AWS re:Invent

DECEMBER 1 – 5, 2025 | LAS VEGAS, NV

AMZ402

Evaluating AI agents: real-world lessons from Amazon's agent systems

Yunfei Bai

He/him

Principal Solutions Architect

Amazon Web Services

Kashif Imran

He/him

Senior Manager, Cloud/Applied AI Architecture

Amazon Web Services

Allie Colin

She/her

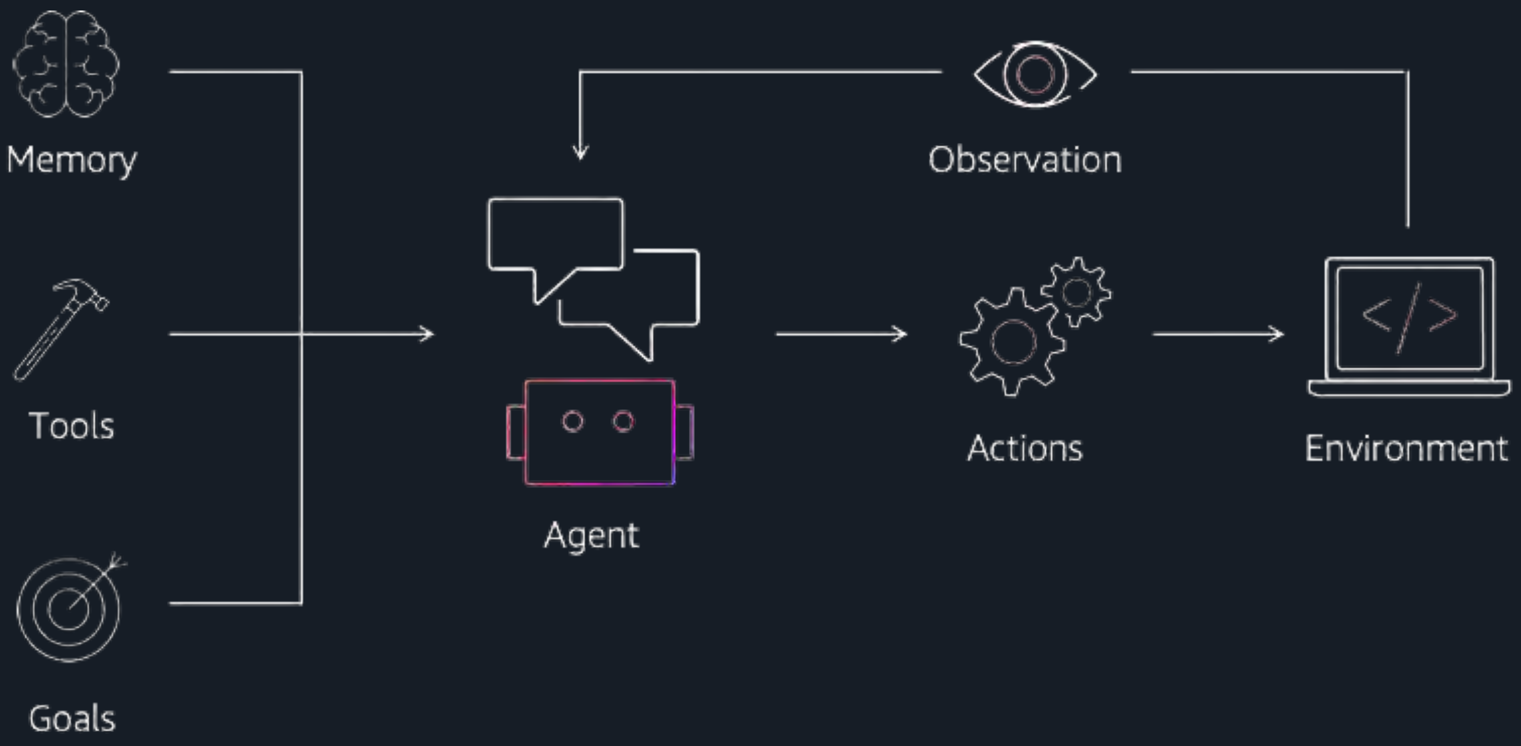
Senior Manager, Head of Product and Science

Amazon

Agenda

- AI agent evaluation at Amazon
- Example 1: evaluating agent tool-use
- Example 2: evaluating agent reasoning
- Example 3: evaluating multi-agent system
- Key takeaways

