Enabling complex reasoning and action with ReAct, LLMs, and LangChain

Giuseppe Zappia
Principal Solutions Architect
Amazon Web Services

Shelbee Eigenbrode
Principal AI/ML Specialist Solutions Architect
Amazon Web Services

Agenda

- Langchain Overview
- ReAct Framework: High-Level Overview
- Workshop Introduction

LangChain Overview

LangChain Components

Component	Function
Document Loaders	Load and manipulate documents
Vector Stores	Store and query unstructured data through vectors
Prompt Templates	Build templates to optimize LLM queries
LLMs	Interfaces for LLMs
Chains	Combine LLMs and prompt templates to build workflows
Memory	State management of chains/agents to preserve context
Agents	Use LLMs to choose which activities to perform
Tools	Used by agents to perform a specific task (Google Search, DB lookups, etc.)

ReAct Framework: High Level Overview

If you were asked you the following question, how would you solve it?

What is the 4th largest planet in our solar system,
and how many Earths can fit inside it?

Things you need to know:

- Which planet is the 4^{th} largest in the solar system?
- What is the volume of that planet?
  - What's the radius of the planet?
  - What's the formula for the volume of a sphere?
- What is the volume of Earth?
  - What's the radius of Earth?
  - What's the formula for the volume of a sphere?
- What is the ratio of the Earth to that planet?

