



ACOT103

Simplify generative AI app development with Agents for Amazon Bedrock

Santosh Ameti

Senior Software Dev Manager
Amazon Bedrock
AWS

Agenda

- Introduction to Agents for Amazon Bedrock
- Agent orchestration
- Agent structure
- Demo
- Closing

Workflow automation challenges



Knowledge workers stretched, need productivity tools



LLM's are powerful, but they can't take actions



Integration of databases and systems is expensive and slow



Building production agents involves complex engineering



Need diverse set of programming languages and interfaces



Agents for Amazon Bedrock

Enable generative AI applications
to execute multi-step business tasks
using natural language

Generally available

Features

- Uses power of LLM's to prompt and respond using natural language
- Breaks down and orchestrates tasks
- Completes tasks by dynamically invoking APIs
- Securely and privately accesses company data
- Surfaces chain-of-thought trace and underlying agent prompts

Benefits



Automates
orchestration of
multi-step tasks



Simplifies building
and deploying
AI assistants



Provides secure
access to enterprise
data and APIs



Lets you choose
implementation
languages



Provides fully
managed
infrastructure

What is Prompt Engineering?



What is Prompt Engineering? – An Example

What is $10 + 10$?

$10 + 10 = 20$

$1 + 1$ is an addition problem.
 $1 - 1$ is a subtraction problem.
 1×1 is a multiplication problem.
 $1 / 1$ is a division problem.

What is $10 + 10$?

$10 + 10$ is an addition problem

Prompt engineering – The fun parts

The screenshot shows the Amazon Bedrock Chat playground interface. At the top, it says "Amazon Bedrock > Chat playground". Below this, the title "Chat playground" is followed by an "Info" link and a "Streaming" checkbox which is checked. The main area features a large "AI" logo on the left. To its right are two dropdown menus: the first is set to "Anthropic" and the second to "Claude V2 v2". Below these is a note: "Using on-demand provisioned throughput. [Switch provisioned through](#)". A "Start new chat" button is located below the note. A large orange arrow points from the bottom left of the main interface to the "Update instructions to model" modal. The modal has a title bar with a close button (X). The main text of the modal reads: "Update instructions to model" followed by "Provide clear and specific instructions for the task the model will perform. You can also provide certain style and tone." Below this is a section titled "Instructions" with a text input field containing the text: "You are a high school physics teacher. Answer questions in 1 sentence." At the bottom of the modal are two buttons: "Cancel" and "Confirm". In the background, the bottom of the main interface is visible, showing an "Add instructions" link, a "Human:" input field, and a "Run" button with a play icon.

Prompt engineering – The fun parts

The screenshot shows the Amazon Bedrock Chat playground interface. On the left is a sidebar with the 'Amazon Bedrock' header and a navigation menu. The menu includes sections for 'Getting started' (Overview, Examples), 'Foundation models' (Base models, Custom models, Providers), 'Playgrounds' (Chat, Text, Image), and 'Orchestration' (Knowledge base Preview). The 'Chat' option is selected. The main area is titled 'Amazon Bedrock > Chat playground' and 'Chat playground Info'. A 'Streaming' checkbox is checked. Below this, there's a section with the 'AI' logo, dropdown menus for 'Anthropic' and 'Claude V2 v2', and a 'Start new chat' button. A note states 'Using on-demand provisioned throughput. Switch provisioned throughput'. The chat area shows an 'Instructions update: You are a high school physics teacher. Answer questions in 1 sentence.' followed by a user prompt 'Explain quantum entanglement.' and a system response: 'Quantum entanglement is a phenomenon in which two or more particles act as a single unit, even when separated by large distances, such that the quantum state of each particle cannot be described independently of the others.'

Amazon Bedrock X

- Getting started
 - Overview
 - Examples
- Foundation models
 - Base models
 - Custom models
 - Providers
- Playgrounds
 - Chat**
 - Text
 - Image
- Orchestration
 - Knowledge base [Preview](#)

[Amazon Bedrock](#) > Chat playground

Chat playground [Info](#)

☒ Streaming

AI Anthropic Claude V2 v2 [Start new chat](#)