AI agent evaluation



Evaluation strategy

- Task completion
- Planning/multi-step reasoning
- Function call and tool-use
- Memory management
- Operations, costs and RAI
- App-specific agent

Common challenges in AI agent evaluation

Real-world performance

Black box frustration

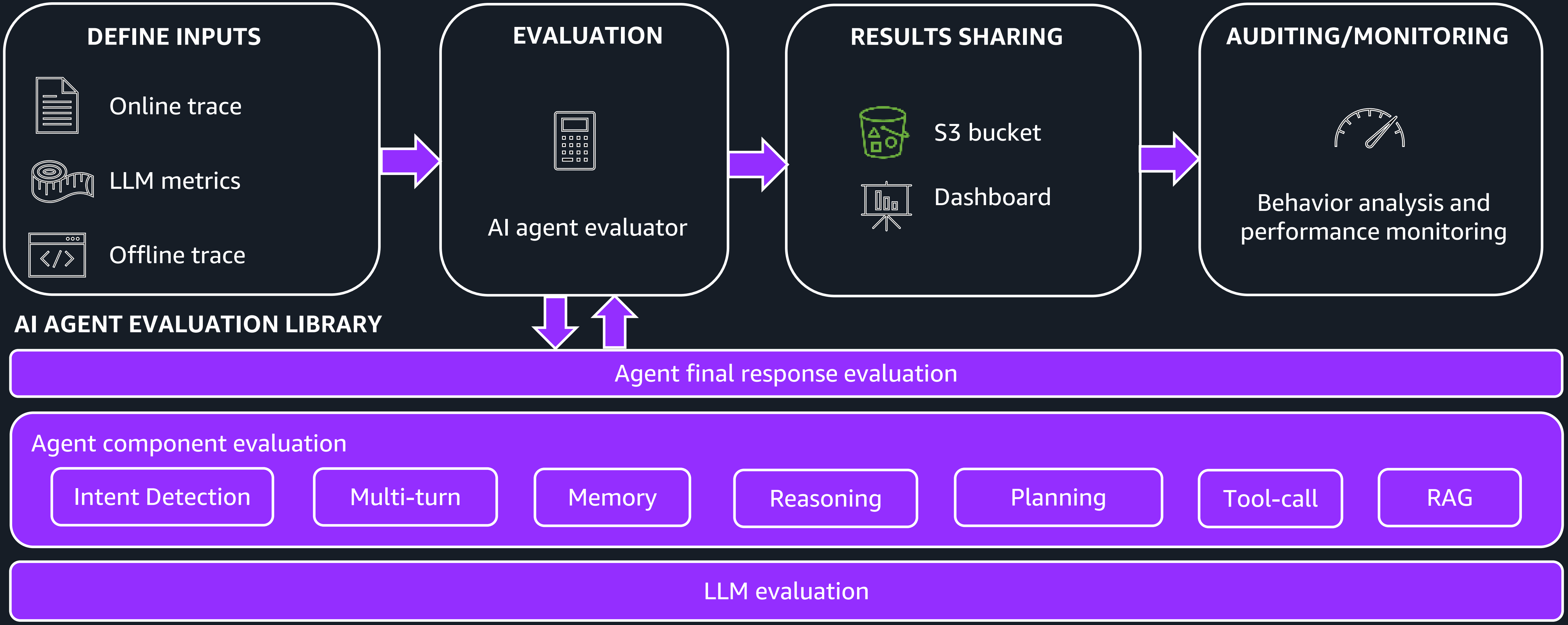
Complexity overwhelm

Framework lock-in

Performance monitoring

Evaluation data quality

AI agent evaluation at Amazon



DEFINE INPUTS

- Online trace
- LLM metrics
- Offline trace

EVALUATION

AI agent evaluator

RESULTS SHARING

- S3 bucket
- Dashboard

AUDITING/MONITORING