Wikipedia

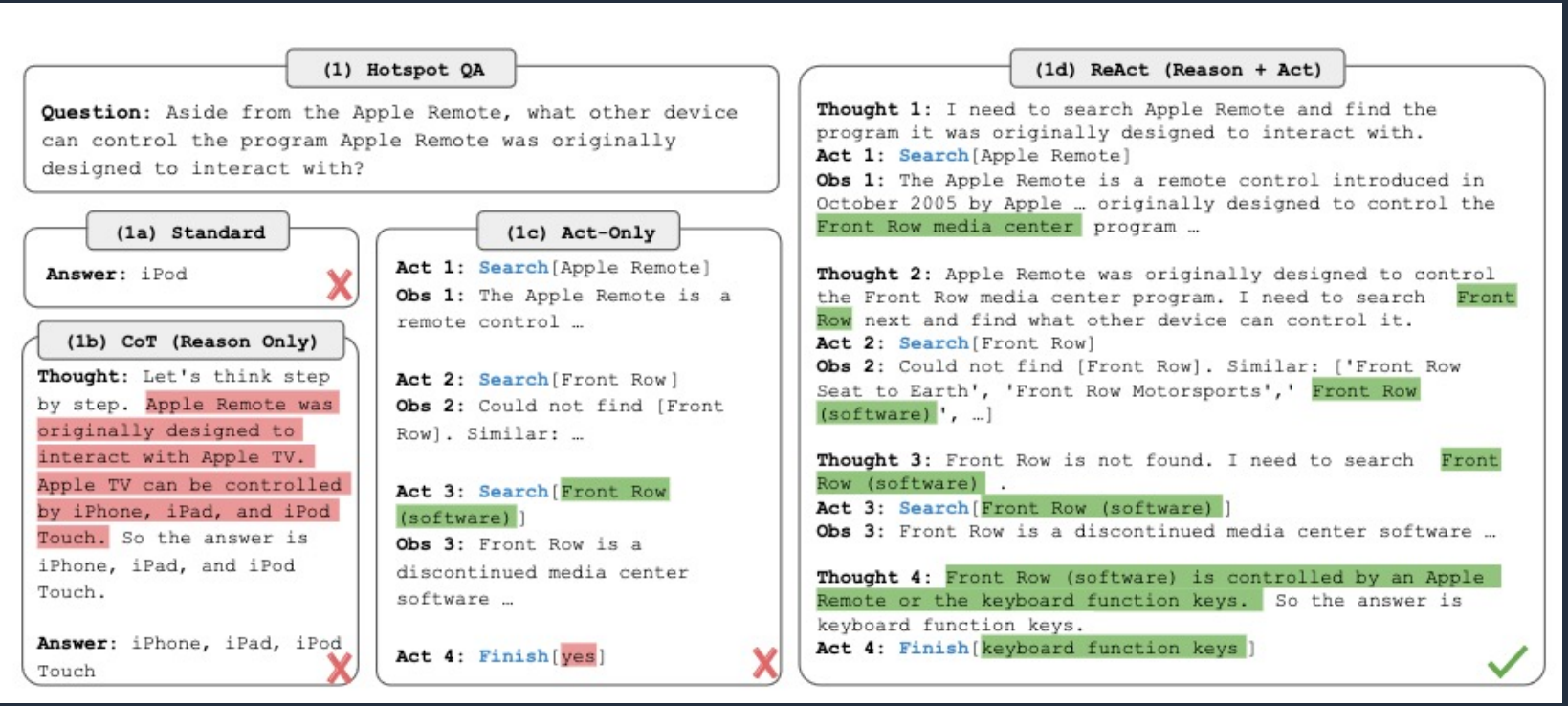
Calculations

Things you look up:

- Neptune
- $V = 6.253 \times 10^{13}\ km^3$
  - Radius = 24,622 km
  - $V = 4/3\ \pi\ r^3$
- $V = 1.08321 \times 10^{12}\ km^3$
  - Radius = 6,371 km
  - $V = 4/3\ \pi\ r^3$
- 62.53/1.08321 = ~ 57.7

