# AWS State, Local, and Education Learning Days

## Madison, Wisconsin

**3:15pm – 4:15pm**

**300** level

**Generative AI Masterclass**

A comprehensive masterclass on AI, covering technology evolution, implementation strategies, responsible practices.

aws Learning Days
State, Local, and Education

# Generative AI Master Class


**Shashank Tanksali**

Sr. Solutions Architect
GenAI + Security

**Norman Owens**

Sr. Solutions Architect
GenAI + Security



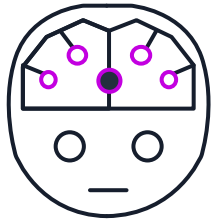

aws Learning Days
State, Local, and Education

# AI Masterclass

- Setting the stage
- What is Amazon Bedrock?
- GenAI Governance
- Responsible AI
- Agentic AI
- Closing Comments

# Setting the stage

# AIML/GenAI hierarchy

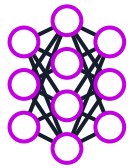


## Artificial Intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning

## Machine learning (ML)

A subset of AI that uses machines to search for patterns in data to build logic models automatically



## Deep learning (DL)

A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

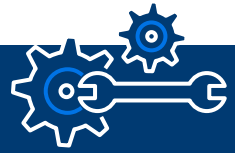
## Generative AI

Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

# How does Gen AI work?

AI that can produce original content close enough to human generated content for real-world tasks

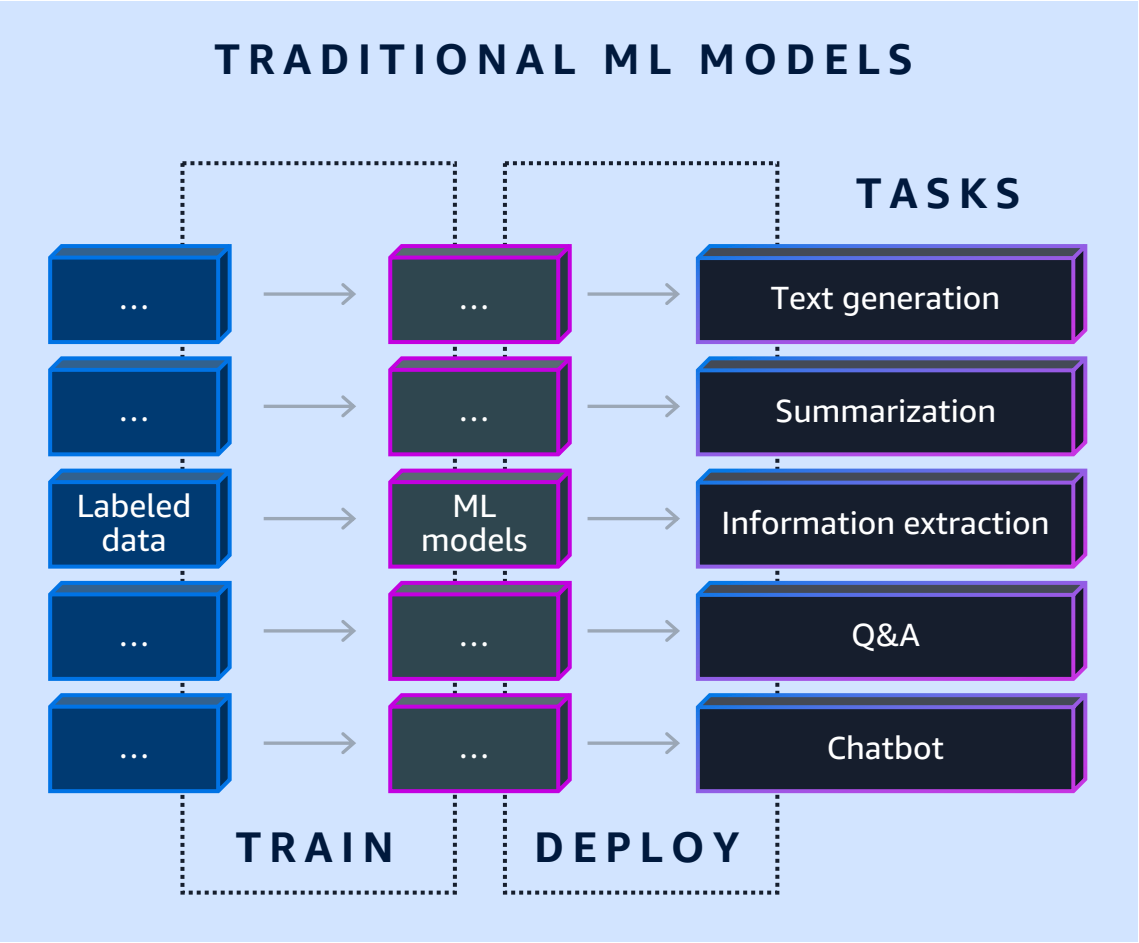
Powered by foundation models pre-trained on large sets of data with several hundred billion parameters

