ZURICH | September 4, 2024

aws SUMMIT

AIM303

Deep Dive: Building AI agents using Amazon Bedrock

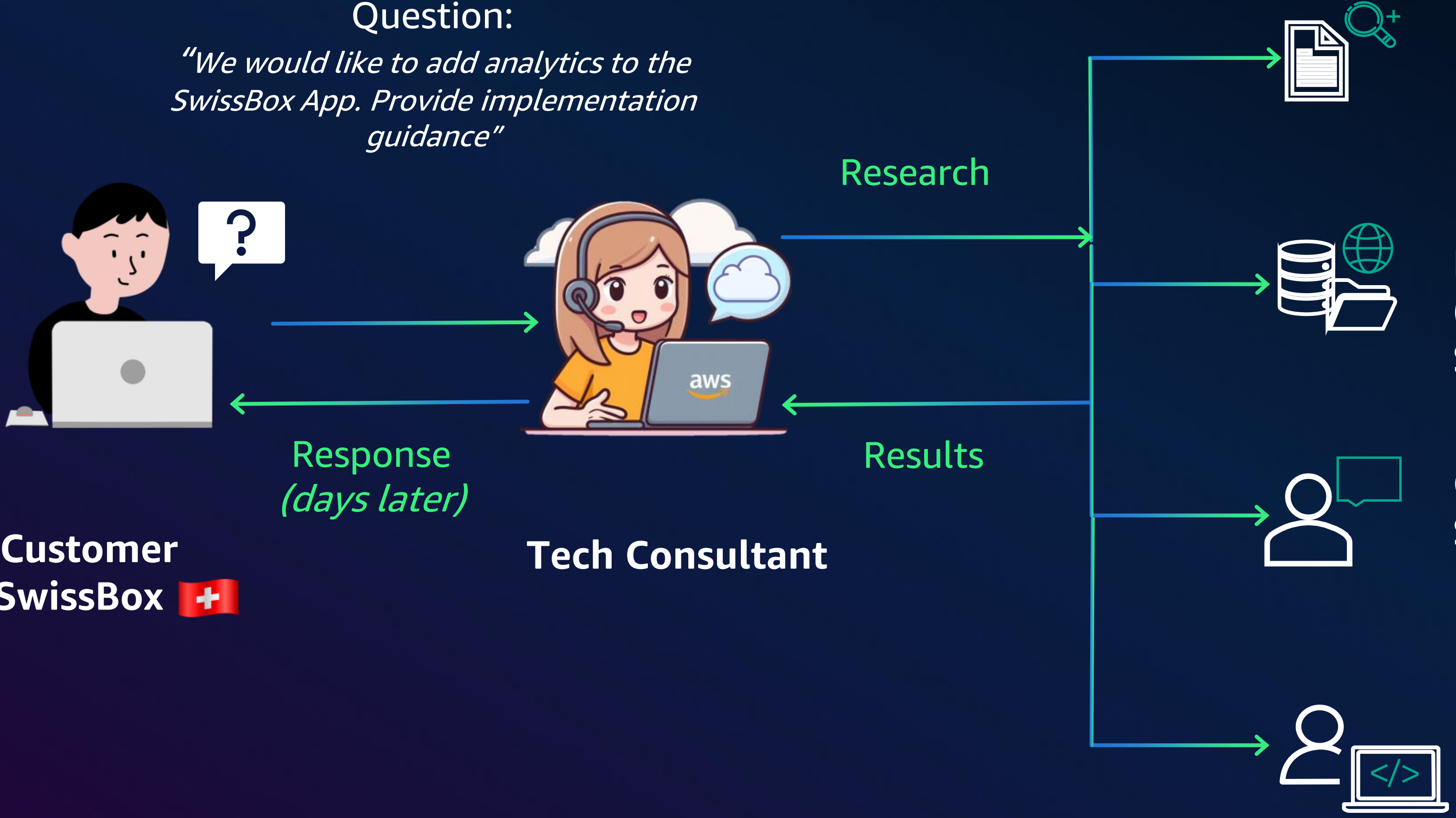
Viktor Vedmich
Senior Developer Advocate
AWS

Viktoria Semaan
Senior Developer Advocate
AWS

Human Workflow

Question:

"We would like to add analytics to the SwissBox App. Provide implementation guidance"



Research

Results

Response *(days later)*

Customer SwissBox 🇨🇭

Tech Consultant

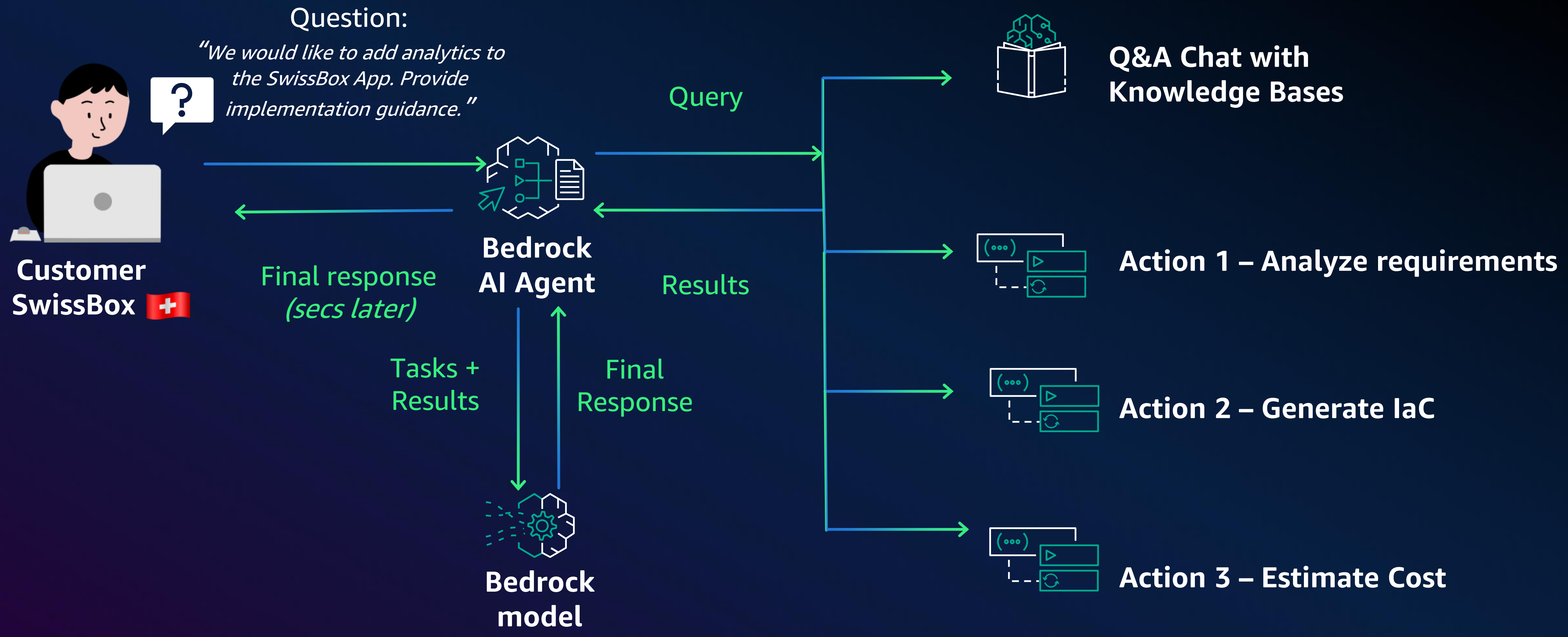
Analyze requirements

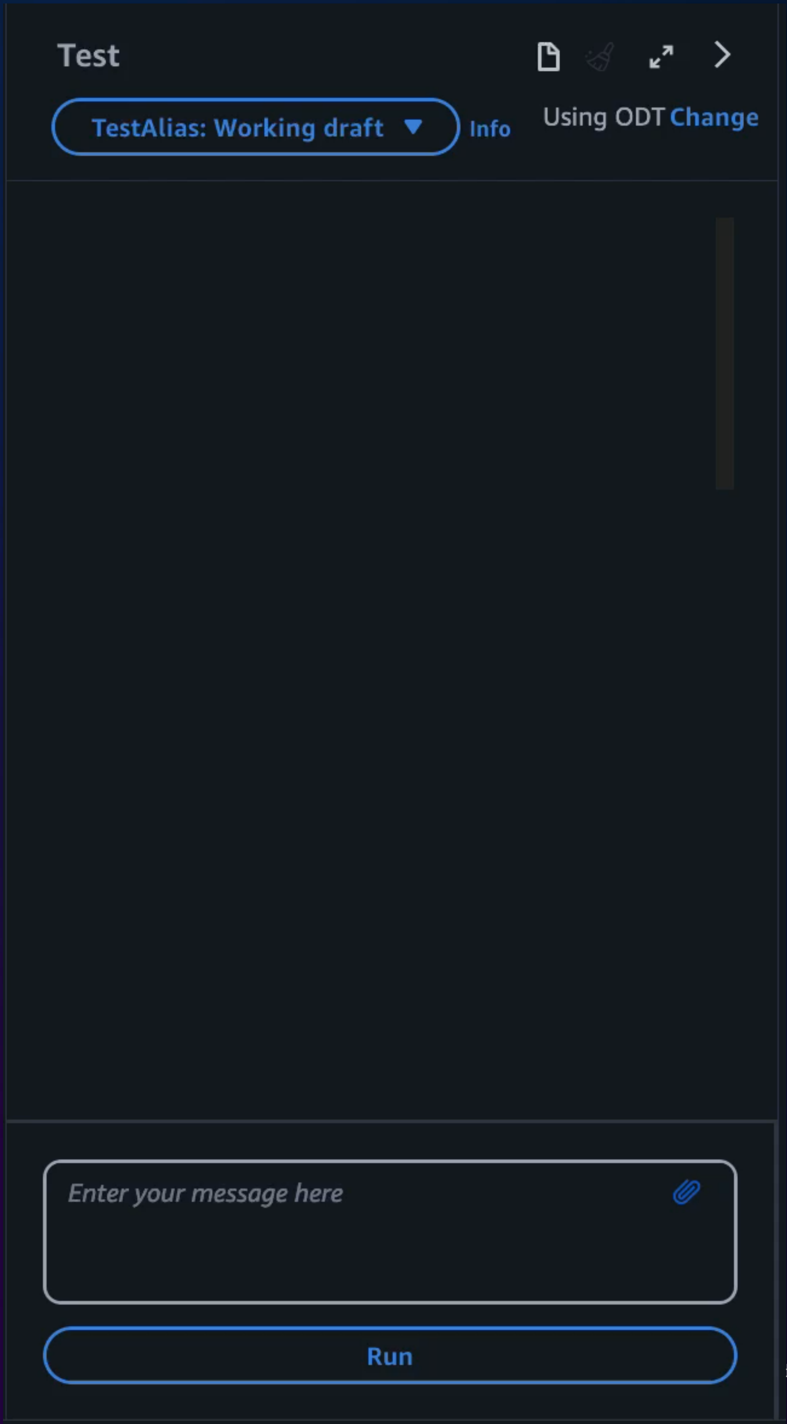
Knowledge Bases (Tech Docs, Customer Studies)