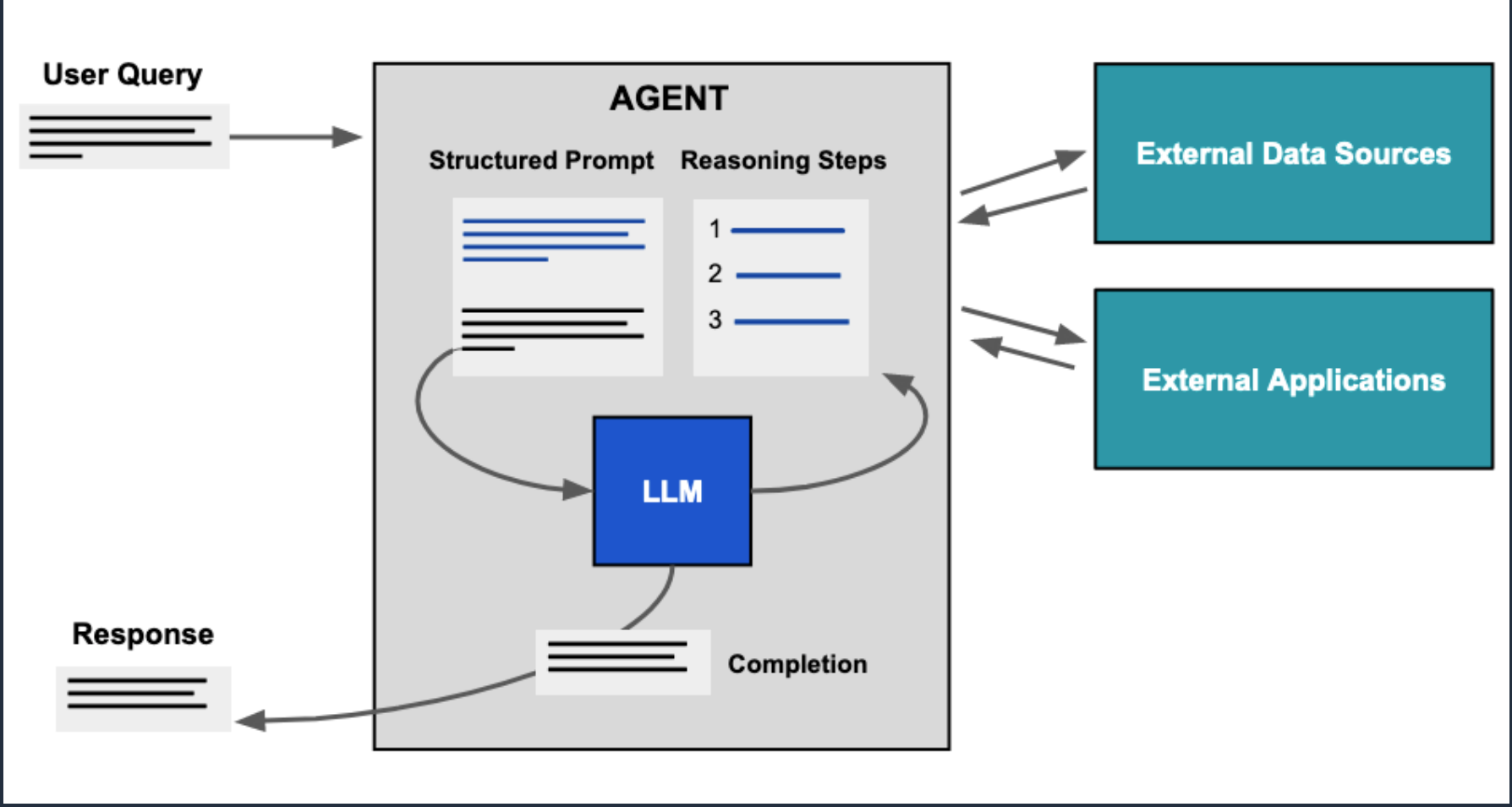
ReAct Framework

ReAct: Synergizing Reasoning and Acting in Language Models



Source: https://arxiv.org/pdf/2210.03629.pdf

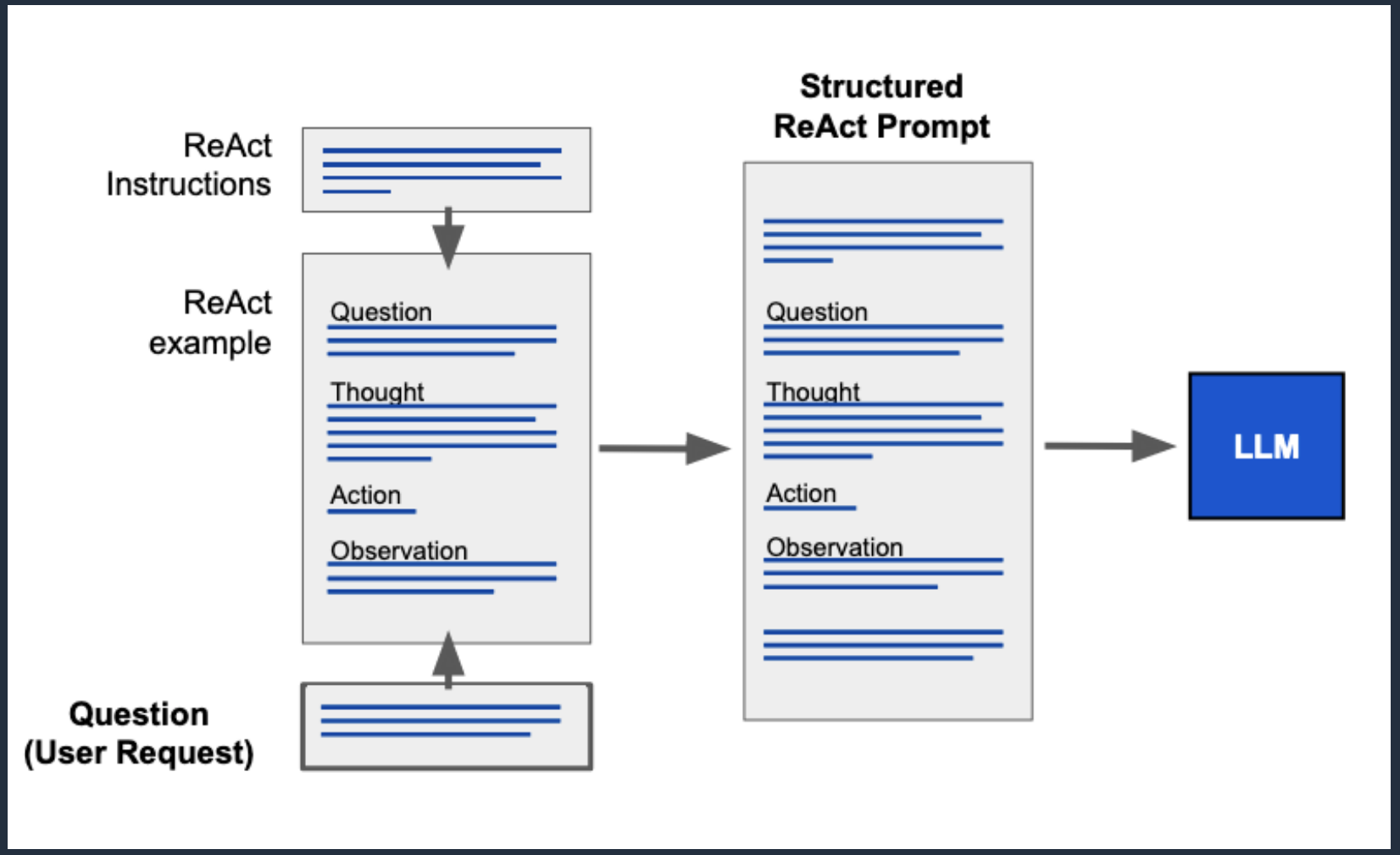
ReAct: The role of Agents



Agents orchestrate prompt-completion workflows between user requests, the foundation model, and external data sources and/or applications

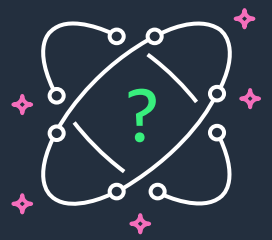
Source: Generative AI on AWS, O'Reilly

ReAct: Prompt Structure



Source: Generative AI on AWS, O'Reilly

ReAct: Prompt Breakdown

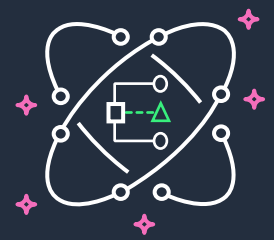


Question

Problem or query that requires reasoning and one or more steps to solve

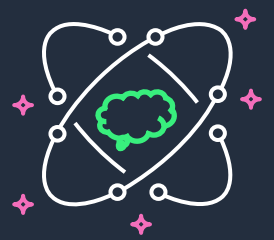
Thought

Reasoning steps to guide the model in how to solve the problem and which actions to take



Action

External actions that can be carried out through allowed actions and tools



Observation

Result of completing the action

Final Answer

ReAct: Prompt Breakdown

Question

Which candy was created first, Twix or Snickers?

```
Tools Available: [Wikipedia]
Actions Allowed:  search[entity], lookup[string], finish[answer]
```

Thought Need to search for Twix and Snickers and see which one was created first

Action search[Twix]

Observation *“The product was first produced in the United Kingdom in 1967...”*

Thought Twix was first produced in 1967. Search for Snickers next.