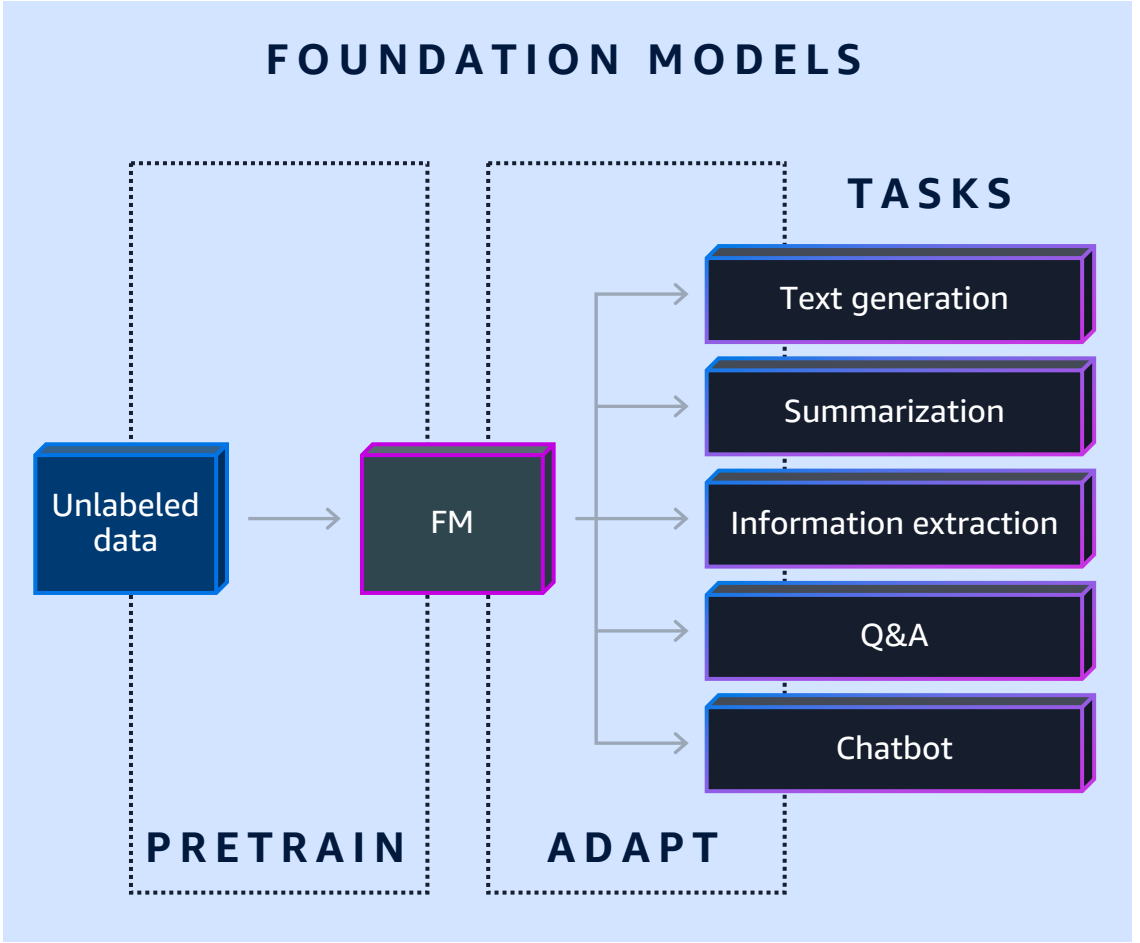


Tasks can be customized for specific domains with minimal fine-tuning

Applicable to many use cases like text summarization, question answering, digital art creation, code generation, etc.

# Why foundation models?



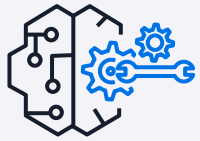



# GenAI in Action

# What is Amazon Bedrock

# Amazon Bedrock is a service that has everything you need to build, deploy and scale generative AI applications

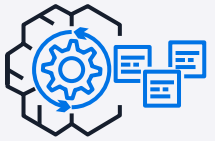
Choose the best model



Optimize for cost, latency, and accuracy

Securely customize with your data

Apply safety and responsible AI checks



Deploy and operate Agents

# Customers are using Amazon Bedrock to deploy and scale business-critical use cases

## ENHANCE CUSTOMER EXPERIENCES

- Chatbots
- Virtual Assistants
- Conversation Analytics
- Personalization

## BOOST EMPLOYEE PRODUCTIVITY & CREATIVITY

- Conversational Search
- Summarization
- Content Creation
- Code Generation
- Data To Insights

## OPTIMIZE BUSINESS PROCESSES

- Document Processing
- Data Augmentation
- Fraud Detection
- Process Optimization

# GenAI Implementation Strategies

# Strategies for implementation and their trade-offs

1

**Prompt engineering**

Using context instructions, examples and output indicators

2

**Retrieval Augmented Generation (RAG)**

Text generation based on specified corpus of data, to generate accurate responses with no hallucination

3

**Agentic Systems**

Using LLMs to understand tasks, gather information, decide best path and execute actions

4

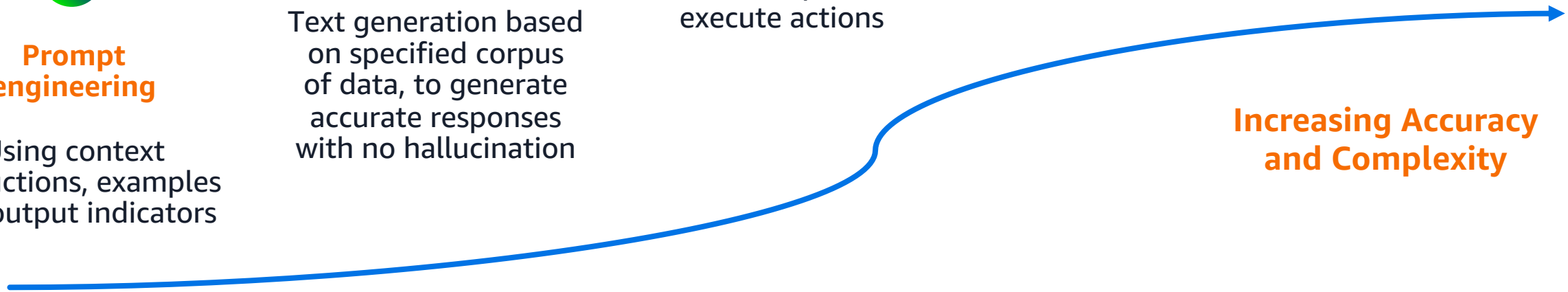
**Fine-tuning**

Give a model company or domain specific data, or instructions examples

5

**Re-training**

Give a model a different dataset, or building a model from scratch



# Prompt Engineering

## Zero shot learning

Prompt

Review: “Earnings per share have beaten analyst expectations"

What is the sentiment?