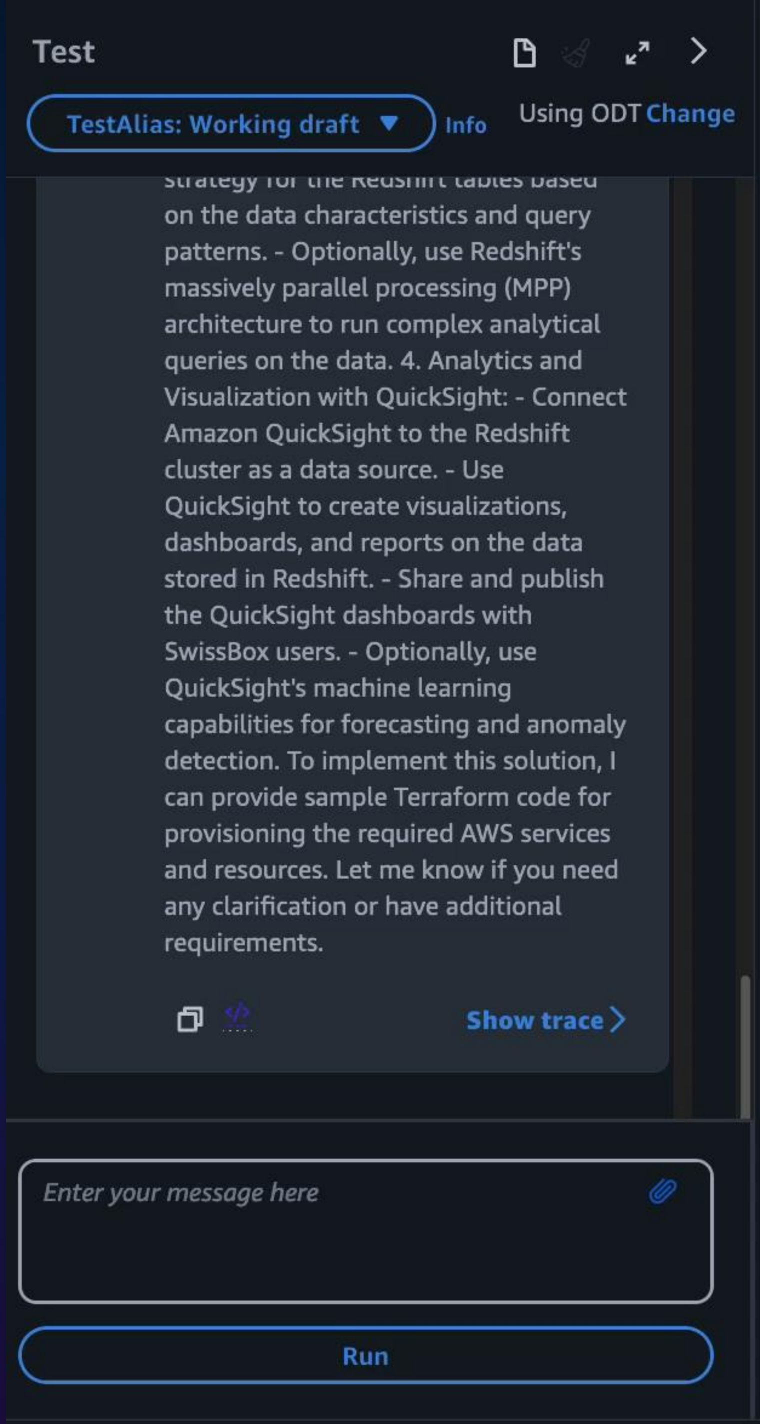
Connect with Specialist teams

Build a prototype

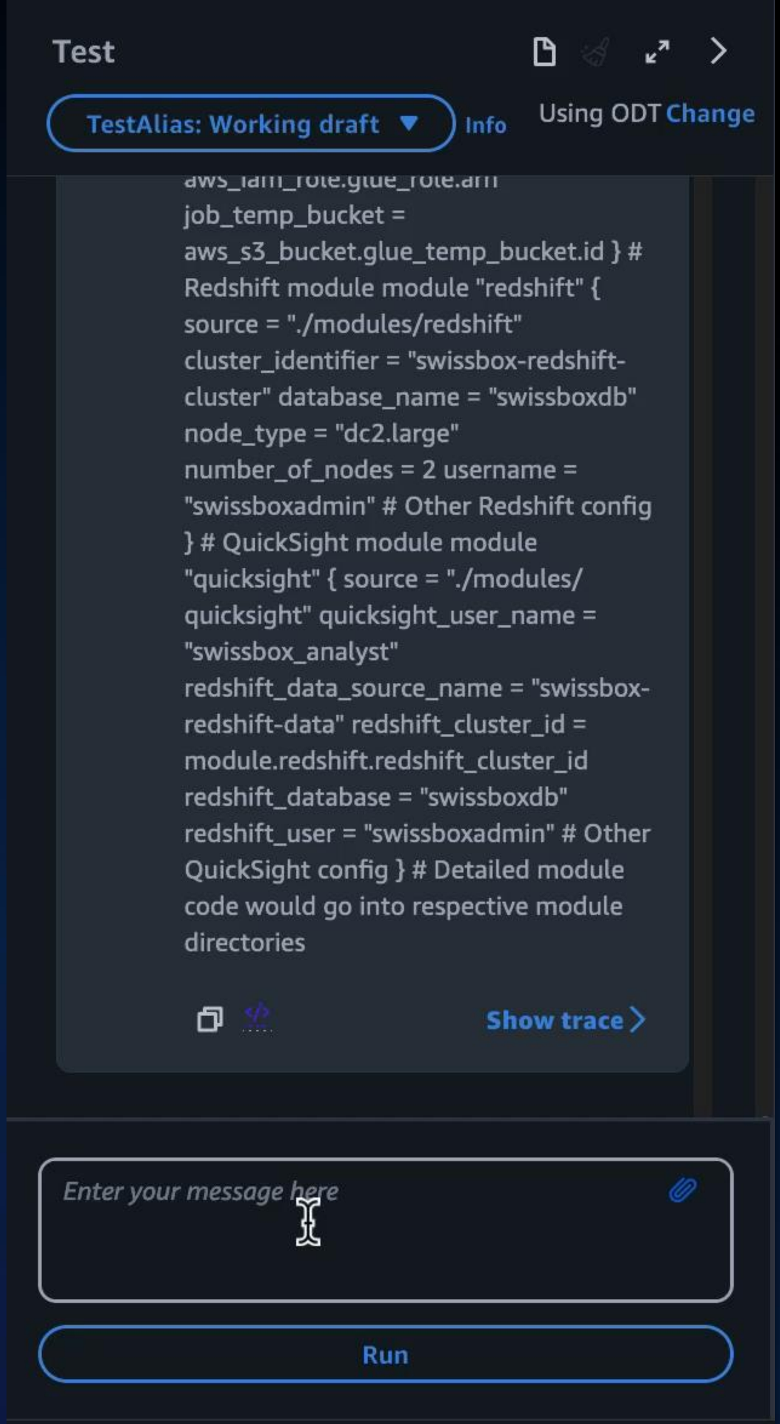
AI Agent Workflow


Question:
"We would like to add analytics to the SwissBox App. Provide implementation guidance."
Customer SwissBox 🇨🇭
Final response (secs later)
Bedrock AI Agent
Query
Results
Tasks + Results
Final Response
Bedrock model
Q&A Chat with Knowledge Bases
Action 1 – Analyze requirements
Action 2 – Generate IaC
Action 3 – Estimate Cost





strategy for the Redshift tables based on the data characteristics and query patterns. - Optionally, use Redshift's massively parallel processing (MPP) architecture to run complex analytical queries on the data. 4. Analytics and Visualization with QuickSight: - Connect Amazon QuickSight to the Redshift cluster as a data source. - Use QuickSight to create visualizations, dashboards, and reports on the data stored in Redshift. - Share and publish the QuickSight dashboards with SwissBox users. - Optionally, use QuickSight's machine learning capabilities for forecasting and anomaly detection. To implement this solution, I can provide sample Terraform code for provisioning the required AWS services and resources. Let me know if you need any clarification or have additional requirements.



aws_iam_role.glue_role.arn job_temp_bucket = aws_s3_bucket.glue_temp_bucket.id } # Redshift module module "redshift" { source = "./modules/redshift" cluster_identifier = "swissbox-redshift-cluster" database_name = "swissboxdb" node_type = "dc2.large" number_of_nodes = 2 username = "swissboxadmin" # Other Redshift config } # QuickSight module module "quicksight" { source = "./modules/quicksight" quicksight_user_name = "swissbox_analyst" redshift_data_source_name = "swissbox-redshift-data" redshift_cluster_id = module.redshift.redshift_cluster_id redshift_database = "swissboxdb" redshift_user = "swissboxadmin" # Other QuickSight config } # Detailed module code would go into respective module directories