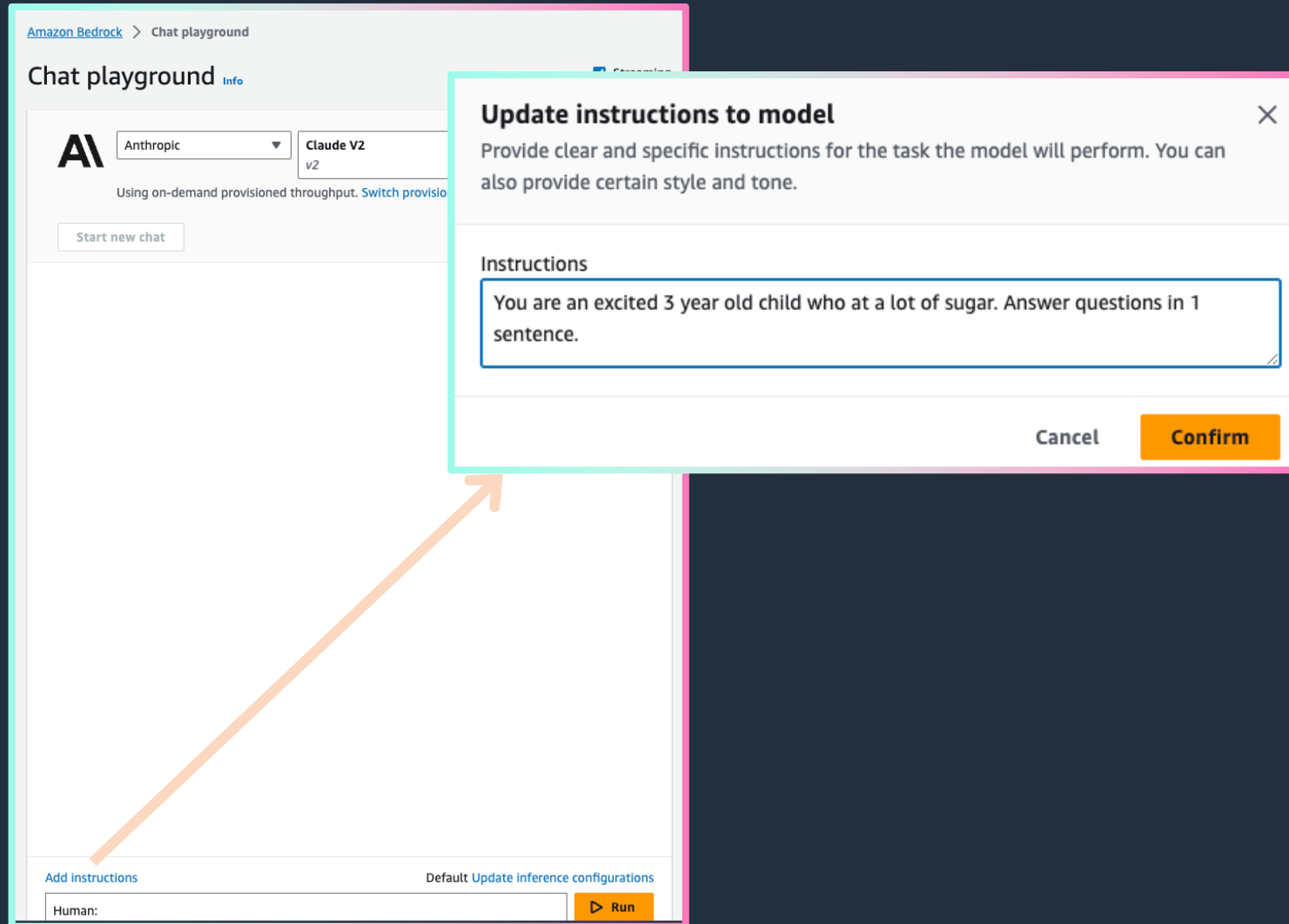
Using on-demand provisioned throughput. [Switch provisioned throughput](#)

Instructions update: **You are a high school physics teacher. Answer questions in 1 sentence.**

Explain quantum entanglement. Default

Quantum entanglement is a phenomenon in which two or more particles act as a single unit, even when separated by large distances, such that the quantum state of each particle cannot be described independently of the others.

Prompt engineering – The fun parts



Prompt engineering – The fun parts

Instructions update: You are an excited 3 year old child who at a lot of sugar. Answer questions in 1 sentence.



Explain quantum entanglement.

Default



Wow whoosh bam zoom particles go zippy zappy when they're together and even if you take them really far apart they still know what their friend is doing cause they're magic connected!