Input

The text explains that earnings have been expectations, that is generally a good signal in financial reporting, therefore the review is positive.

Output

## Few shot learning

Prompt

Review: " Earnings per share have beaten analyst expectations "
Sentiment: positive

Review: “sales remained constant over the past quarter, but EBITDA has decreased"
Sentiment: negative

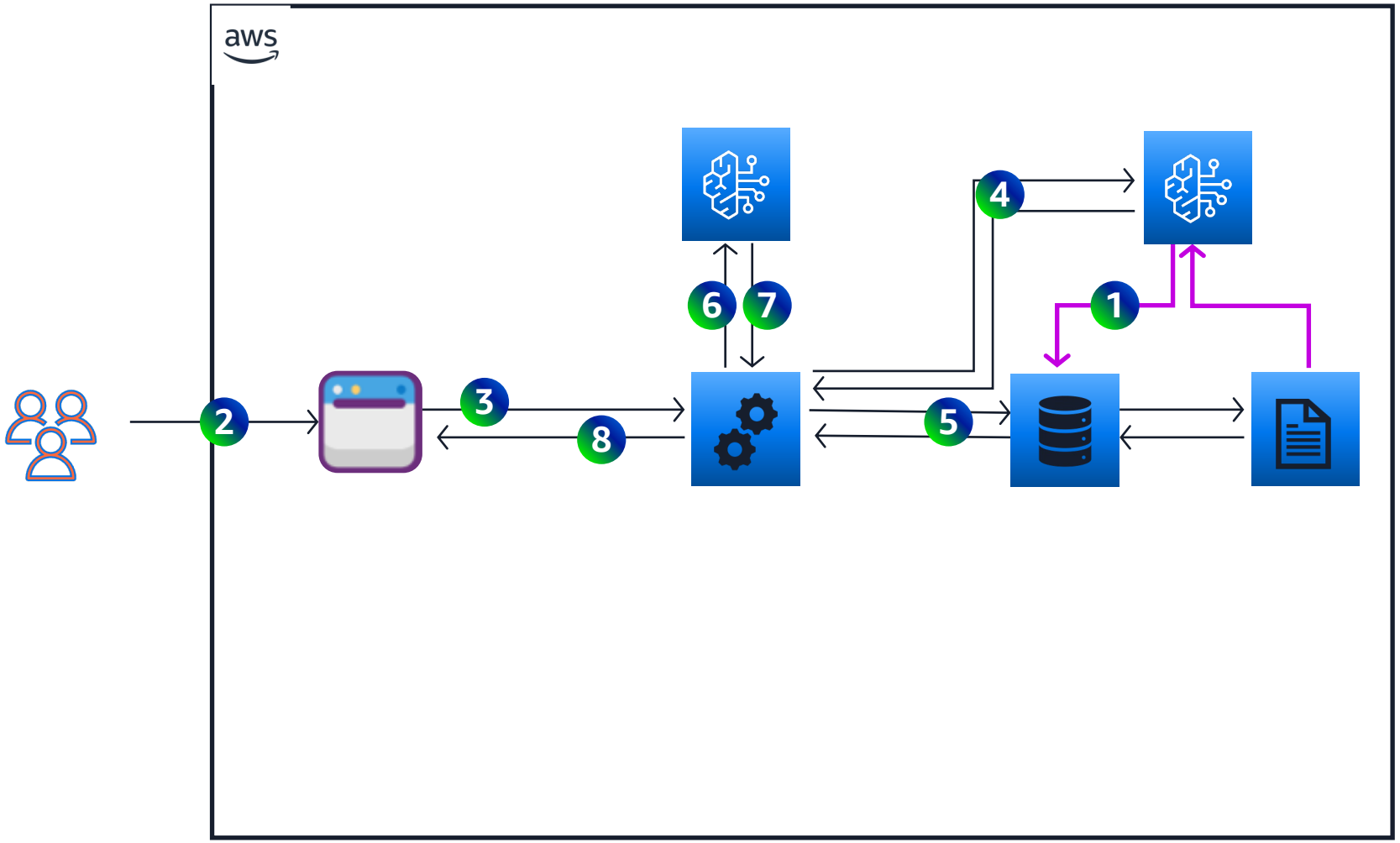
Review: “S&P500 Tops 5,600 for first time as tech rallies"
Sentiment:

Positive

Input

Output

# Retrieval augmented generation (RAG) – “Chat with my Docs”




1. Documents are converted into vectors using the embeddings model on Bedrock and then they are stored in the vector DB
2. User sends a question
3. Question is sent for processing
4. Request sent to converts the prompt in embeddings
5. With the embedding we search the paragraphs that contain the answer to the question
6. Those paragraphs together with the prompt are sent to the model
7. The model generates the answer
8. The answer is sent it back