AMAZON BEDROCK

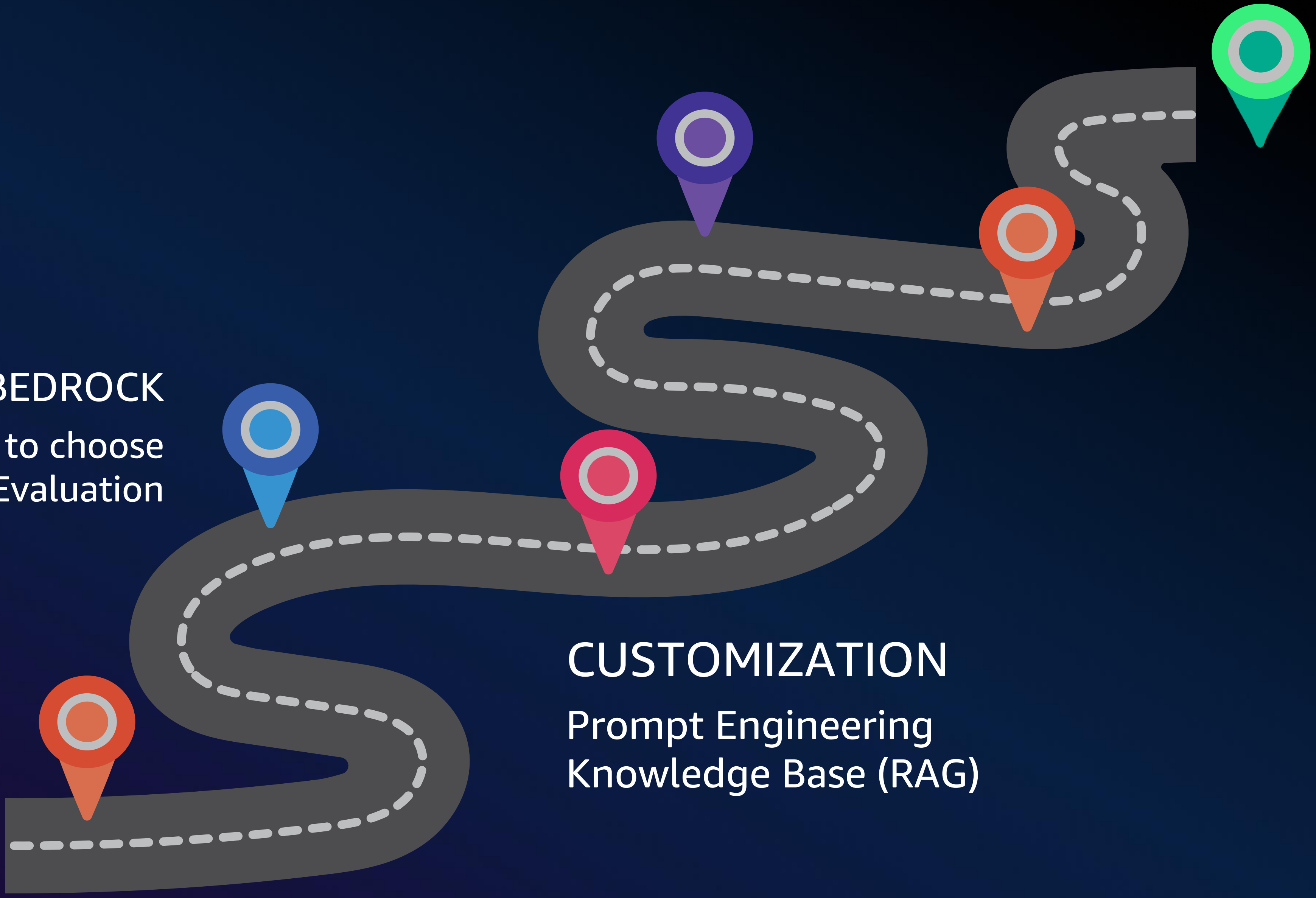
Models to choose
Model Evaluation

INTRO
Use Case

Amazon Bedrock Foundation Models

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, text classification, text completion, code generation, Q&A**	**High-quality images and art**
Jamba-Instruct Jurassic-2 Ultra Jurassic-2 Mid	Amazon Titan Text Premier Amazon Titan Text Lite Amazon Titan Text Express Amazon Titan Text Embeddings Amazon Titan Text Embeddings V2 Amazon Titan Multimodal Embeddings Amazon Titan Image Generator	Claude 3.5 Sonnet Claude 3 Opus Claude 3 Sonnet Claude 3 Haiku Claude 2.1 Claude 2 Claude Instant	Command Command Light Embed English Embed Multilingual Command R+ Command R	Llama 3 8B Llama 3 70B Llama 3.1 8B Llama 3.1 70B Llama 3.1 405B	Mistral Small Mistral Large Mistral 7B Mixtral 8x7B	Stable Diffusion XL1.0 Stable Diffusion XL 0.8



INTRO

Use Case

AMAZON BEDROCK

Models to choose

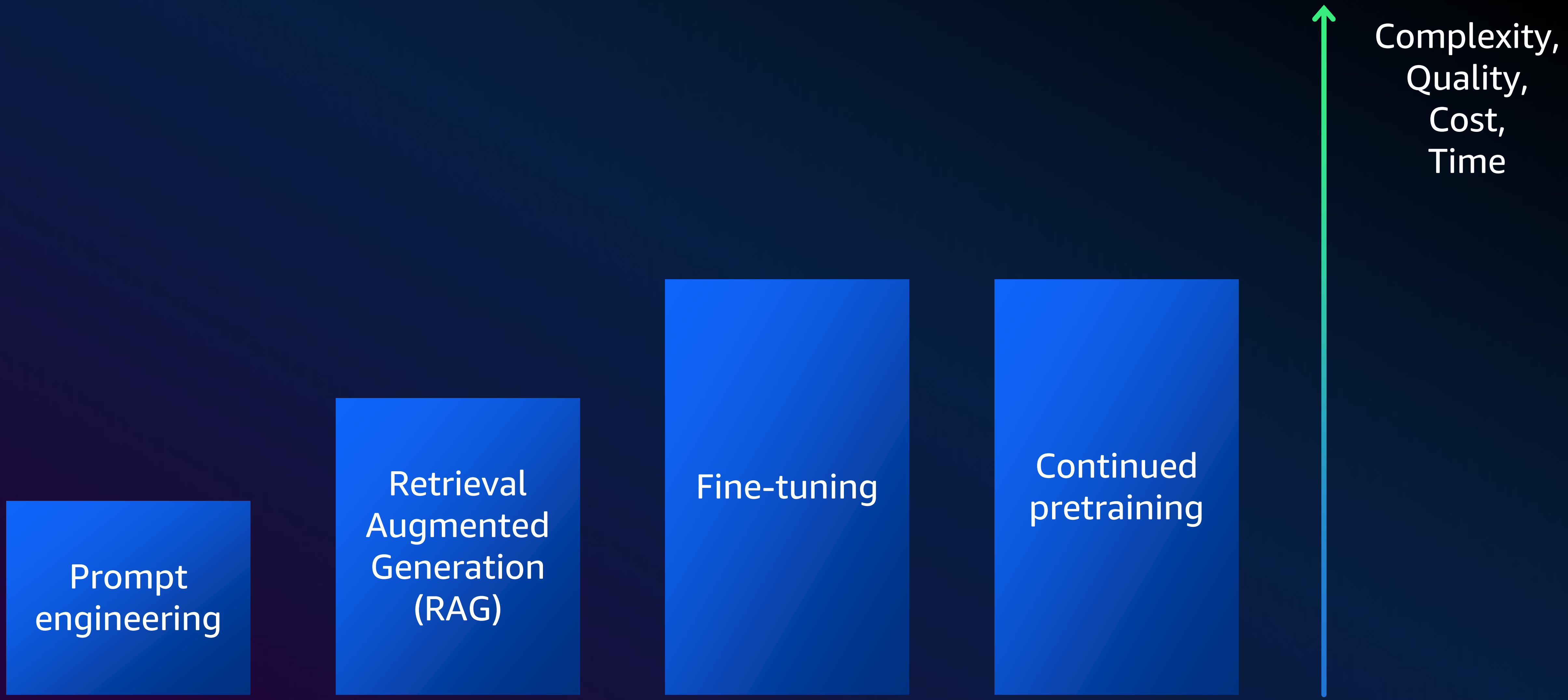
Model Evaluation

CUSTOMIZATION

Prompt Engineering

Knowledge Base (RAG)

4 approaches to customize FMs



Prompt engineering

Retrieval Augmented Generation (RAG)