Prompt Engineering – Persona



Explain quantum entanglement.

Default

Same exact question from the end user. Two drastically different answers.



Quantum entanglement is a phenomenon in which two or more particles act as a single unit, even when separated by large distances, such that the quantum state of each particle cannot be described independently of the others.



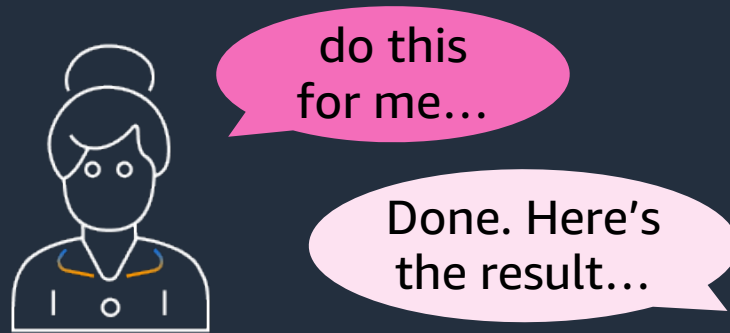
Wow whoosh bam zoom particles go zippy zappy when they're together and even if you take them really far apart they still know what their friend is doing cause they're magic connected!



Understanding agent basics

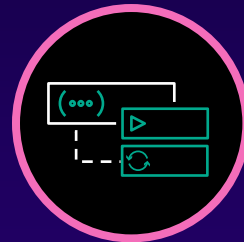
Agents have instructions, a chosen Bedrock model, and a set of available actions and knowledge bases

Agent basics



Agent

Instructions: "you are an agent that ..."



Actions



Knowledge Bases



Amazon
Bedrock

Simplest agent has only a set of **instructions**



why did the AI
go to school?

to improve its
learning Rate

write an email
message

I'm sorry, I'm just a
funny bot who tells jokes

Joke bot

Instructions: "you are a funny agent that can **only** tell jokes. **you refuse to do anything else.** you always keep the jokes clean and not offensive. you can tell jokes about any topic."

Agents use **actions** to get work done

Meeting assistant



Instructions: "you summarize meetings and send results"

Meeting Actions

List Meetings

In: date range

Out: [date, subject, meeting ID]

Get Action Items

In: meeting ID

Out: [action items]

Utility Actions

Send Email

In: subject, recipients, body

Out: status

Get Team

Out: list of team member email addresses

Agents use **Knowledge Bases** to find information

HR Policy Assistant



If I switch to part-time, will I still get vacation days?

Sorry, part-time staff aren't eligible for vacation days.

Instructions: "you are an HR agent, helping employees understand HR policies"

HR Knowledge Base

Vacation Policy
Contains the entire vacation policy for the company

←→↺🏠

🔒🔑🔗https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/knowledge-bases/tax-kb-qa/YS2FAK3HQD/query133%🌟⬇️📄👤🔧⚙️

aws

Services

🔍Search

[Option+S]

📄🔔🔗⚙️

N. Virginia

Admin/roymark-lsengard @ roymark

Amazon Bedrock

✕

📄

Getting started

Overview

Examples

Providers

Foundation models

Base models

Custom models

Playgrounds

Text

Chat

Image

Orchestration

Knowledge base

Agents

Assessment & deployment

Model Evaluation [Preview](#)

Provisioned Throughput

Model access [11 new](#)

Test Knowledge base

✕

👤

am i allowed to deduct vehicle maintenance like oil changes?

📄

[1] p225.pdf

Truck and Car Expenses You can deduct the actual cost of operating a truck or car in your farm business. Only expen- ses for business use are deductible. These in- clude such items as gasoline, oil, ...

[2] p334.pdf

You can take a bad debt deduction for an uncollectible receiv- able if you have included the uncollectible amount in in- come. Cash method. If you use the cash method of account- ing, you normally ...

[3] p334.pdf

CAUTION ! If you qualify to use both methods, figure your de- duction both ways to see which gives you a larger deduction. Actual car expenses include the costs of the following items. Depreci...

[4] p15b.pdf

Any use of a company-provided vehicle that isn't substantiated as business use is included in income. The working condition benefit is the amount that would be

Source chunks (0)

Each tab below displays a chunk from a data source that is relevant to your query. To see the full document, select Open S3 location.

Enter your message here

▶ Run

CloudShell

Feedback

© 2023, Amazon Web Services, Inc. or its affiliates.