# Agents combine Actions and Knowledge Bases

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

## HR Policy Assistant

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

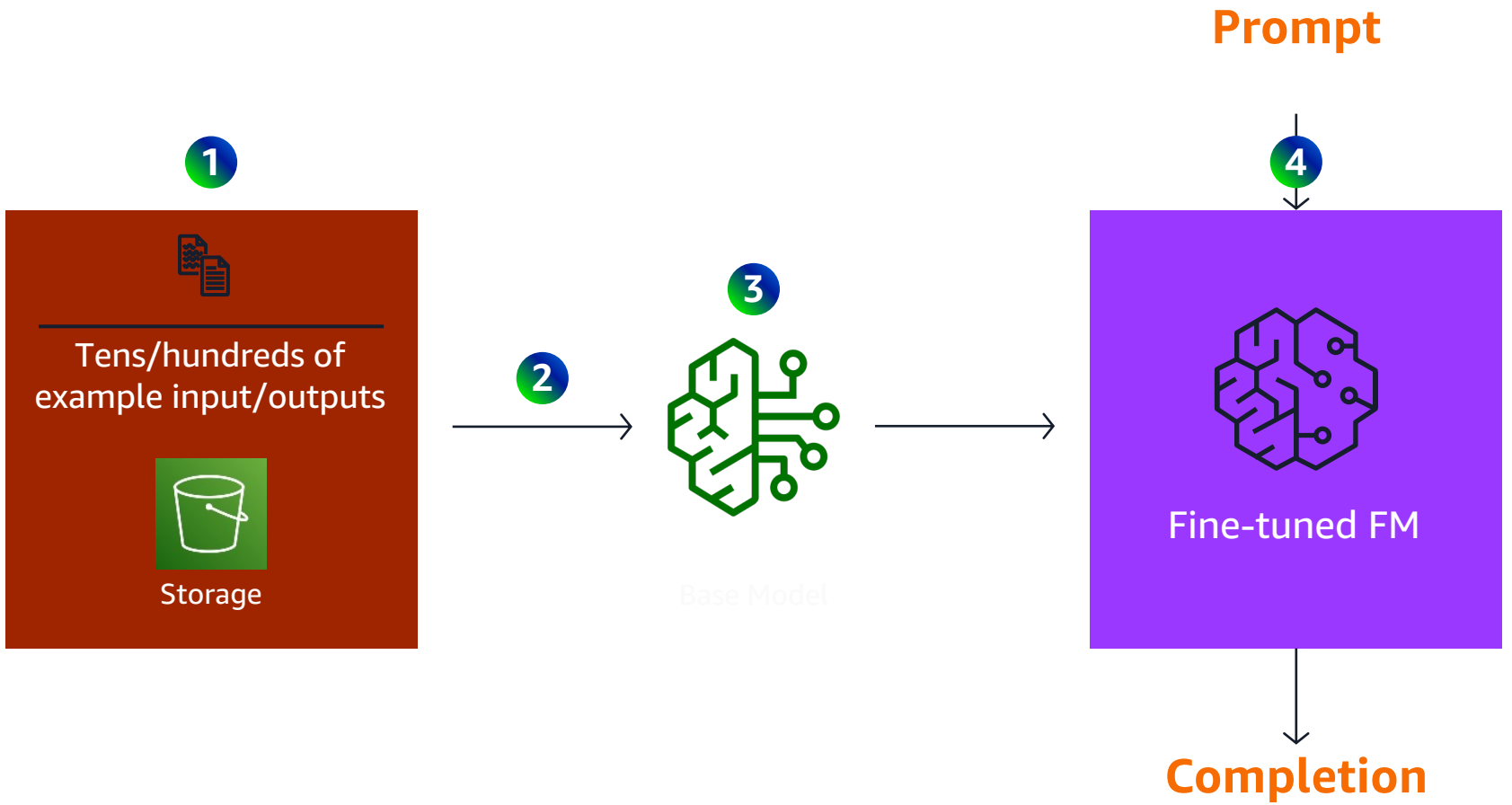
### HR Knowledge Base

Vacation Policy
Contains the entire vacation policy for the company

### HR Actions

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

# Fine-tuning (Task specific)

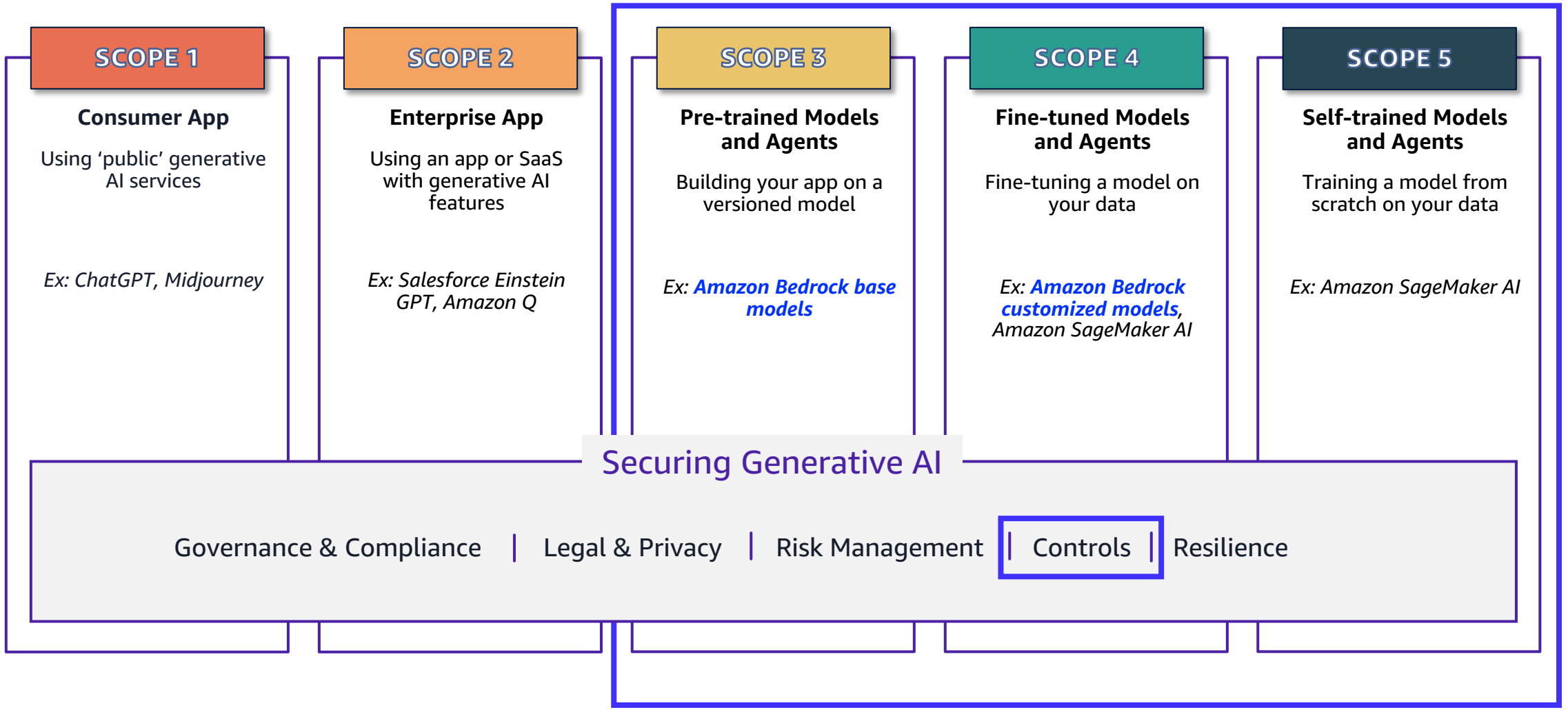



1. Tens/hundreds of instructions or examples are stored
2. Instructions are sent to the base model via APIs
3. A base model is fine-tuned with these instructions (minutes, hours, days)
4. The new fine-tuned model is available for the customer to start inferencing

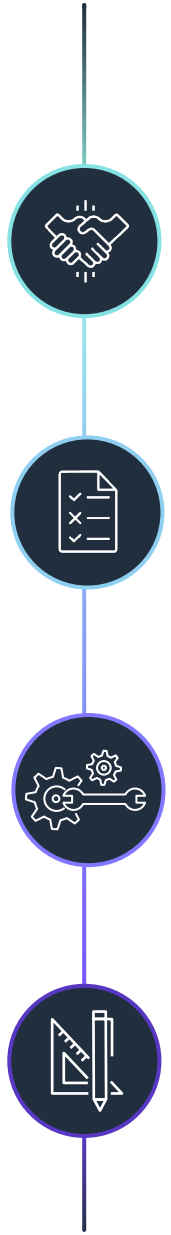