Fine-tuning

Continued pretraining

Complexity,
Quality,
Cost,
Time

Customizing model responses for your business

Fine-tuning

PURPOSE

Maximizing accuracy for **specific tasks**

DATA NEED

Small number of labeled examples

Continued pretraining

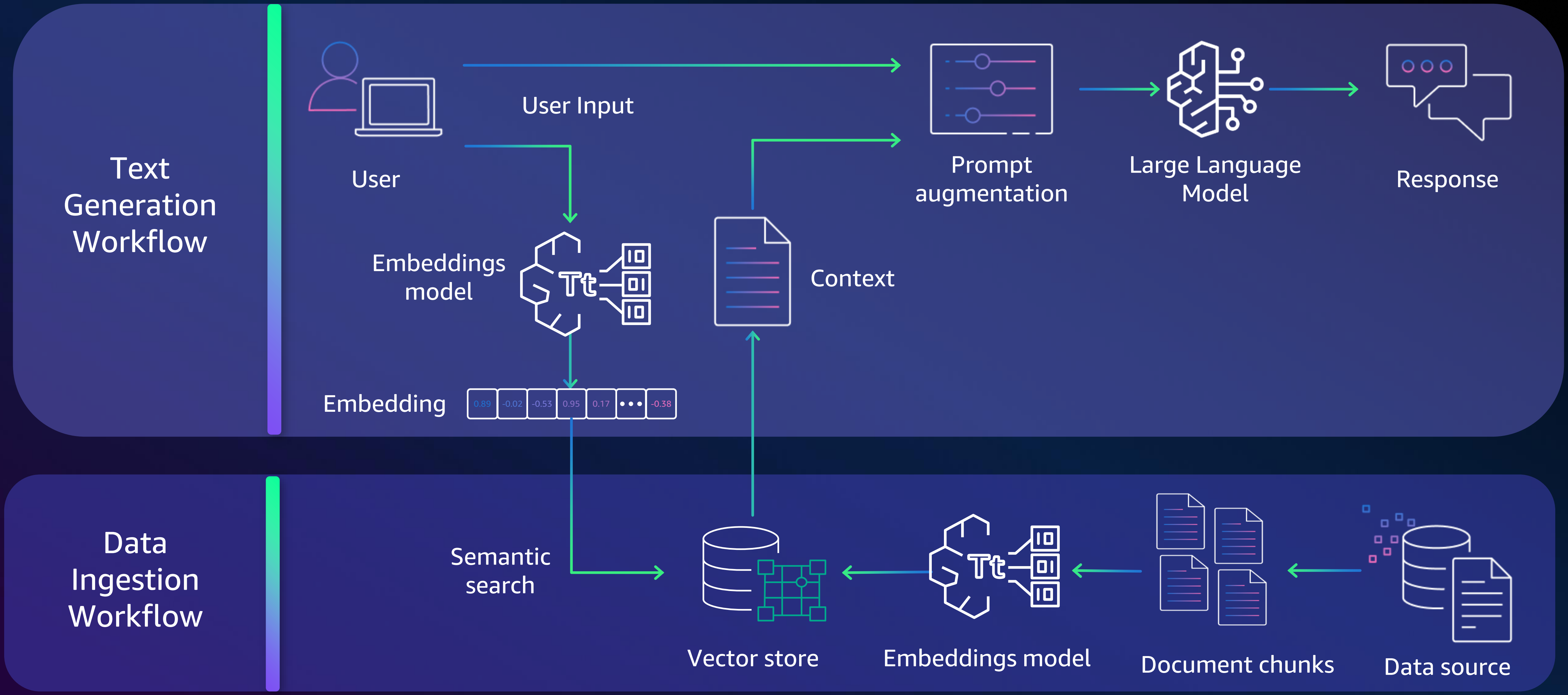
PURPOSE

Maintaining model accuracy for **your domain**

DATA NEED

Large number of unlabeled datasets

Knowledge Base: End-to-End RAG



INTRO

Use Case

AMAZON BEDROCK

Models to choose

Model Evaluation

CUSTOMIZATION

Prompt Engineering

Knowledge Base (RAG)

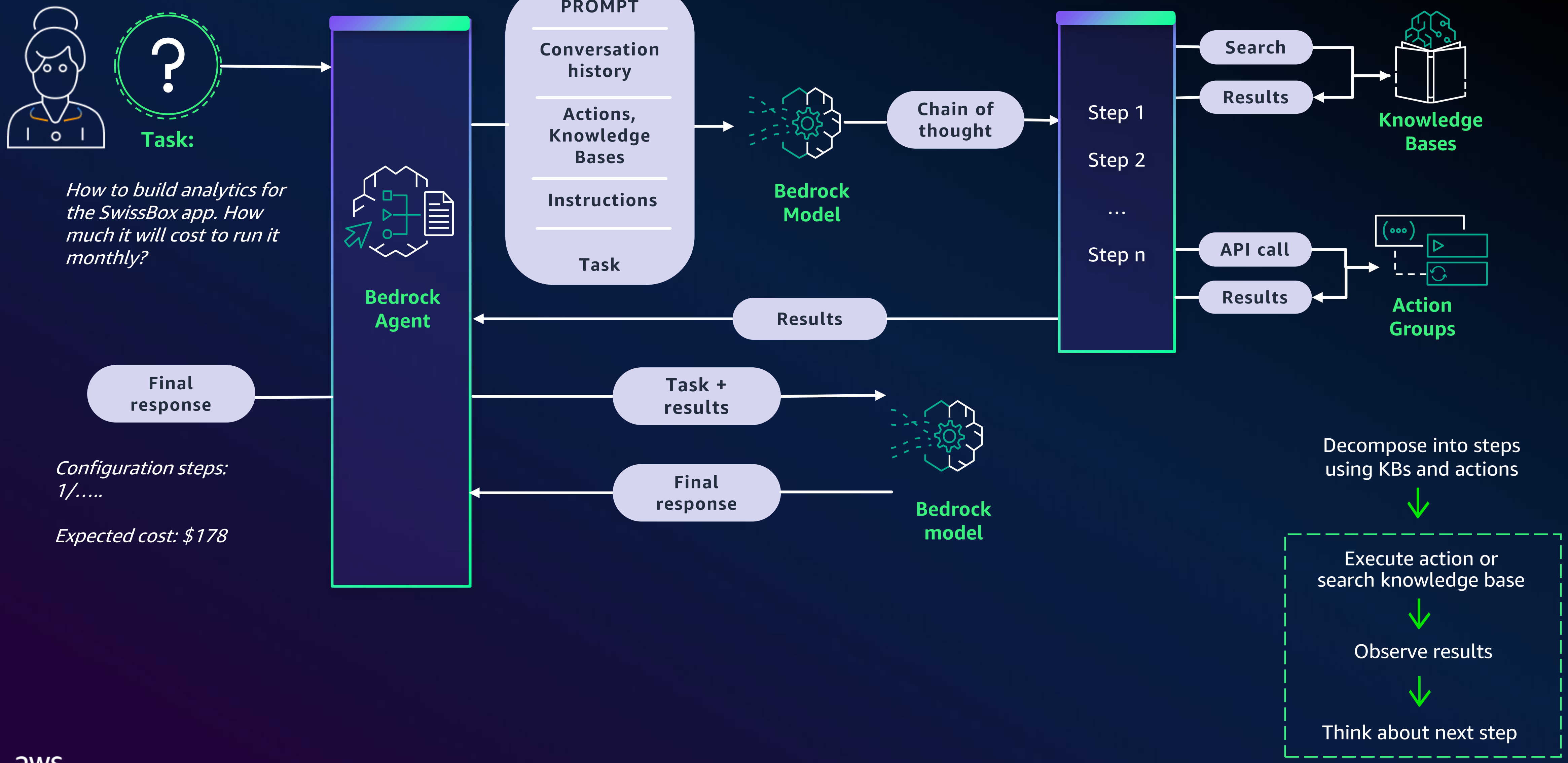
AGENT TOOLS

SA Q&A

Generate IaC

Cost Estimation

AI Agent: Detailed Flow



How to build analytics for the SwissBox app. How much it will cost to run it monthly?

Configuration steps:
1/…..

Expected cost: $178

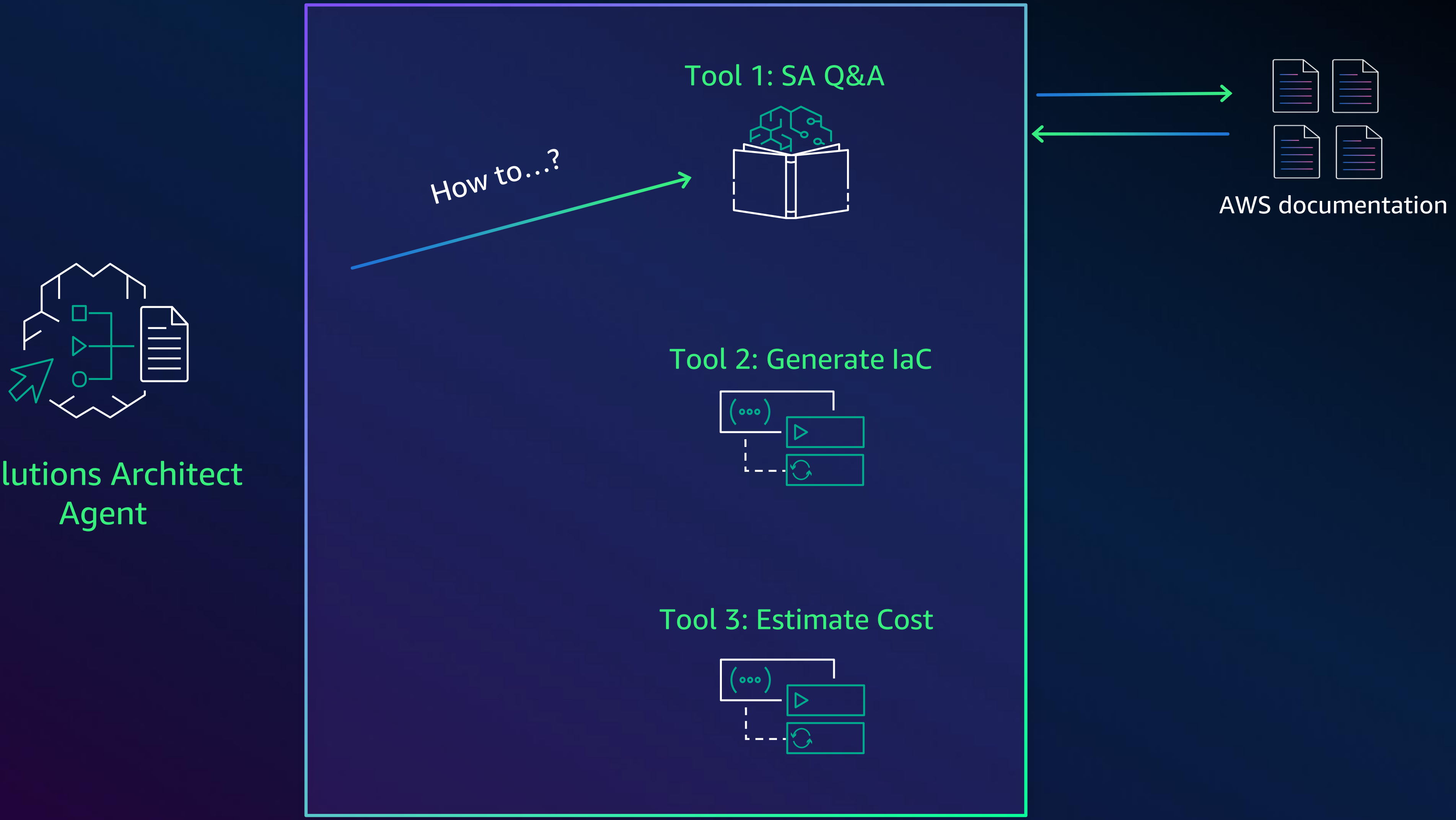
Decompose into steps using KBs and actions

