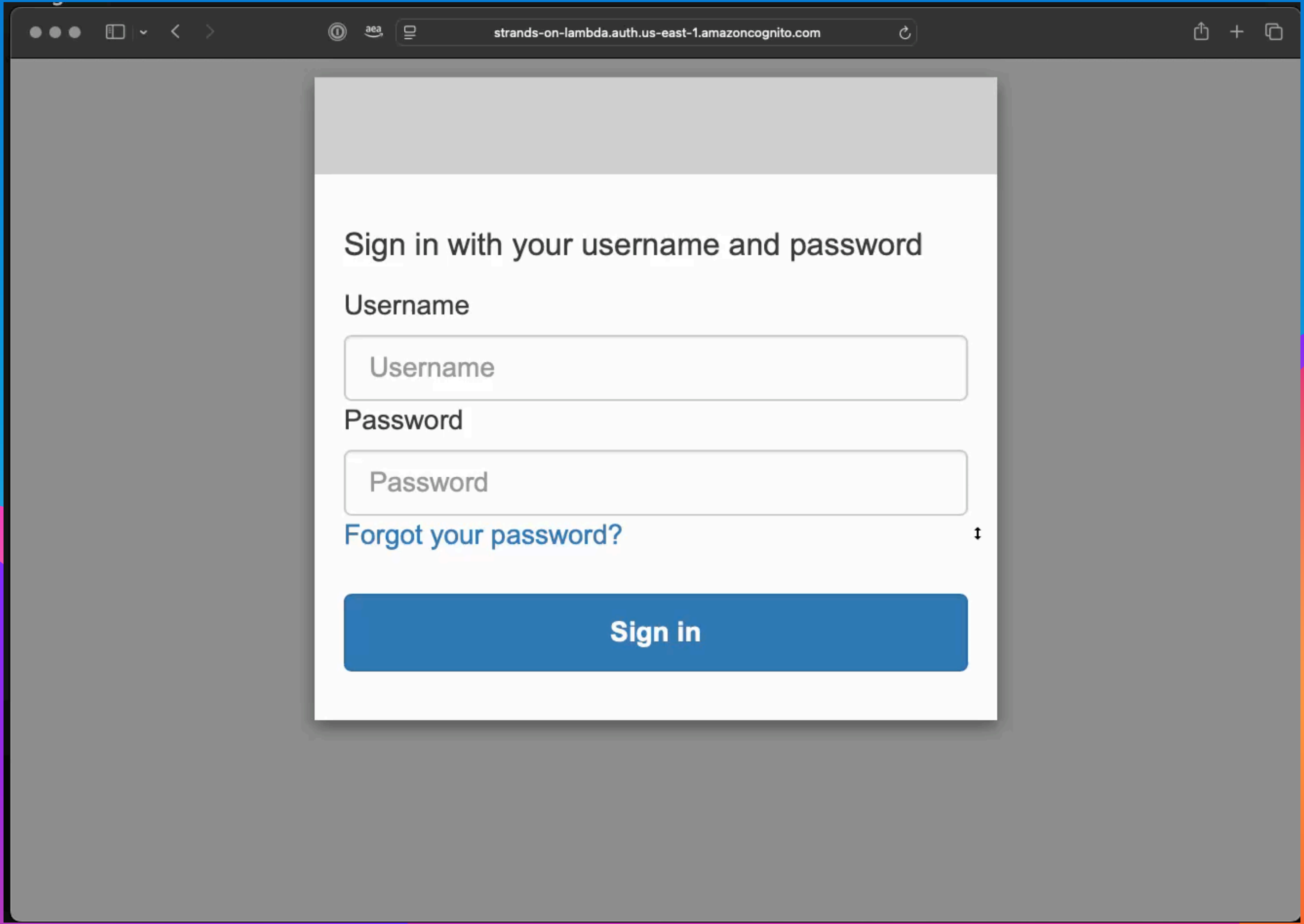
Book me a trip to Paris

Agent returned authorization error.

Try to re-login. Status code: 403

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Demo time



strands-on-lambda.auth.us-east-1.amazoncognito.com

Sign in with your username and password

Username

Password

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Conclusion

1

Agents are applications calling LLMs in loops. Your existing knowledge is your accelerator.

2

Use the right tool for the job. Decouple when needed, use proper protocols.

3

AWS services provide built-in enterprise-grade capabilities for building AI agents.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Continue your learning journey

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!

Please complete the session survey in the mobile app

**Anton Aleksandrov**
in antonal80
aal80.github.io/whoami

**Debasis Rath**
in rathd

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.