# GenAI Governance

# Generative AI Security Scoping Matrix

## A MENTAL MODEL TO CLASSIFY USE-CASES



# Responsible AI:
# Best practices



Put your people first

Assess risk on a (use) case-by-case basis

Iterate across the AI lifecycle

Test, test again, and then test again

# Responsible AI Dimensions

### FAIRNESS

Considering impacts on different groups of stakeholders

### EXPLAINABILITY

Understanding and evaluating system outputs

### CONTROLLABILITY

Having mechanisms to monitor and steer AI system behavior

### SAFETY

Preventing harmful system output and misuse

### PRIVACY & SECURITY

Appropriately obtaining, using and protecting data and models

### GOVERNANCE

Incorporating best practices into the AI supply chain, including providers and deployers

### TRANSPARENCY

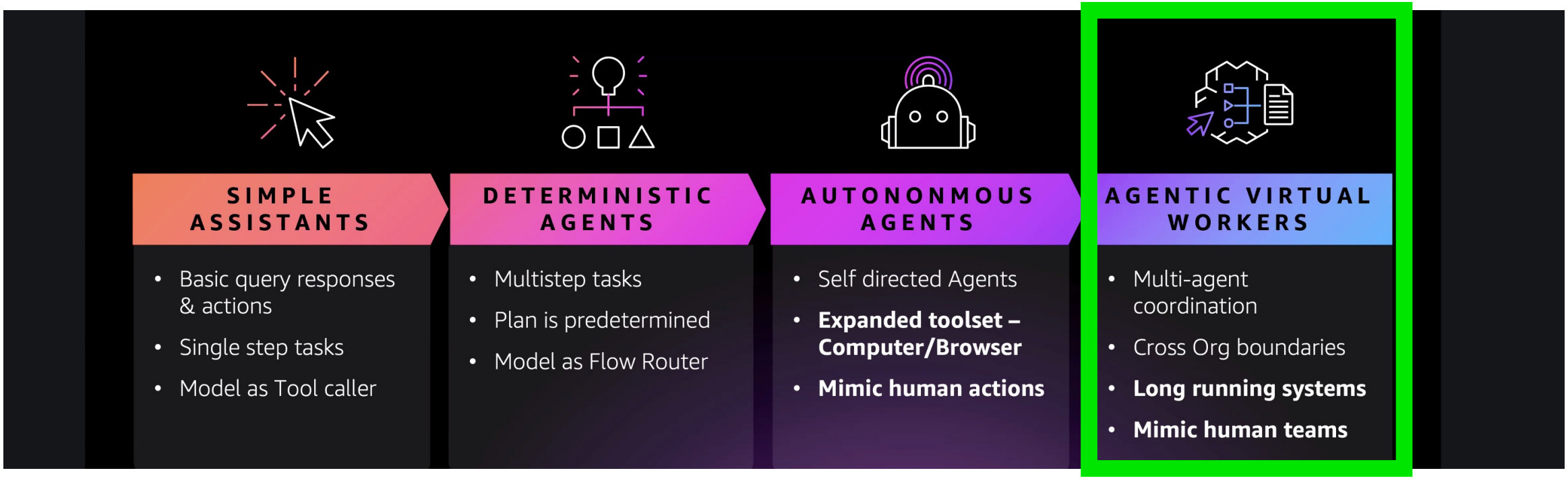
Enabling stakeholders to make informed choices about their engagement with an AI system