Behavior analysis and performance monitoring

AI AGENT EVALUATION LIBRARY

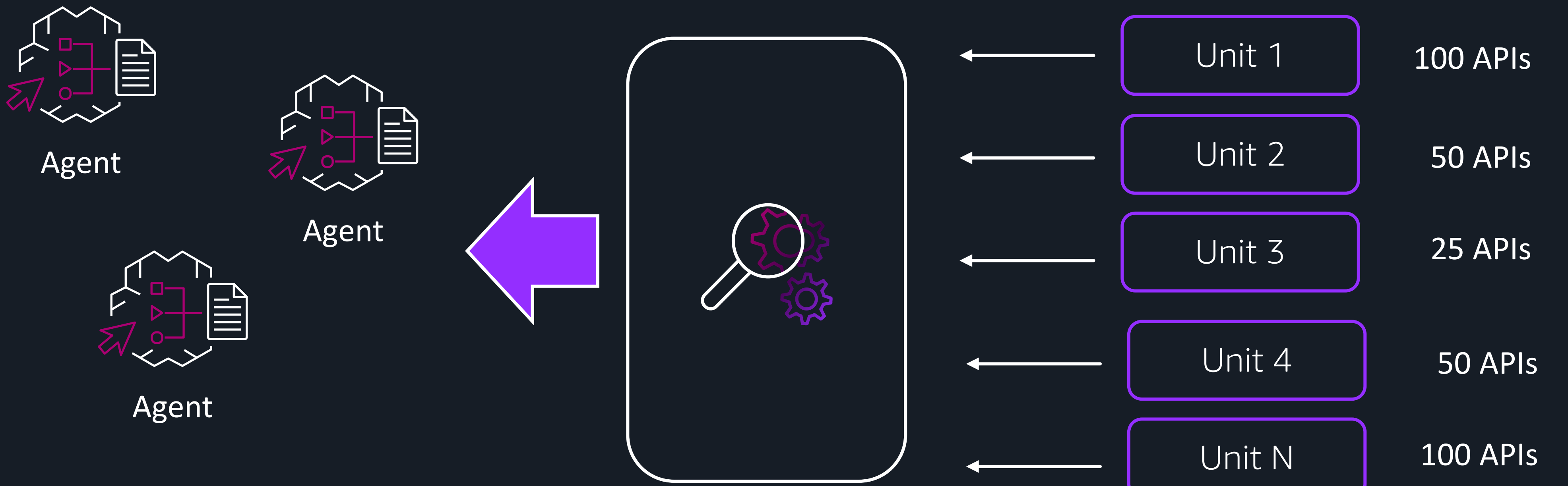
Agent final response evaluation

Agent component evaluation

Intent Detection | Multi-turn | Memory | Reasoning | Planning | Tool-call | RAG

LLM evaluation

Example 1: evaluating agent tool-use



Agent

Agent

Agent

Unit 1 100 APIs

Unit 2 50 APIs

Unit 3 25 APIs

Unit 4 50 APIs

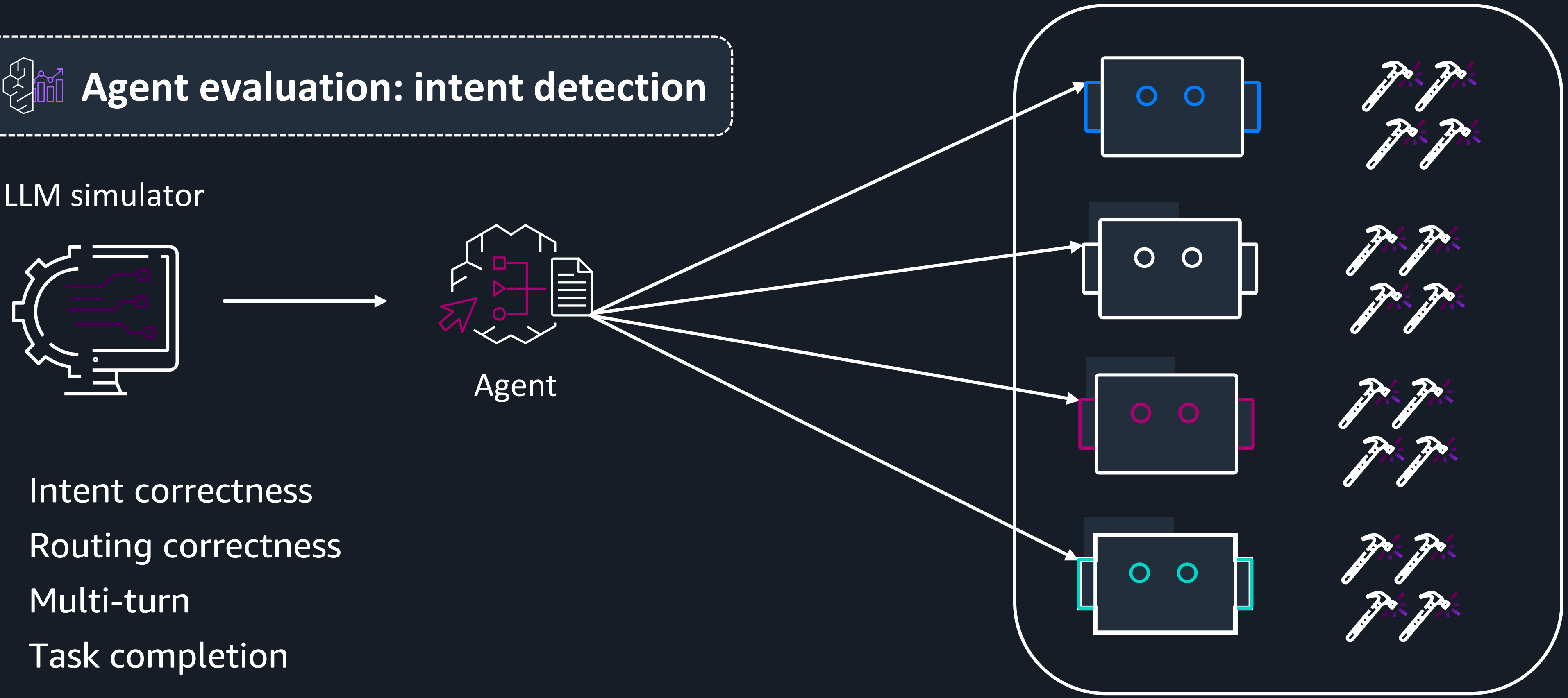
Unit N 100 APIs

Agent evaluation: tool-use

- Tool-call accuracy
- Response correctness
- Function relevance
- Multi-turn function call support