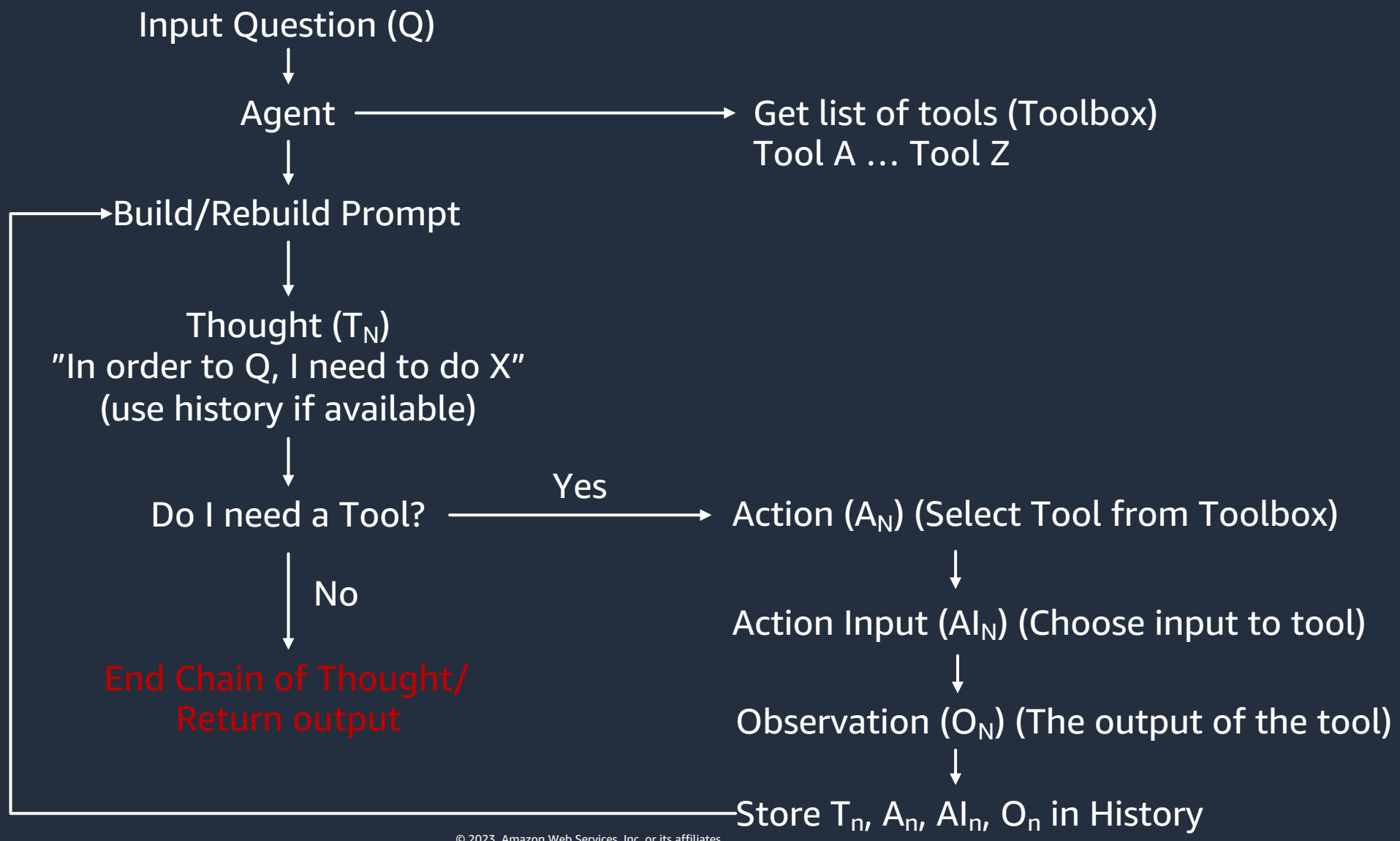
Action search[Snickers]

Observation “In 1930, Mars introduced Snickers,...”

Thought Twix was first produced in 1967. Snickers was first produced introduced in 1930 so Snickers was created first.