### VERACITY & ROBUSTNESS

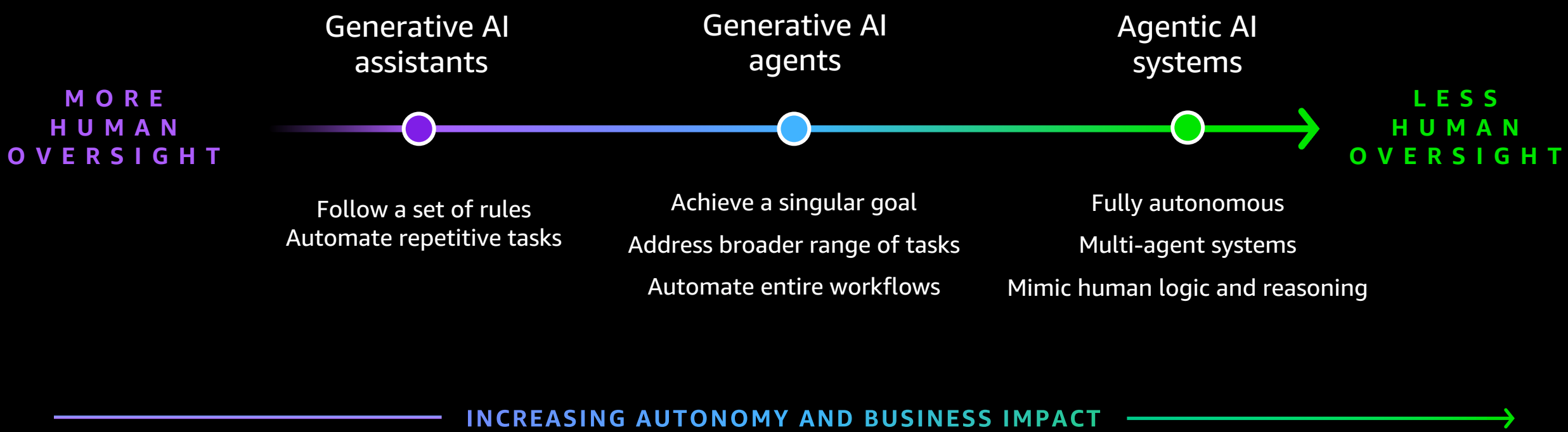
Achieving correct system outputs, even with unexpected or adversarial inputs