Execute action or search knowledge base

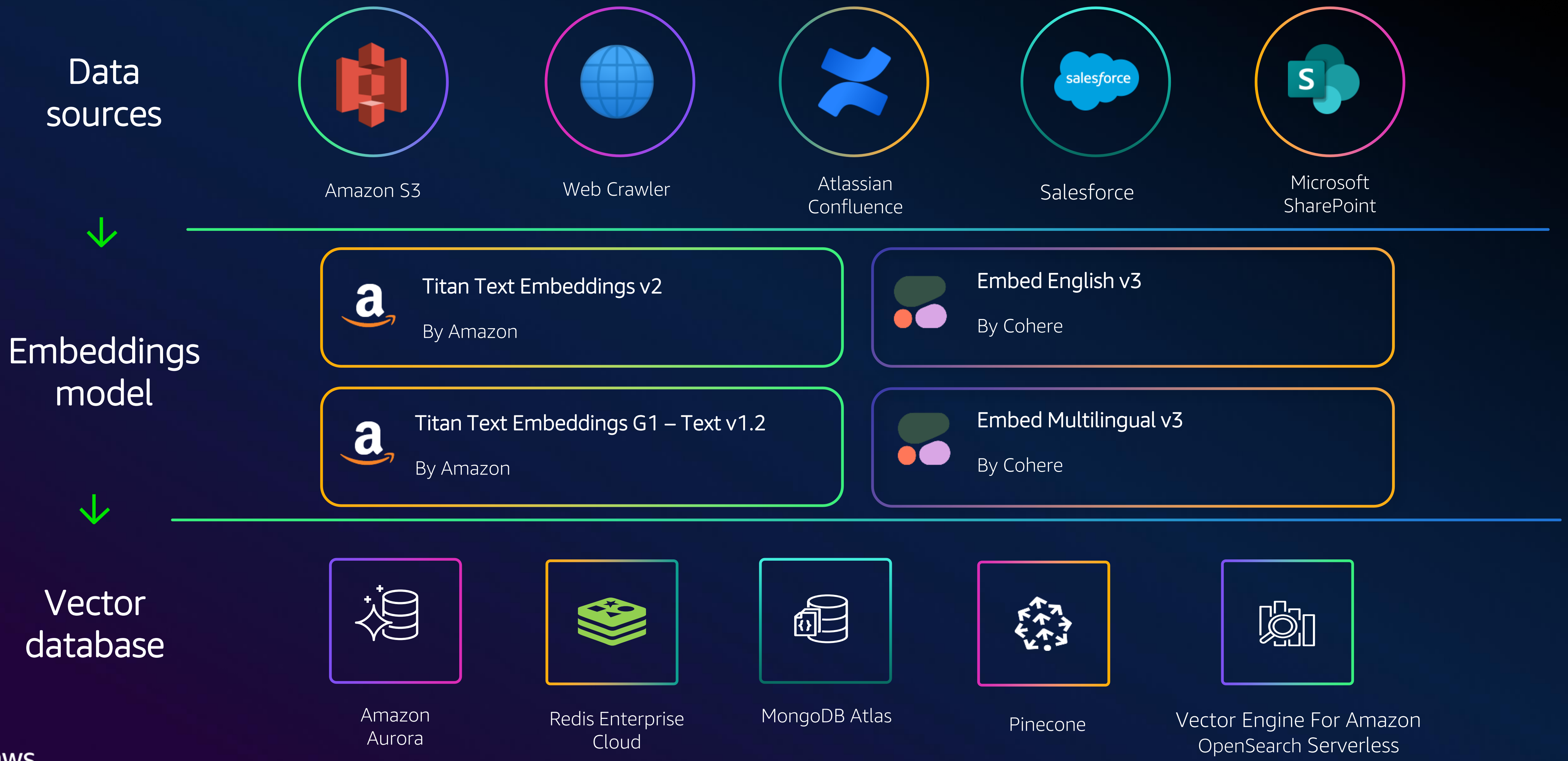
Observe results

Think about next step

Tool 1: Question & Answer Bot



Knowledge Bases



Data sources

Amazon S3

Web Crawler

Atlassian Confluence

Salesforce

Microsoft SharePoint

Embeddings model

Titan Text Embeddings v2

By Amazon

Embed English v3

By Cohere

Titan Text Embeddings G1 – Text v1.2

By Amazon

Embed Multilingual v3

By Cohere

Vector database

Amazon Aurora

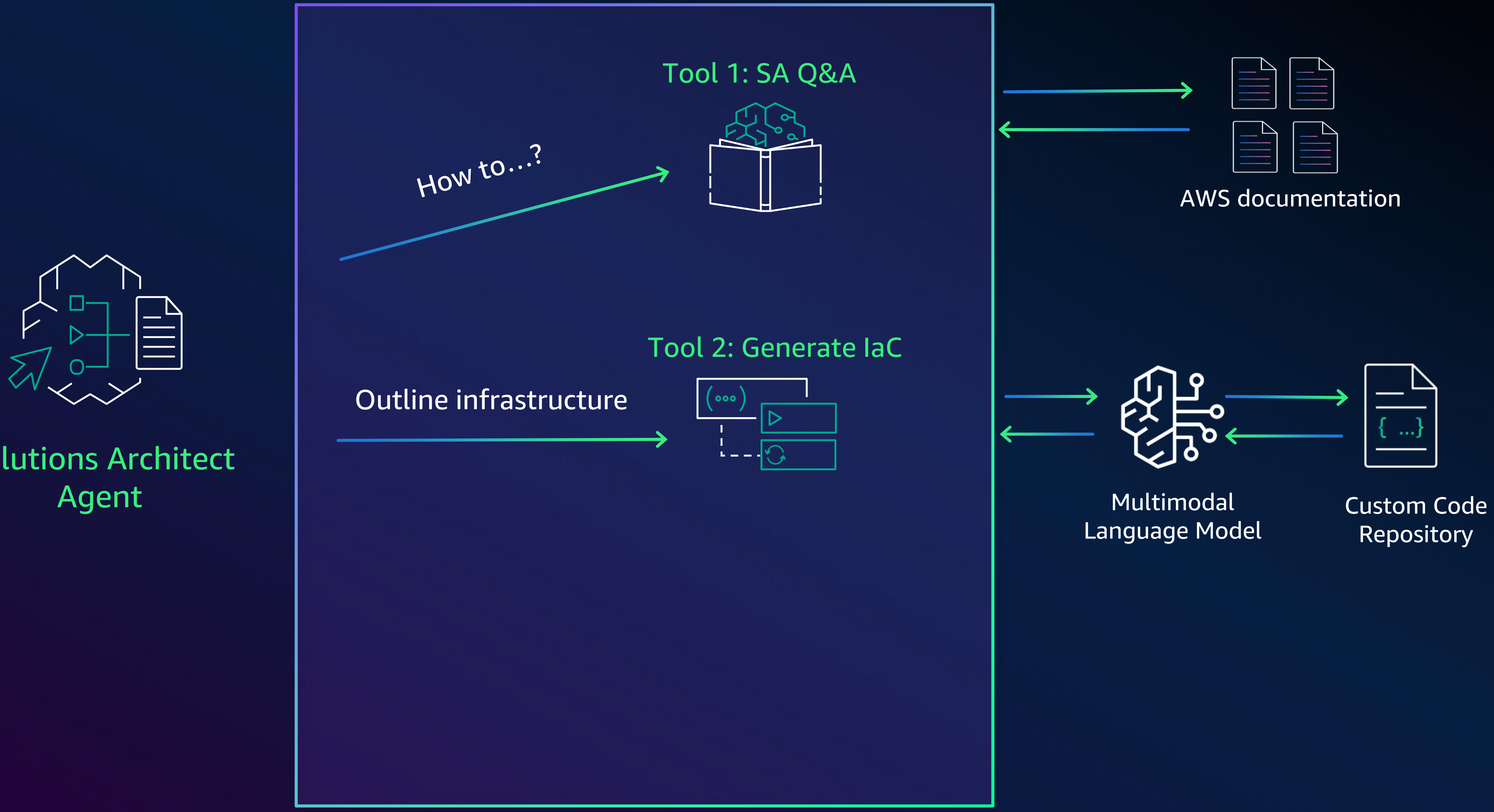
Redis Enterprise Cloud

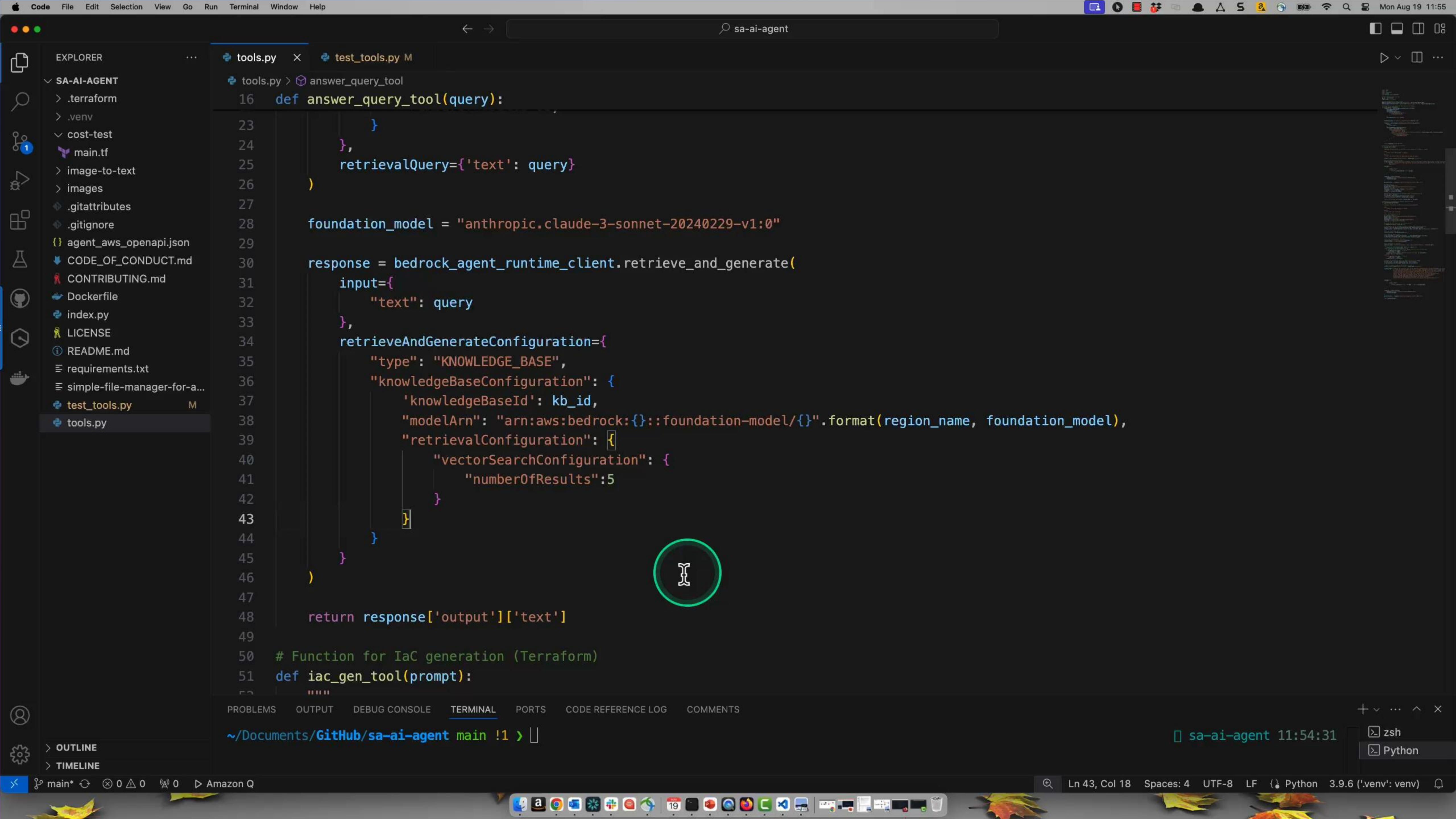
MongoDB Atlas

Pinecone

Vector Engine For Amazon OpenSearch Serverless

Tool 2: Generate IaC



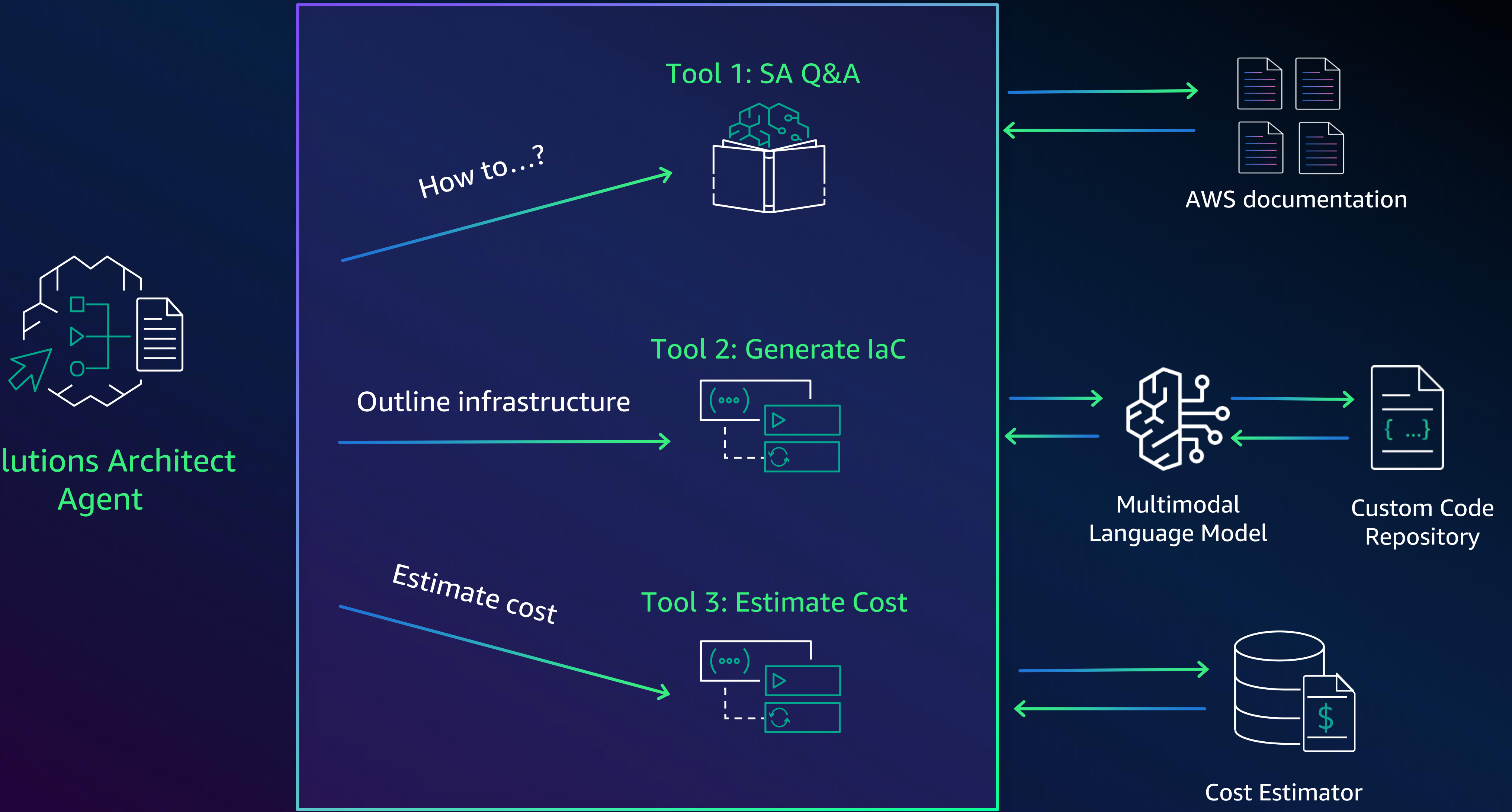


```python
def answer_query_tool(query):
```

```python
            }
        },
        retrievalQuery={'text': query}
    )

    foundation_model = "anthropic.claude-3-sonnet-20240229-v1:0"