Example 2: evaluating agent reasoning

Agent evaluation: intent detection

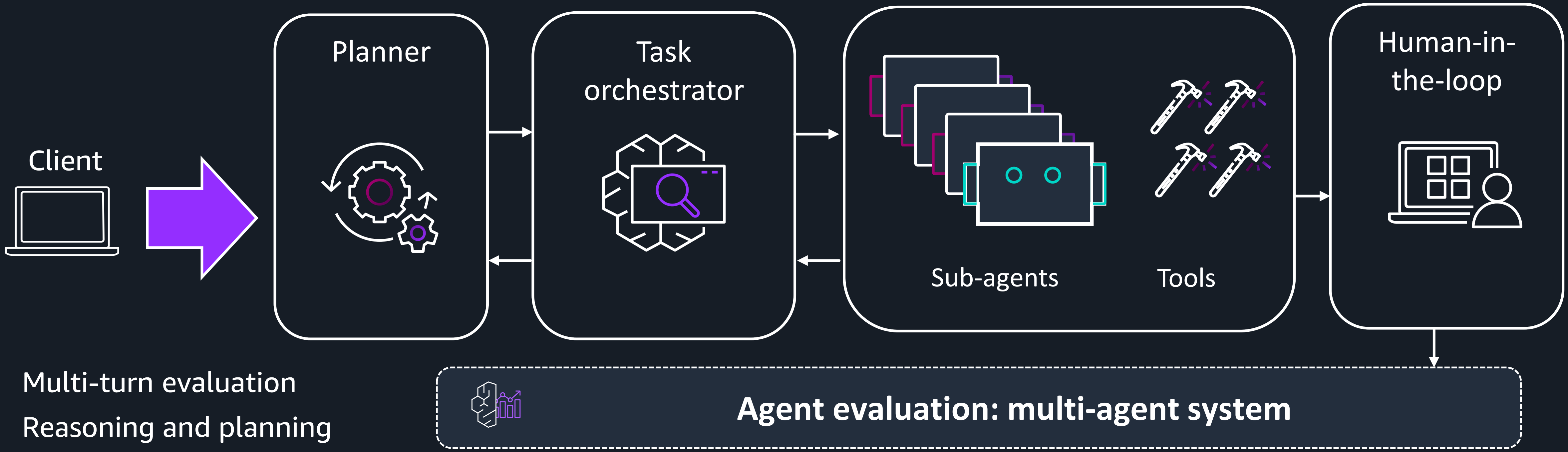


LLM simulator

Agent

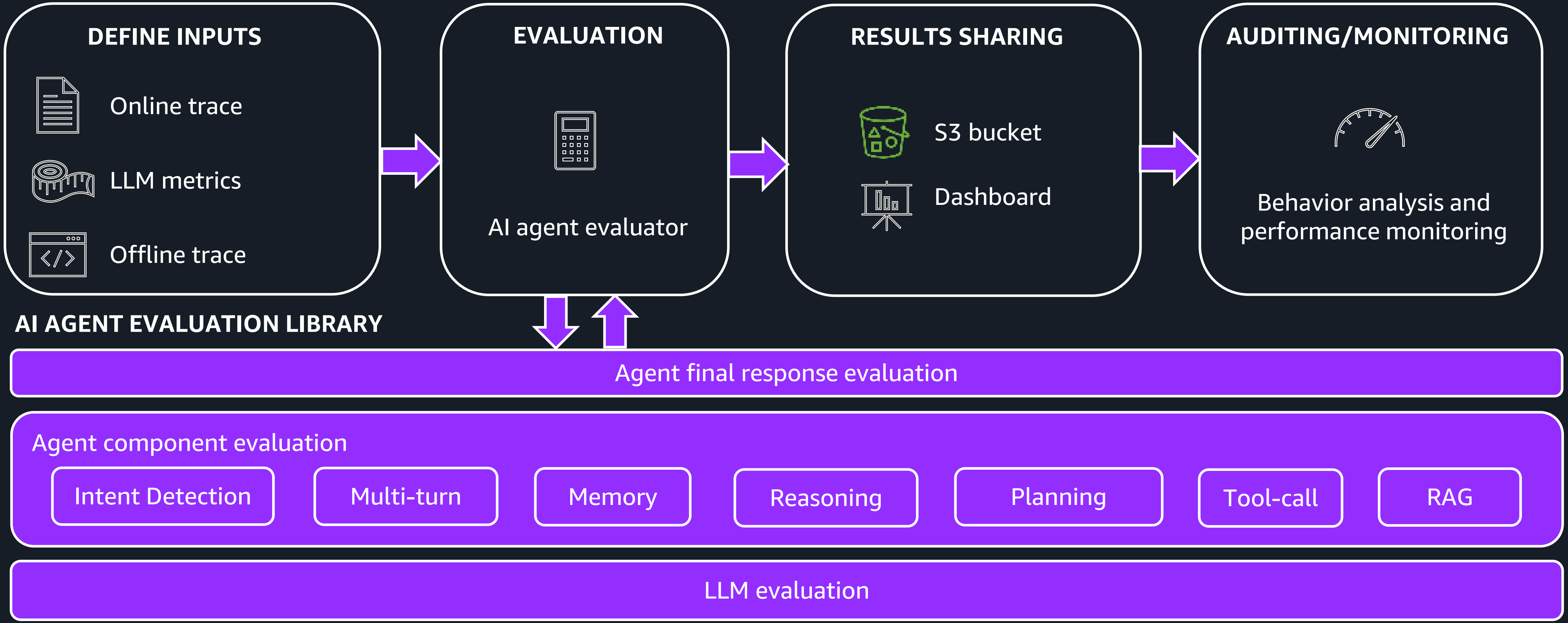
- Intent correctness
- Routing correctness
- Multi-turn
- Task completion

Example 3: evaluating multi-agent system



- Multi-turn evaluation
- Reasoning and planning
- Orchestration
- Sub-agent performance
- Human-in-the-loop

AI agent evaluation at Amazon



DEFINE INPUTS

Online trace

LLM metrics

Offline trace

EVALUATION

AI agent evaluator

RESULTS SHARING

S3 bucket

Dashboard

AUDITING/MONITORING

Behavior analysis and performance monitoring

AI AGENT EVALUATION LIBRARY

Agent final response evaluation

Agent component evaluation

Intent Detection

Multi-turn

Memory

Reasoning

Planning

Tool-call

RAG

LLM evaluation

Key takeaways

- Holistic evaluation on agent quality, performance, guardrail and responsibility
- Use case specific evaluation
- Combine automated evaluation with human-in-the-loop (HITL)
- Continuous evaluation

Thank you

Please complete the session survey in the mobile app