# Agentic AI : Bedrock AgentCore

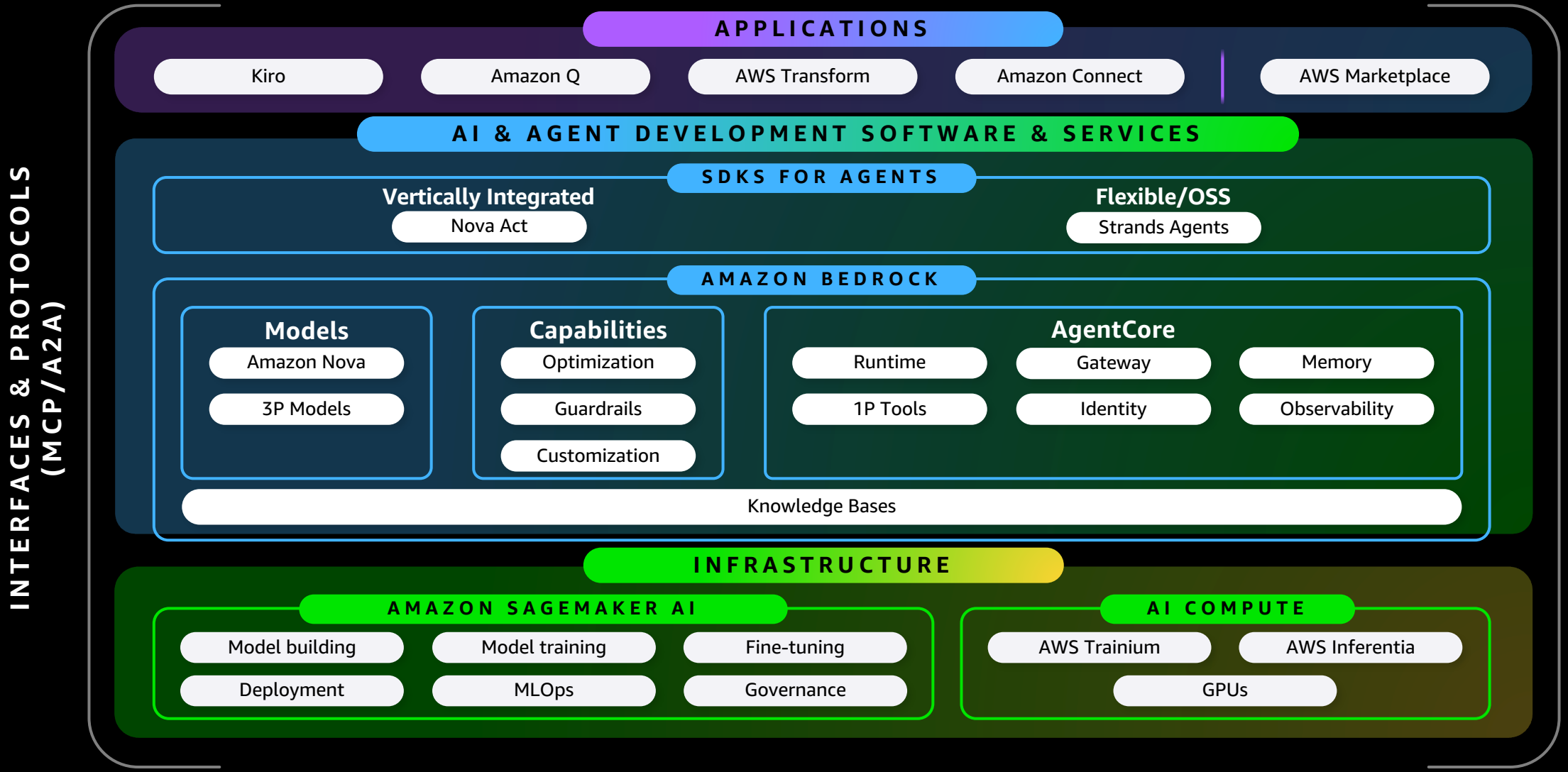
# Types of Gen AI



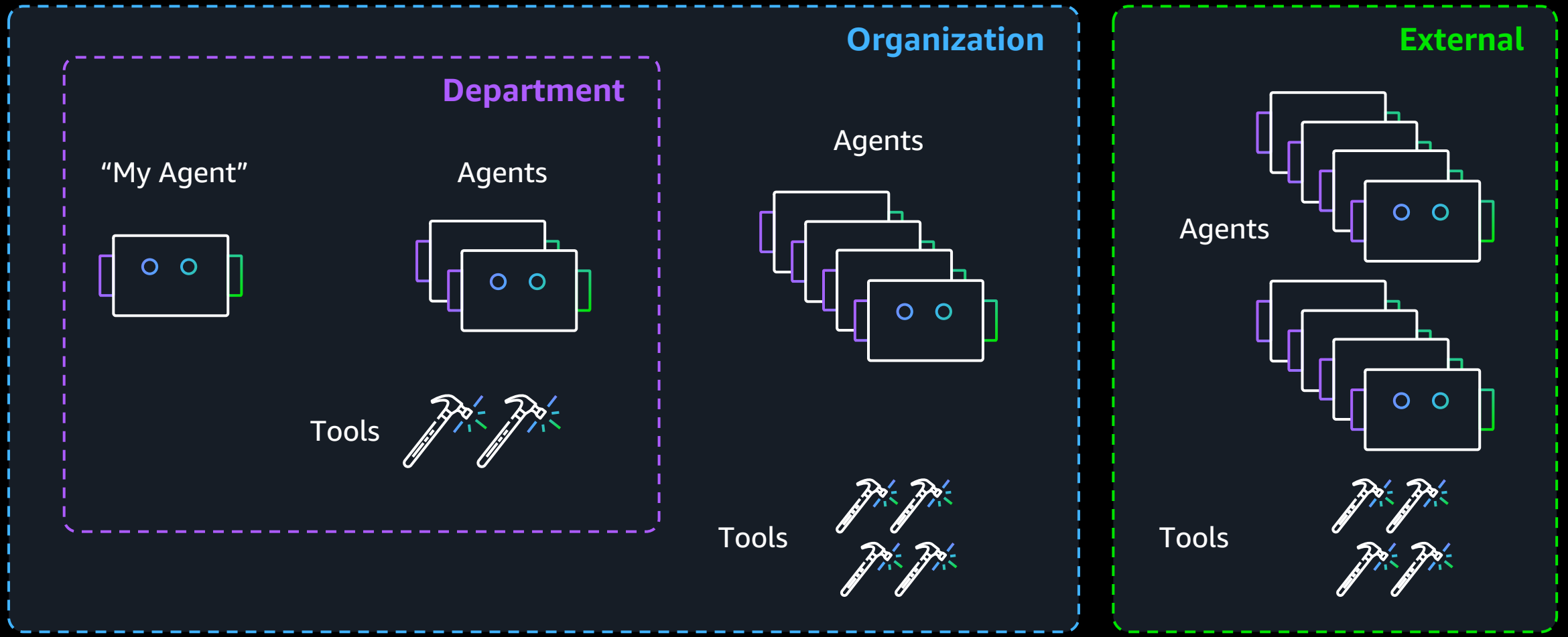
# The evolution into Agentic AI



# AWS AI Stack



# Building an agent is the start, production scale is the goal



# The prototype to production “chasm”

Excitement and potential

POC

Challenges on the path to production

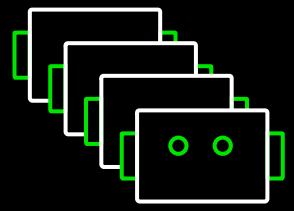
Performance

Scalability

Security

Governance

Meaningful business value



AI production agents

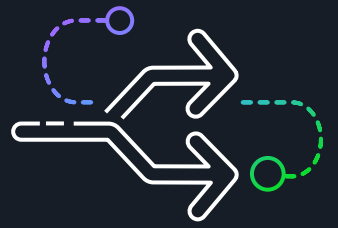
# Amazon Bedrock AgentCore

## Deploy and operate highly capable agents securely, at scale using any framework and model

### TIME TO VALUE

Build powerful AI agents without the infrastructure and operational headaches

### FLEXIBLE

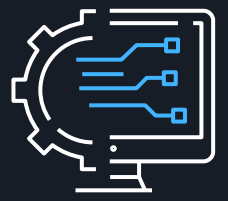


Create agents with any framework or model

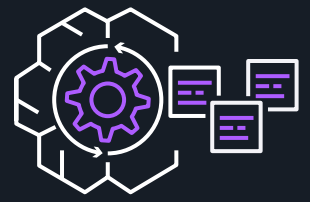
### TRUSTED

Deploy secure, scalable, and reliable agents your organization can trust

# Foundational services for running highly capable agents, securely at scale



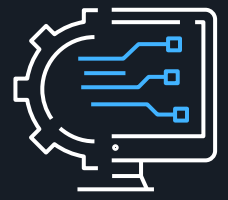
Deploy securely at scale



Enhance with tools and memory

Monitor

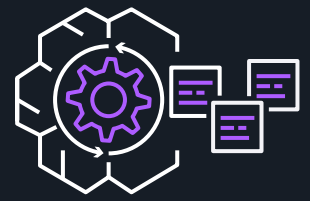
# Secure, scalable runtime for agents



AgentCore Runtime

AgentCore Identity

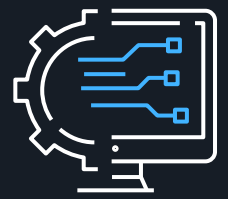
Deploy securely at scale



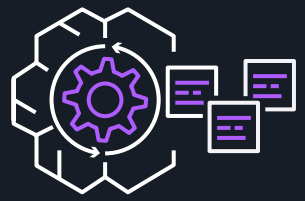
Enhance with tools and memory

Monitor

# Essential tools and capabilities to build highly effective agents



Deploy securely at scale

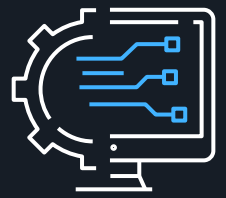


Enhance with tools and memory

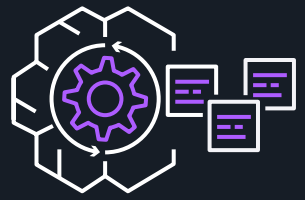
- AgentCore Gateway
- AgentCore Memory
- AgentCore Browser
- AgentCore Code Interpreter

Monitor

# Visibility to operate agents you can trust



Deploy securely at scale



Enhance with tools and memory

Monitor

AgentCore Observability

# Success Stories

# Top education agentic AI use cases

| | Accelerated transcript processing | Streamlined education help desk | Supply chain optimization |
|---|---|---|---|
| PROBLEM | Manual processing of diverse educational documents across multiple formats and languages creates inefficient, costly, and error-prone communication between institutions and stakeholders | Staff shortages at higher education institutions are hampering their ability to effectively deliver student support services and manage varying inquiry volumes | Teachers and advisors spend significant time manually gathering and analyzing student information from various sources, limiting their time to spend more time on teaching and advising |