    response = bedrock_agent_runtime_client.retrieve_and_generate(
        input={
            "text": query
        },
        retrieveAndGenerateConfiguration={
            "type": "KNOWLEDGE_BASE",
            "knowledgeBaseConfiguration": {
                'knowledgeBaseId': kb_id,
                "modelArn": "arn:aws:bedrock:{}::foundation-model/{}".format(region_name, foundation_model),
                "retrievalConfiguration": {
                    "vectorSearchConfiguration": {
                        "numberOfResults":5
                    }
                }
            }
        }
    )

    return response['output']['text']

# Function for IaC generation (Terraform)
def iac_gen_tool(prompt):
    """
```

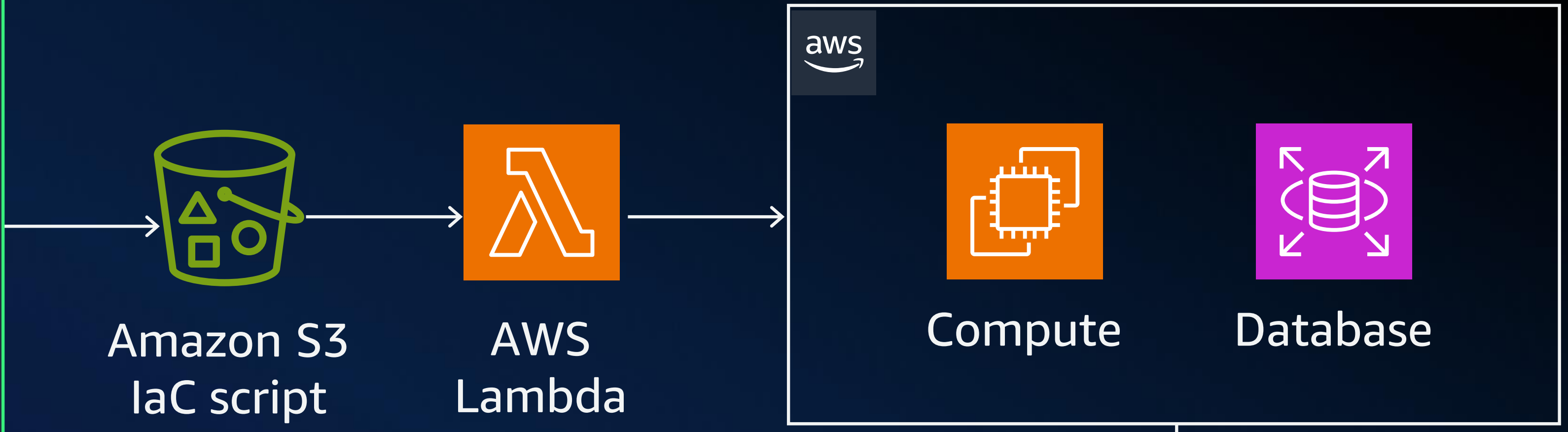
```
~/Documents/GitHub/sa-ai-agent main !1 ❯                sa-ai-agent 11:54:31
```