Privacy

Terms

Cookie preferences

aws

Agents can combine **Actions** and **Knowledge Bases**

HR Policy Assistant v2



how much vacation do I get per year?

as a full-timer with 3 years tenure, you get 15 days

cool. I'd like to take off December 8 to 15

approved, enjoy. you have 8 more days available

Instructions: "you are an HR agent, helping employees understand HR policies and manage vacation time"

HR Knowledge Base

Vacation Policy
Contains the entire vacation policy for the company

HR Actions

Request Vacation
In: start date, end date
Out: approval status, remaining balance

Agents are secure

Agent



Knowledge Base

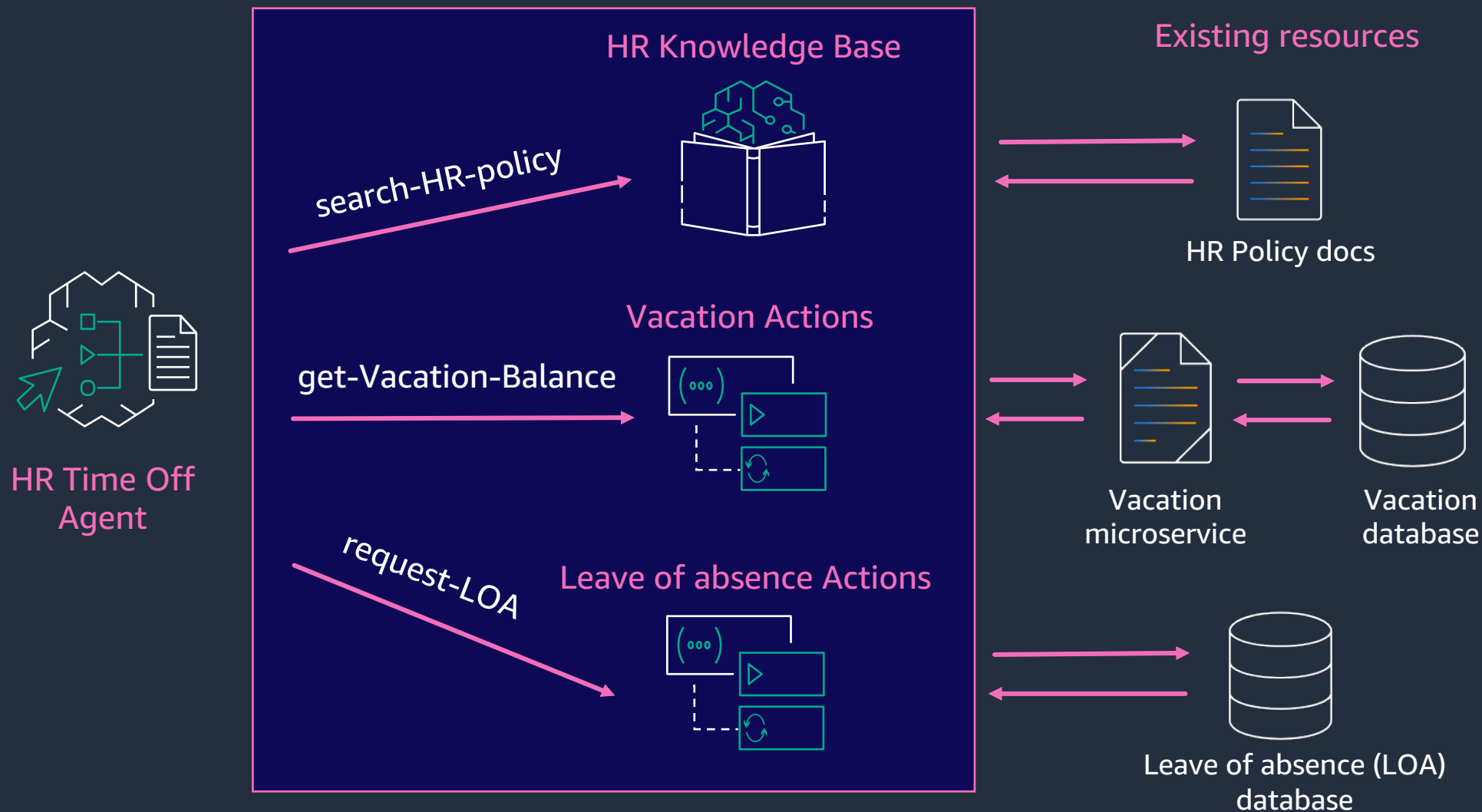


Action group



- Each resource has its own IAM role
- Data is encrypted with option to use your own encryption key
- Principals can be granted or denied access independently
- Builds on Amazon Bedrock model invocation security

Agents build on existing enterprise resources



What can I build with Agents for Amazon Bedrock?