| USE CASE | | | |
| OUTCOME | Streamlined evaluation process for faster enrollment decision-making<br>Enabled new approaches for predictive analytics to support enrollment and fiscal management<br>Reduced manual efforts and time spent on reviewing international transcripts and course articulation | Deliver personalized 24/7 student support across multiple departments<br>Automatically route inquiries and understand intent through natural language<br>Virtual agents resolve routine issues while escalating complex inquiries to faculty and staff | Generate student insight through natural language queries<br>Generate student information and analysis in minutes<br>Proactively identify at risk students with recommendations based on individual student goals |

# Top government agentic AI use cases

| | | | |
|---|---|---|---|
| PROBLEM | Complex engineering design cycles create lengthy development timelines in regulated environments | Remote autonomous systems require extensive human oversight and control<br>. | Regulated operations face inefficiencies in compliance documentation and analysis |
| USE CASE | **AI-powered engineering design for mission-critical components** | **Edge-optimized autonomous systems with real-time decision making** | **ITAR-compliant AI platform for regulated operations** |
| OUTCOME | Reduce development cycles while maintaining security and compliance<br>Increase first-time quality rates through AI-assisted design validation<br>Enable secure knowledge sharing across regulated organizations | Enable efficient autonomous operations with reduced ground control requirements<br>Decrease time-to-deployment for critical systems<br>Reduce operational risk through predictive analytics | Achieve faster compliance analysis and reductions in procedure documentation time<br>Improve audit readiness through automated compliance tracking<br>Enable rapid response to regulatory changes through automated updates |

# Action to take now

- Work backwards from what your business NEEDs.
- Real world next steps
- What data do you need access to

# Thank you!


**Shashank Tanksali**

Sr Solutions Architect
GenAI + Security

**Norman Owens**

Sr Solutions Architect
GenAI + Security


**Please complete the survey for this session**

**Track : Artificial Intelligence and Machine Learning**

**Session : Gen AI Master Class**

**3:15pm – 4:15pm**

**300** level

**Generative AI Masterclass**

A comprehensive masterclass on AI, covering technology evolution, implementation strategies, responsible practices.

aws Learning Days
State, Local, and Education