Tool 3: Estimate Cost

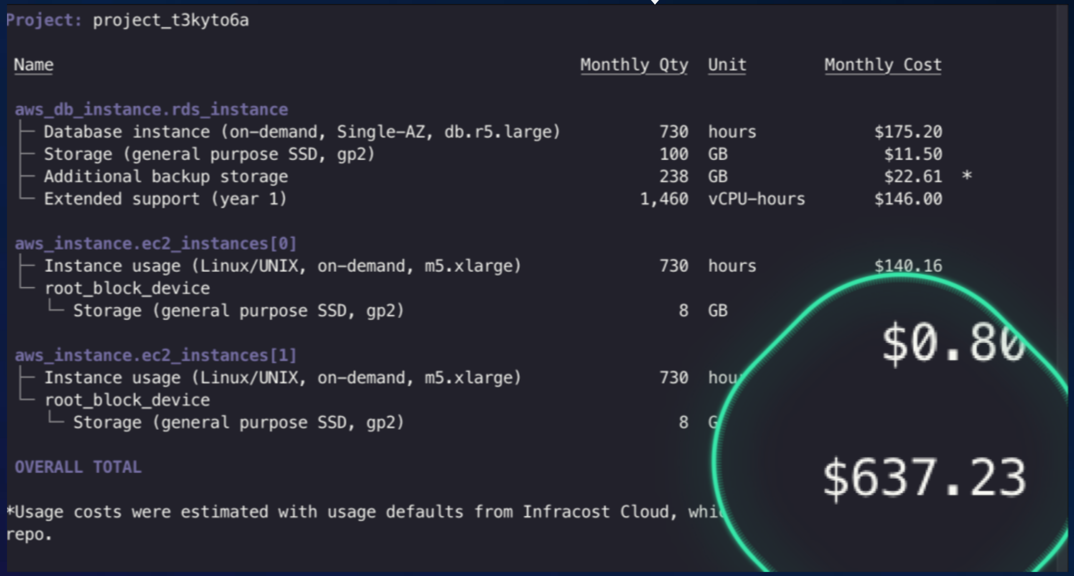


Cost Estimator: Configuration

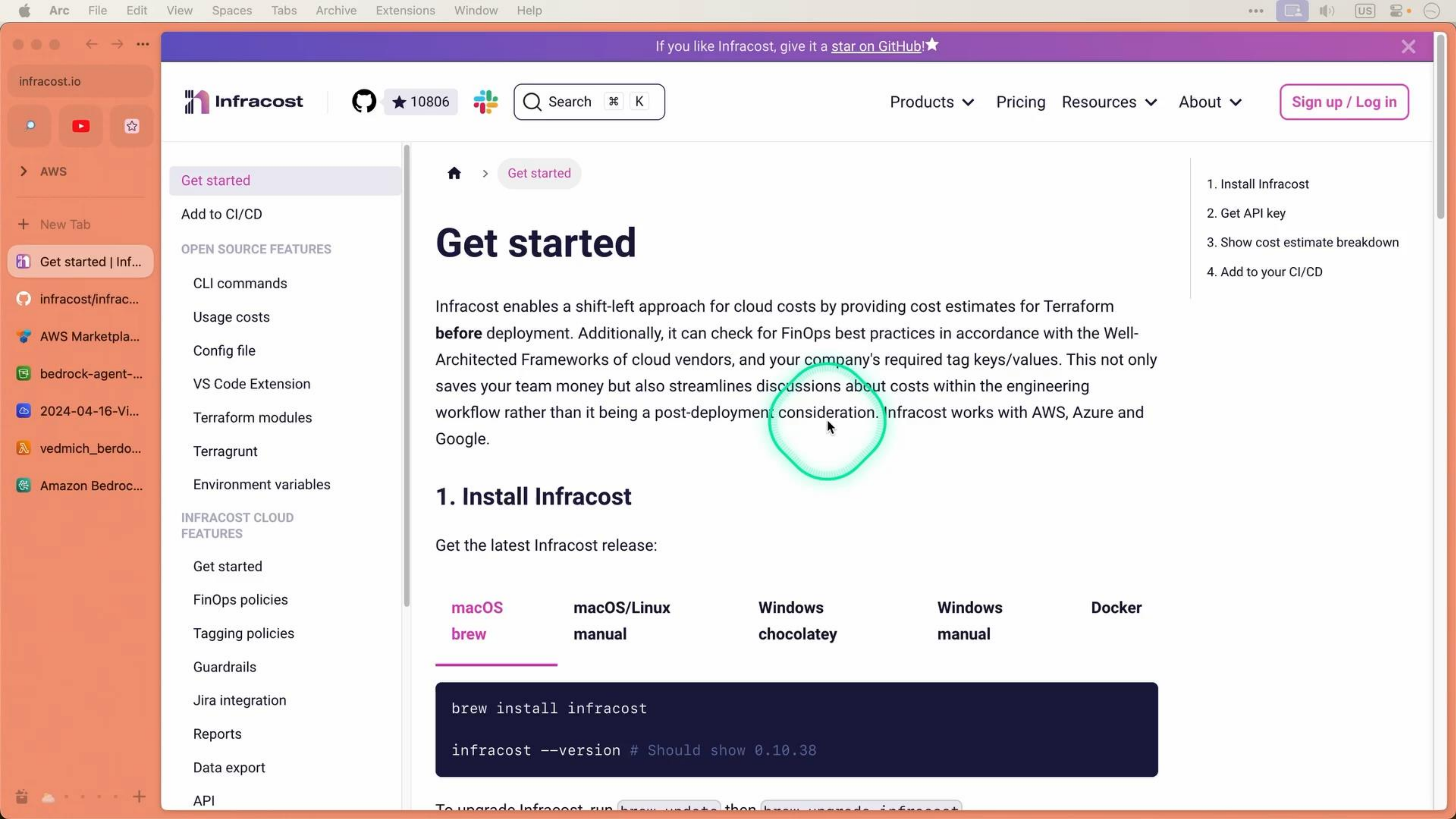
```
resource "aws_db_instance" "db" {
  allocated_storage = 100
  instance_class    = db.r5.large
  ...
}

resource "aws_instance" ”ec2" {
  count         = 2
  instance_type = m5.xlarge
  ...
}
```



Cost – $637



```
Project: project_t3kyto6a

 Name                                                    Monthly Qty  Unit         Monthly Cost

 aws_db_instance.rds_instance
 ├─ Database instance (on-demand, Single-AZ, db.r5.large)        730  hours             $175.20
 ├─ Storage (general purpose SSD, gp2)                           100  GB                 $11.50
 ├─ Additional backup storage                                    238  GB                 $22.61  *
 └─ Extended support (year 1)                                  1,460  vCPU-hours        $146.00

 aws_instance.ec2_instances[0]
 ├─ Instance usage (Linux/UNIX, on-demand, m5.xlarge)            730  hours             $140.16
 └─ root_block_device
    └─ Storage (general purpose SSD, gp2)                          8  GB                  $0.80

 aws_instance.ec2_instances[1]
 ├─ Instance usage (Linux/UNIX, on-demand, m5.xlarge)            730  hou
 └─ root_block_device
    └─ Storage (general purpose SSD, gp2)                          8  G

 OVERALL TOTAL                                                                          $637.23

*Usage costs were estimated with usage defaults from Infracost Cloud, whi
repo.
```



Get started

Infracost enables a shift-left approach for cloud costs by providing cost estimates for Terraform **before** deployment. Additionally, it can check for FinOps best practices in accordance with the Well-Architected Frameworks of cloud vendors, and your company's required tag keys/values. This not only saves your team money but also streamlines discussions about costs within the engineering workflow rather than it being a post-deployment consideration. Infracost works with AWS, Azure and Google.

1. Install Infracost

Get the latest Infracost release:

macOS brew | **macOS/Linux manual** | **Windows chocolatey** | **Windows manual** | **Docker**

```
brew install infracost

infracost --version # Should show 0.10.38
```

To upgrade Infracost, run `brew update` then `brew upgrade infracost`

INTRO
Use Case

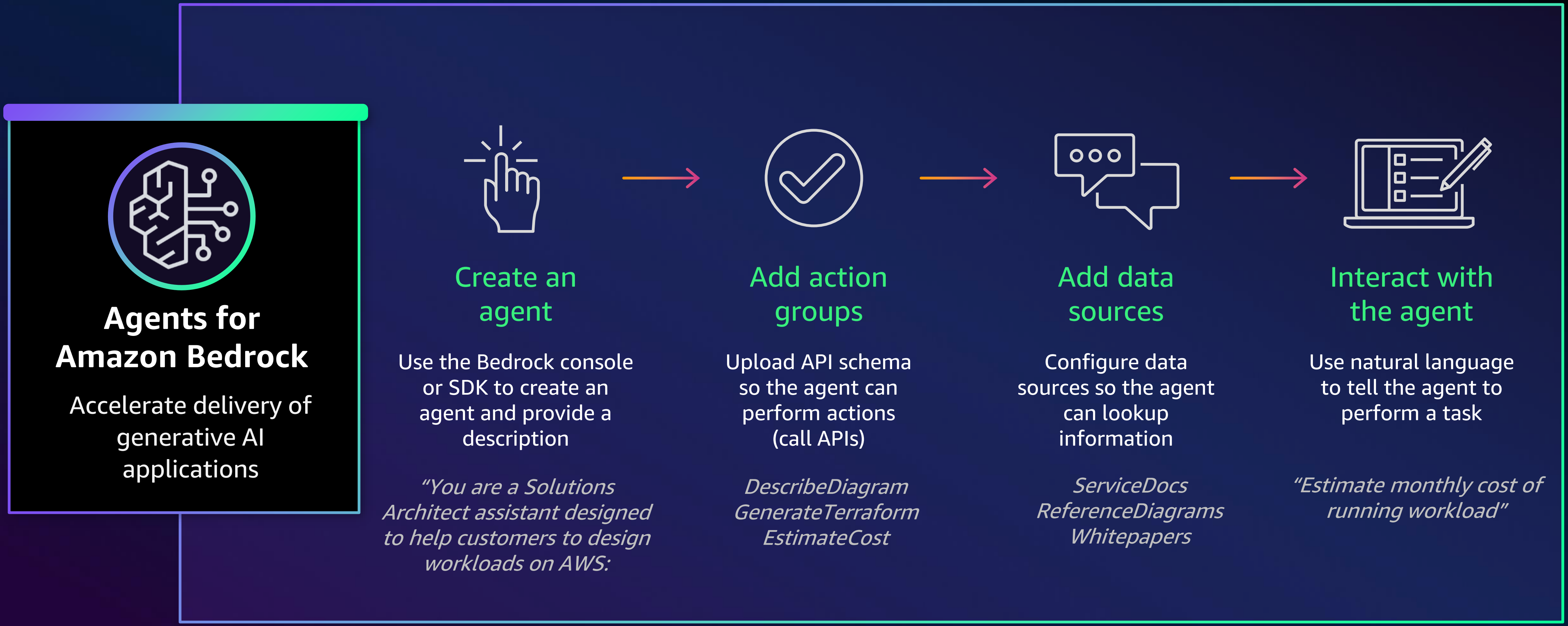
AMAZON BEDROCK
Models to choose
Model Evaluation

CUSTOMIZATION
Prompt Engineering
Knowledge Base (RAG)

AGENT TOOLS
SA Q&A
Generate IaC
Cost Estimation

ORCHESTRATION
Agents

Bedrock Agent



Agents for Amazon Bedrock

Accelerate delivery of generative AI applications

Create an agent

Use the Bedrock console or SDK to create an agent and provide a description

“You are a Solutions Architect assistant designed to help customers to design workloads on AWS:

Add action groups

Upload API schema so the agent can perform actions (call APIs)