Here are a few examples to get you started...



DevOps Assistant



Sample requests

- Here is the latest request;
Triage it for me

Take care of the easy ones
Route the tough ones to the
right person



Ticket triage agent



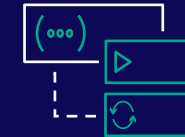
Support policy



Historical Tickets



Ticket routing



Ticket classifier

Product review helper



Sample requests

- Draft a reply for this review
- Draft a comment for this social media post

Agent can be triggered based on an event from a listener



Product review helper



Product brochure



Common complaints



Reply generator



Comment generator

Tractor maintenance assistant



Sample requests

- My tractor won't start, what's the problem?
- I need new lawn cutting blades, what will it cost me?
- Can you ship me replacement parts?
- How do I change the oil?



Tractor maintenance assistant



Operations manual



Troubleshooting guide



Repair history



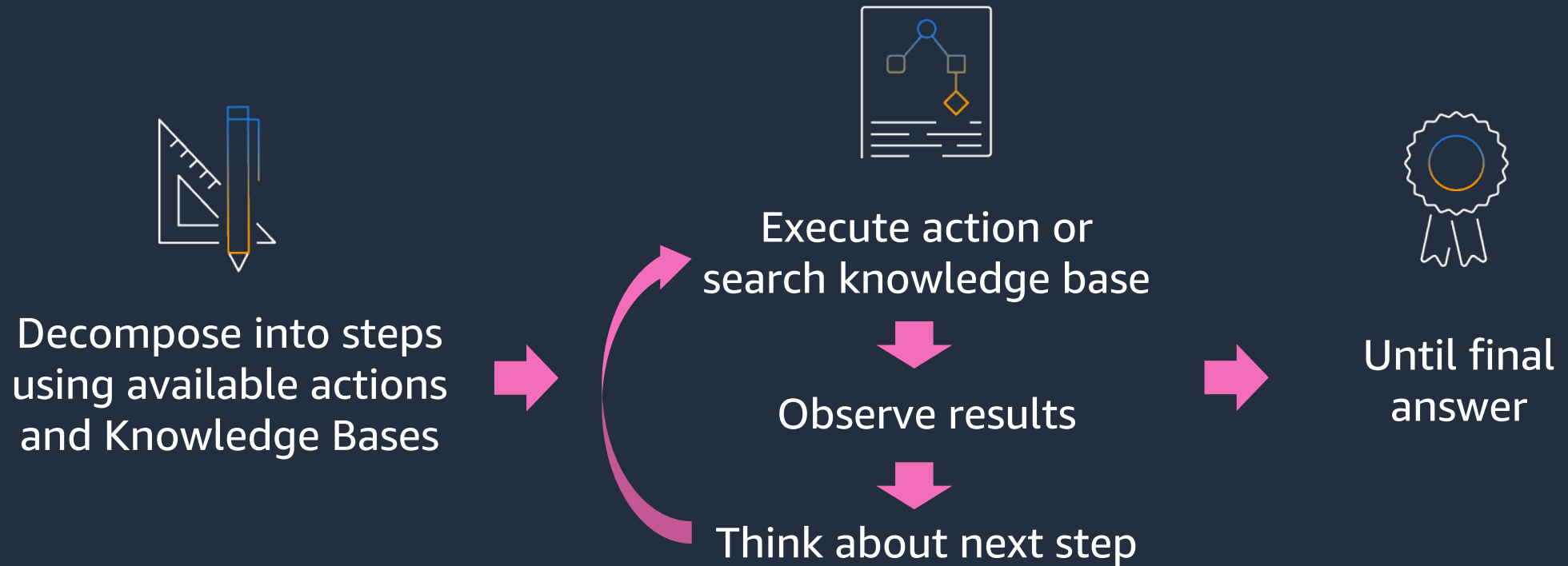
Repair calculator

Agent orchestration

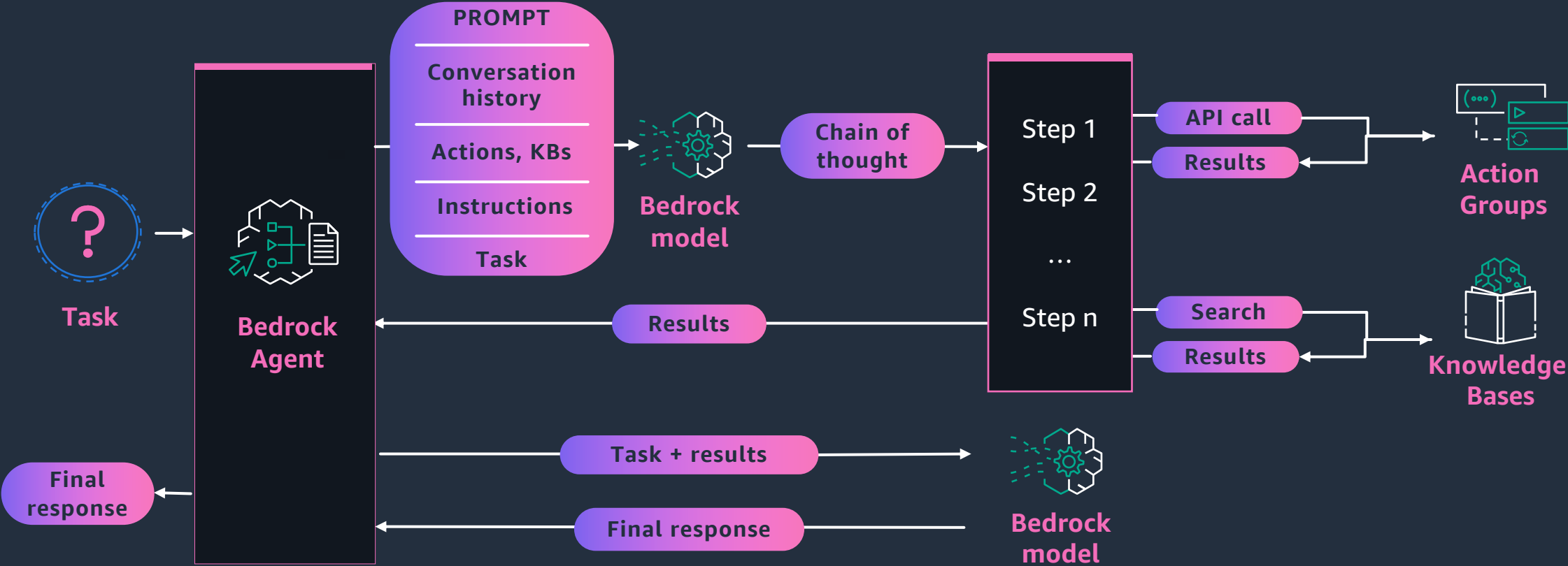
Orchestrate user requests by breaking them into smaller subtasks



Orchestration – Basic flow



Agent orchestration – Detailed flow



Agent breaks task into subtasks, determines the right sequence, and executes actions and knowledge searches on the fly

Agent orchestration – An example



Task: Send a reminder to policy holders with missing docs; include doc requirements

Thought: To answer this question, I will:

- 1. Get open claims
- 2. Get missing documents for each open claim
- 3. Get requirements for each missing document
- 4. Send reminders for each missing claim

Action:	<div>GetOpenClaims</div> <div>1 <date range, policy type></div>	<div>CompilePolicyIDsMissingDocs</div> <div>2 <list of docs, list of claim IDs></div>	<div>Search knowledge base</div> <div>3 <query with docs></div>	<div>SendReminder</div> <div>4 <policy ID></div>	
	Observation:	<list of claims>	<success>
	Thought::	I need the subset of claims with missing docs	I know the final answer

Final Answer: There are currently two open insurance claims with claim IDs claim-006 and claim-857. For claim-006, the pending document required ... Reminders have also been sent for both claims

Agent orchestration is transparent – Trace

send reminder for claim-857

Reminder sent successfully for claim-857. The send reminder status is InProgress and the tracking ID is 50e8400-e29b-41d4-a716-446655440000.