Action finish[Snickers]

ReAct: Visual Workflow



Workshop Tooling

Amazon SageMaker

PREPARE

Geospatial: Visual geospatial data

Ground Truth: Create high quality datasets for ML

Data Wrangler: Aggregate and prepare data for ML

Processing: Built-in Python, BYO R/Spark

Feature Store: Store, catalog, search, and reuse features

Clarify: Detect bias and understand model predictions

Data Scientist ML Engineer

Business Analyst

BUILD

Studio Notebooks & Notebook Instances: Fully managed Jupyter Notebooks with elastic compute

Studio Lab: Free ML development environment

Built-in Algorithms: Integrated tabular, NLP, and vision algorithms

JumpStart: UI based discovery, training, and deployment of models, solutions, and examples

Autopilot: Automatically create ML models with full visibility

Bring Your Own: Bring your own container and algorithms

Local Mode: Test and prototype on your local machine

TRAIN & TUNE

Fully Managed Training: Broad hardware options, easy to setup and scale

Distributed Training Libraries: High performance training for large datasets and models

Training Compiler: Faster deep learning model training

Automatic Model Tuning: Hyperparameter optimization

Managed Spot Training: Reduce training cost by up to 90%

Debugger and Profiler: Debug and profile training runs

Experiments: Track, visualize, and share model artifacts across teams

Customization Support: Integrate with popular open-source frameworks and libraries

Studio | RStudio
Integrated development environment (IDE) for ML

MLOps: Pipelines | Projects | Model Registry
Workflow automation, CI/CD for ML, central model catalog

Canvas
Generate accurate machine learning predictions—no code required

Governance
Model cards, Dashboard, Permissions

DEPLOY & MANAGE

Fully Managed Deployment: Ultra low latency, high throughput inference

Real-Time Inference: For steady traffic patterns

Serverless Inference: For intermittent traffic patterns

Asynchronous Inference: For large payloads or long processing times

Batch Transform: For offline inference on batches of large datasets

Multi-Model Endpoints: Reduce cost by hosting multiple models per instance

Multi-Container Endpoints: Reduce cost by hosting multiple containers per instance

Shadow Testing: Validate model performance in production

Inference Recommender: Automatically select compute instance and configuration

Model Monitor: Maintain accuracy of deployed models

Kubernetes & Kubeflow Integration: Simplify Kubernetes-based ML

Edge Manager: Manage and monitor models on edge devices

Amazon SageMaker Studio

brings tools for every step of the ML lifecycle under one unified visual user interface

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

- Choice of leading FMs through a single API
- Model customization
- Retrieval Augmented Generation (RAG)
- Agents that execute multistep tasks
- Security, privacy, and safety

Amazon Bedrock

Broad choice of models

AI21 labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI_	stability.ai
Contextual answers, summarization, paraphrasing	Text summarization, generation, Q&A, search, image generation	Summarization, complex reasoning, writing, coding	Text generation, search, classification	Q&A and reading comprehension	Text summarization, Q&A, text classification, text completion, code generation	High-quality images and art
Jurassic-2 Ultra	Amazon Titan Text Lite	Claude 3 Opus	Command	Llama 3 8B	Mistral Large	Stable Diffusion XL1.0
Jurassic-2 Mid	Amazon Titan Text Express	Claude 3 Sonnet	Command Light	Llama 3 70B	Mistral 7B	Stable Diffusion XL 0.8
	Amazon Titan Text Embeddings	Claude 3 Haiku	Embed English	Llama 2 13B	Mixtral 8x7B	
	Amazon Titan Text Embeddings V2	Claude 2.1	Embed Multilingual	Llama 2 70B		
	Amazon Titan Multimodal Embeddings	Claude 2	Command R+ (Coming Soon)			
	Amazon Titan Image Generator	Claude Instant	Command R (Coming Soon)			

aws

Thank you!

Giuseppe Zappia
Principal Solutions Architect
Amazon Web Services

Shelbee Eigenbrode
Principal AI/ML Specialist Solutions Architect
Amazon Web Services