DescribeDiagram
GenerateTerraform
EstimateCost

Add data sources

Configure data sources so the agent can lookup information

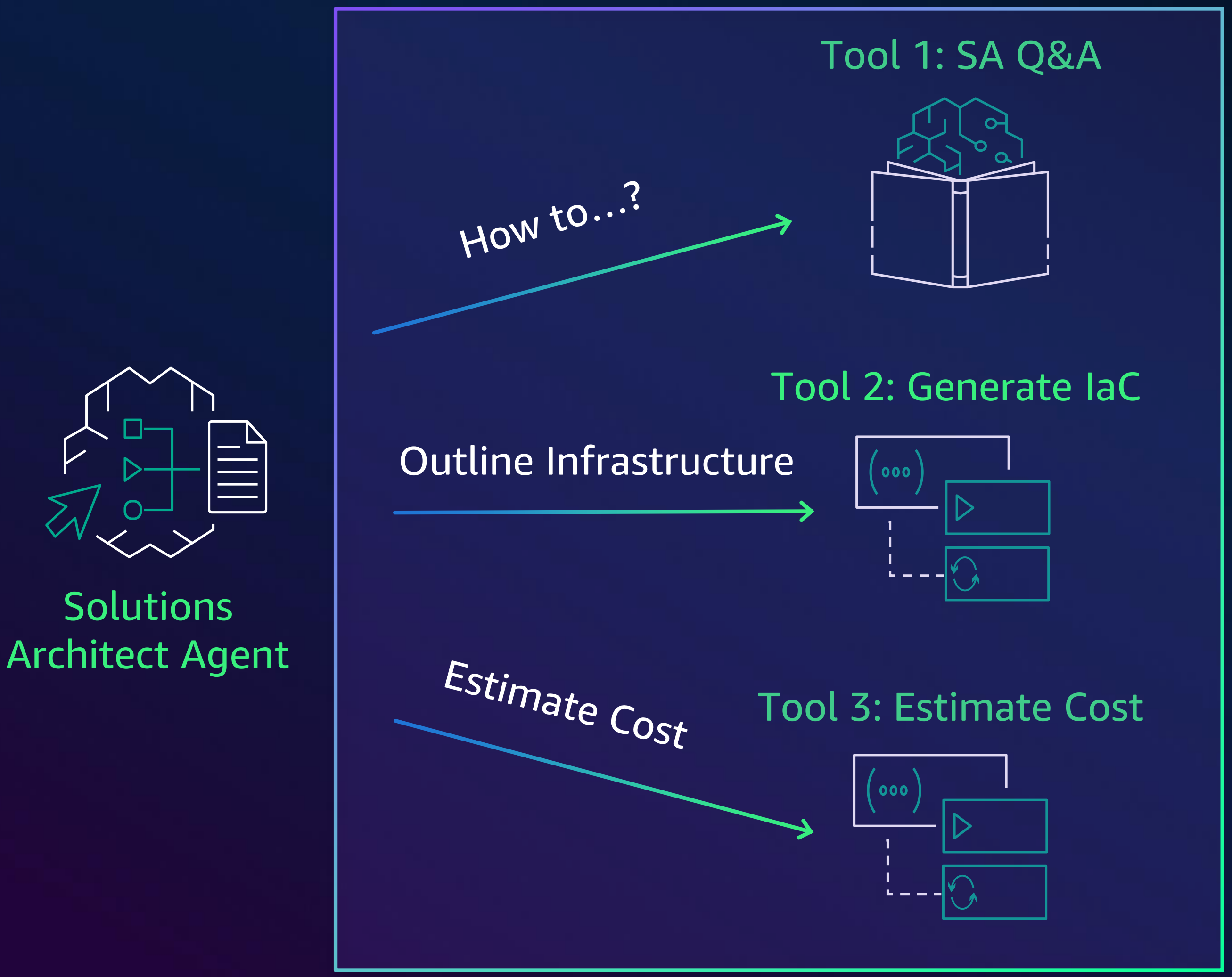
ServiceDocs
ReferenceDiagrams
Whitepapers

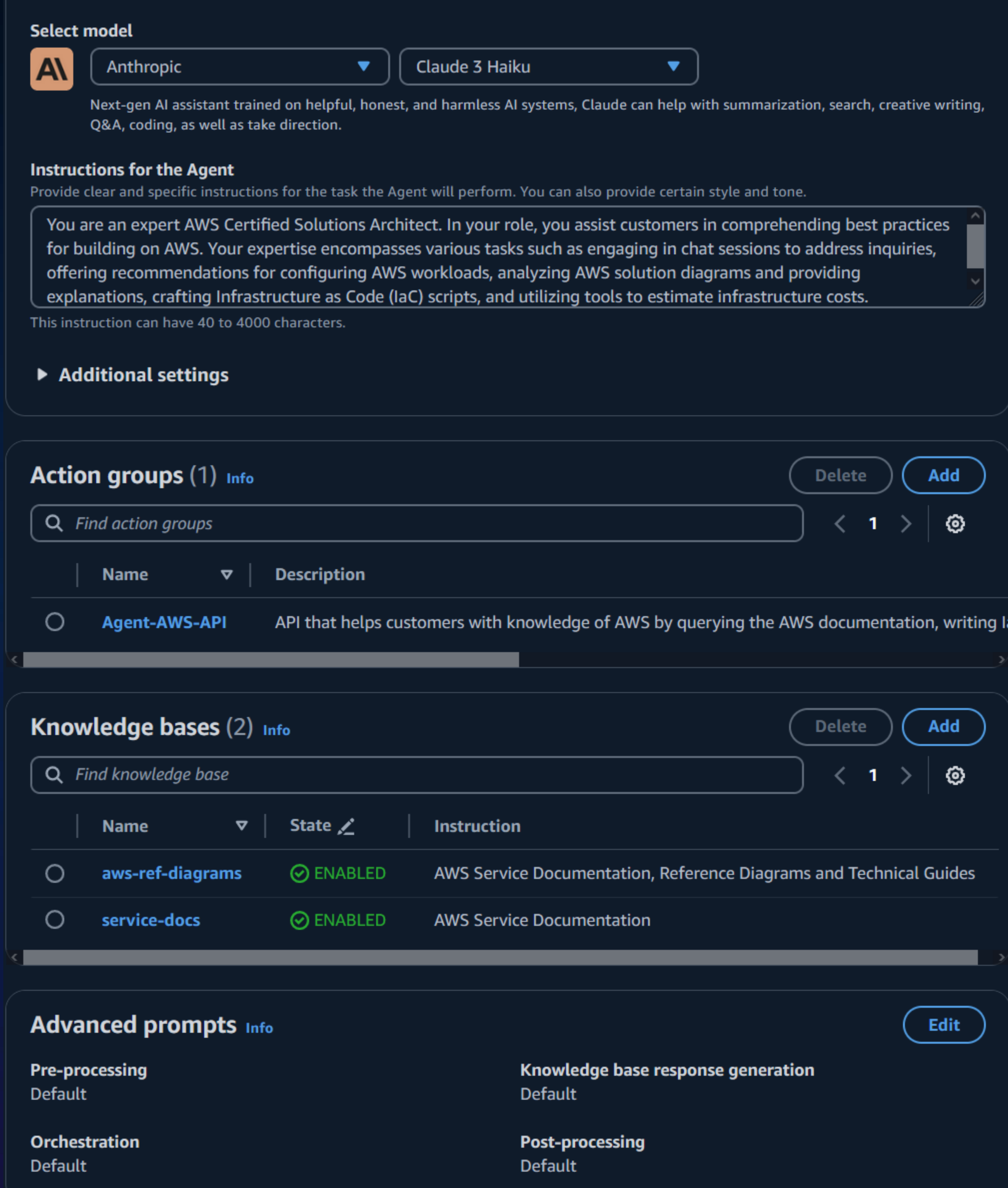
Interact with the agent

Use natural language to tell the agent to perform a task

“Estimate monthly cost of running workload”

Agents builder





Select model

Anthropic ▼

Claude 3 Haiku ▼

Next-gen AI assistant trained on helpful, honest, and harmless AI systems, Claude can help with summarization, search, creative writing, Q&A, coding, as well as take direction.

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

You are an expert AWS Certified Solutions Architect. In your role, you assist customers in comprehending best practices for building on AWS. Your expertise encompasses various tasks such as engaging in chat sessions to address inquiries, offering recommendations for configuring AWS workloads, analyzing AWS solution diagrams and providing explanations, crafting Infrastructure as Code (IaC) scripts, and utilizing tools to estimate infrastructure costs.

This instruction can have 40 to 4000 characters.

▸ **Additional settings**

Action groups (1) Info

	Name	Description
○	Agent-AWS-API	API that helps customers with knowledge of AWS by querying the AWS documentation, writing I

Knowledge bases (2) Info

	Name	State	Instruction
○	aws-ref-diagrams	✓ ENABLED	AWS Service Documentation, Reference Diagrams and Technical Guides
○	service-docs	✓ ENABLED	AWS Service Documentation

Advanced prompts Info

Pre-processing
Default

Knowledge base response generation
Default

Orchestration
Default

Post-processing
Default

Each Action Group has 3 key elements

Action Group Description

Overview of actions provided – helps agent know when this action group is relevant

API Schema

- Rich definition of each action
- Operation name, input parameters, data types, response details
- Helps agents know when to use it, how to call it, and how to use results
- Language agnostic API definition using industry-standard schema

Lambda Function

- Implementation of each action
- Contains either business logic or wraps microservices, databases, or tools
- Serverless, scalable, secure
- Choice of programming language (Python, C#, JavaScript, Java, …)

INTRO
Use Case

AMAZON BEDROCK
Models to choose
Model Evaluation

CUSTOMIZATION
Prompt Engineering
Knowledge Base (RAG)

AGENT TOOLS
SA Q&A
Generate IaC
Cost Estimation

ORCHESTRATION
Agents

WRAP-UP
Takeaways
Resources

3 Key Takeaways

Evaluate different models on cost, speed, and efficiency using Amazon Bedrock's single-API access.

Secure AI Integration with customizable FMs that can be tailored to specific business needs using proprietary data.

Automate complex, multi-step tasks by breaking them down into smaller, manageable actions with Bedrock Agents

aws

Additional resources

GitHub repo: Amazon Bedrock samples

This repository contains pre-built examples to help customers get started with the Amazon Bedrock service including: Knowledge Bases, RAG, Agents, Bedrock Fine-tuning, Security and Governance

GitHub repo: Build AWS SA using Amazon Bedrock agents

This repository contains instructions and code samples to help customers This repository contains instructions and code samples for building an AWS Solutions Architect Agent with Amazon Bedrock (SA Q&A, Generate IaC, Estimate Cost)

PartyRock app: Generative AI Agents For Amazon Bedrock

This prototype app lets you describe an API, and generate code for Agents for Amazon Bedrock. The generated code includes: an OpenAPI schema providing a rich description of the API, Python Lambda function implementation based on the generated API schema, test suite for the API, delivered as a json array of Lambda event payloads based on the generated API schema.

Thank you!

Please complete the session survey in the mobile app

Viktoria Semaan

linkedin.com/in/semaan

Viktor Vedmich

linkedin.com/in/vedmich