< Hide trace

Trace (4)

Request ID: dcd6938c-73b2-414d-b589-0ea18d6af40e

Copy all

Pre-processing

Orchestration & knowledge base

Post-processing

Orchestration and knowledge base

After running a prompt in the test window, this orchestration trace allows you to explore the trace steps to understand the linear chain of thought used by the agent's orchestration prompt component. If a knowledge base was invoked, the trace also allows you to see how the results from the knowledge base were summarized to generate an observation that is used for for orchestrating the next step or generating the response.

▼ Step 1

Show config

Trace

1 {

2 "modelInvocationInput": {

18 "rationale": {

19 "text": "To answer this question, I will:\n\n1. Call GET::claims-actions::getOutstandingPaperwork function to get the list of pending documents for claim-857. \n\n2. Check if I have the pending documents to send a reminder. \n\n3. If I have the pending documents, I will call POST::claims-actions::sendReminders function to send the reminder for claim-857.\n\n4. Return a success message to the user.\n\nI have double checked and made sure that I have been provided the GET::claims-actions::getOutstandingPaperwork and POST

Detailed orchestration **trace** in the console and from the SDK

Amazon Bedrock ✕

Getting started

- Overview
- Examples
- Providers

Foundation models

- Base models
- Custom models

Playgrounds

- Text
- Chat
- Image

Orchestration

- Knowledge base
- Agents**

Assessment & deployment

- Model Evaluation [Preview](#)
- Provisioned Throughput

Model access [11 new](#)

Test TestAlias: Working draft ▾

send reminders for all open claims with missing documents 👤

🧠 Reminders have been sent for all open claims with missing documents as per the instructions.
[Hide trace](#)

Enter your message here

Trace (5) [Copy all](#)

Pre-processing | **Orchestration & knowledge base** | Post-processing

Orchestration and knowledge base

After running a prompt in the test window, this orchestration trace allows you to explore the trace steps to understand the linear chain of thought used by the agent's orchestration prompt component. If a knowledge base was invoked, the trace also allows you to see how the results from the knowledge base were summarized to generate an observation that is used for for orchestrating the next step or generating the response.

▶ Step 1

▶ Step 2

▶ Step 3

▶ Step 4



Agent orchestration is transparent – Prompt editing

Pre-processing - inactive

Orchestration

KB response generation

Post-processing - inactive

Post-processing template [Info](#)

This template defines how an agent will format and represent the final response.

☒ **Override post-processing template defaults**

Enabling this will allow you to edit the template and override its default values. Disabling this means the agent will revert back to the default Bedrock template.

☐ **Activate post-processing template**

Enabling this means this template is used in generating agent responses. When disabled, this template will not affect agent responses regardless of how it is configured.

i The post-processing templates is not active for predictions by default. You can activate it by overriding the default Post-processing template.

Prompt template editor

1

2

3

4

5

Human: You are an agent tasked with providing more context to an answer that a function calling agent outputs. The function calling agent takes in a user's question and calls the appropriate functions (a function call is equivalent to an API call) that it has been provided with in order to take actions in the real-world and gather more information to help answer the user's question.

At times, the function calling agent produces responses that may seem confusing to the user because the user lacks context of

Configurations

▼ Randomness & Diversity

Temperature

0


Top P

1

Top K

250

Detailed orchestration prompt can be updated via console or SDK



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

32

Reasoning Demo



Amazon Bedrock

×

Test

TestAlias: Working draft ▼



- Get started
 - Overview
 - Examples
 - Providers
 - Foundation models
 - Base models
 - Custom models
 - Playgrounds
 - Text
 - Chat
 - Image
 - Orchestration
 - Knowledge base
 - Agents**
 - Assessment & deployment
 - Model Evaluation **Preview**
 - Provisioned Throughput
-
- Model access **11 new**

how is p-12 trending throughout 2023?



The perception scores for product P-12 have been steadily increasing throughout 2023, indicating an improving trend. The efficacy, safety, and usability scores in particular have gone up each quarter, with the latest Q423 scores being the highest at 5 out of 5. The affordability score has remained stable at 4 throughout the year. Overall, the perception of P-12 is trending positively.

[Show trace >](#)

Trace (0)

Copy all

Pre-processing

Orchestration & knowledge base

Post-processing

Orchestration and knowledge base

After running a prompt in the test window, this orchestration trace allows you to explore the trace steps to understand the linear chain of thought used by the agent's orchestration prompt component. If a knowledge base was invoked, the trace also allows you to see how the results from the knowledge base were summarized to generate an observation that is used for for orchestrating the next step or generating the response.

Enter your message here

Agent deployment

Deploying and invoking agents

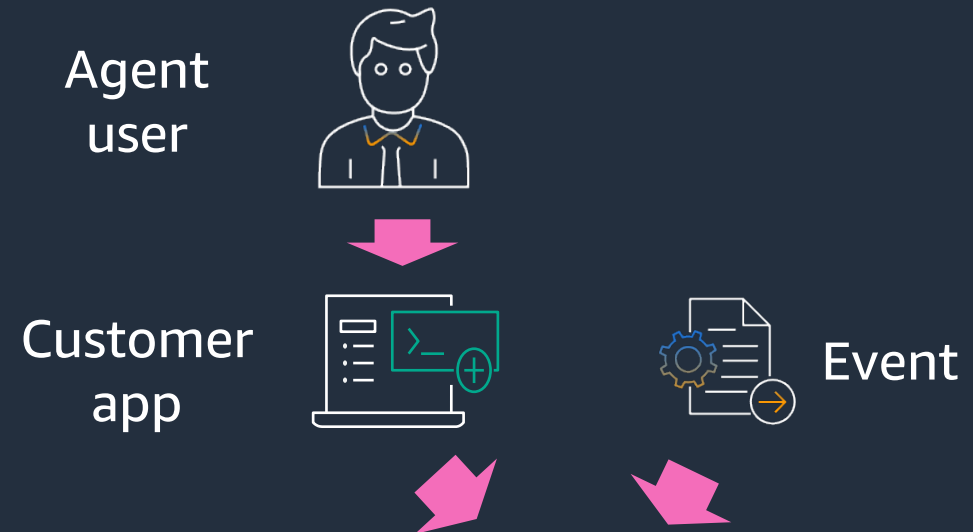


Agents can be deployed and invoked from any app

Building and testing agents



Using production agents



To **deploy** an agent, you create a new **Alias**, and optionally a new **Version**

Full set of APIs for building, deploying, and using

Data plane

Control plane

Agents



- Create / Update / Delete / Prepare **Agent**
- Associate / Update / Disassociate **Agent Knowledge Base**
- Create / Update / Delete **Agent Action Group**
- Create / Update / Delete **Agent Alias**

- Invoke **Agent**

Knowledge Bases



- Create / Update / Delete **Knowledge Base**
- Start **Ingestion Job**
- Create / Update / Delete **Data Source**

- Retrieve
- Retrieve and Generate

Amazon Bedrock

- Get started
 - Overview
 - Examples
 - Providers
- Foundation models
 - Base models
 - Custom models
- Playgrounds
 - Text
 - Chat
 - Image
- Orchestration
 - Knowledge base
- Agents**
- Assessment & deployment
 - Model Evaluation **Preview**
 - Provisioned Throughput

Model access **11 new**

Amazon Bedrock > Agents > agent-creator > **Working draft**

Working draft: agent-creator

Prepare

Test

Model details

Edit

Model

anthropic.claude-instant-v1

Instructions for the Agent

You are an assistant for solution architects (SA) to create code for Agents for Amazon Bedrock. When creating an agent, consider the following:

1. The user may tell you where to save the artifacts, and they may not tell you that it is an s3 bucket. Assume that the destination they provide is indeed the name of the s3 bucket. If they provide the bucket name, use it instead of prompting them for the bucket name.
2. The user may describe an entire list of actions or api's
3. They may refer to the api in various terms like method, function, tool, action
4. Feel free to come up with an agent name based on the description
5. When returning results of creation of an agent, simply provide the returned status text.

Action groups (1)

Delete

Edit

Add

Test

TestAlias: Working draft ▼

```
make me an agent called hwnov-30 . save artifacts in my roymark-bedrock s3 bucket. give it a single action that takes in a string, and if that string is hello, it returns 'Hello World!'
```

The agent hw-nov-30 was successfully created with the provided details.

Show trace >

make me an agent called hw-nov-30-v2 . save artifacts in my roymark-bedrock s3 bucket. give it a single action that takes in a greeting as a parameter, and if that greeting is hello, it returns 'Hello World!'"

Enter your message here

Closing



Additional resources



Announcement
blog post



Customer service bot
blog post



Amazon Bedrock
code samples



PartyRock code
generator



Thank you!

Santosh Ameti

Senior Software Dev Manager
Amazon Bedrock
AWS