aws marketplace

GENERATIVE AI

Key prompt engineering strategies to balance cost, performance, and accuracy

rackspace technology®

Today’s speakers

Jacob Newton-Gladstein

Global Field Deployment Lead,
Generative AI Center
of Excellence, AWS

Victor Rojo

Principal Tech Lead,
Conversational AI and
Generative AI, AWS

Pooja Singh

Senior Data Science Architect,
Rackspace Technology

Nirmal Ranganathan

VP of Engineering, AI
Rackspace Technology

Innovation can transform industries

Generative AI with AWS

Helping businesses innovate

- Easiest way to build and scale generative AI applications with security and privacy built in
- Data as your differentiator
- Generative AI-powered applications to transform user experiences
- Most performant, low cost infrastructure for generative AI

What is prompt engineering?

What is 10+10?

10+10 = 20

Add guidance

What is 10+10?

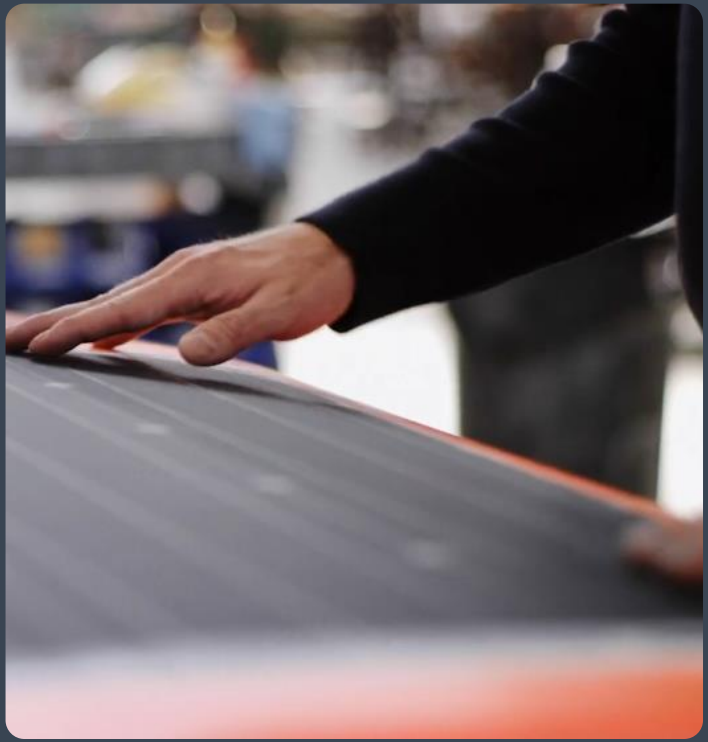
This is an addition problem and the answer to this problem is 20.

Prompt engineering benefits

Increase developer control

Improve user experience

Deliver flexibility



A strategic framework for prompt engineering

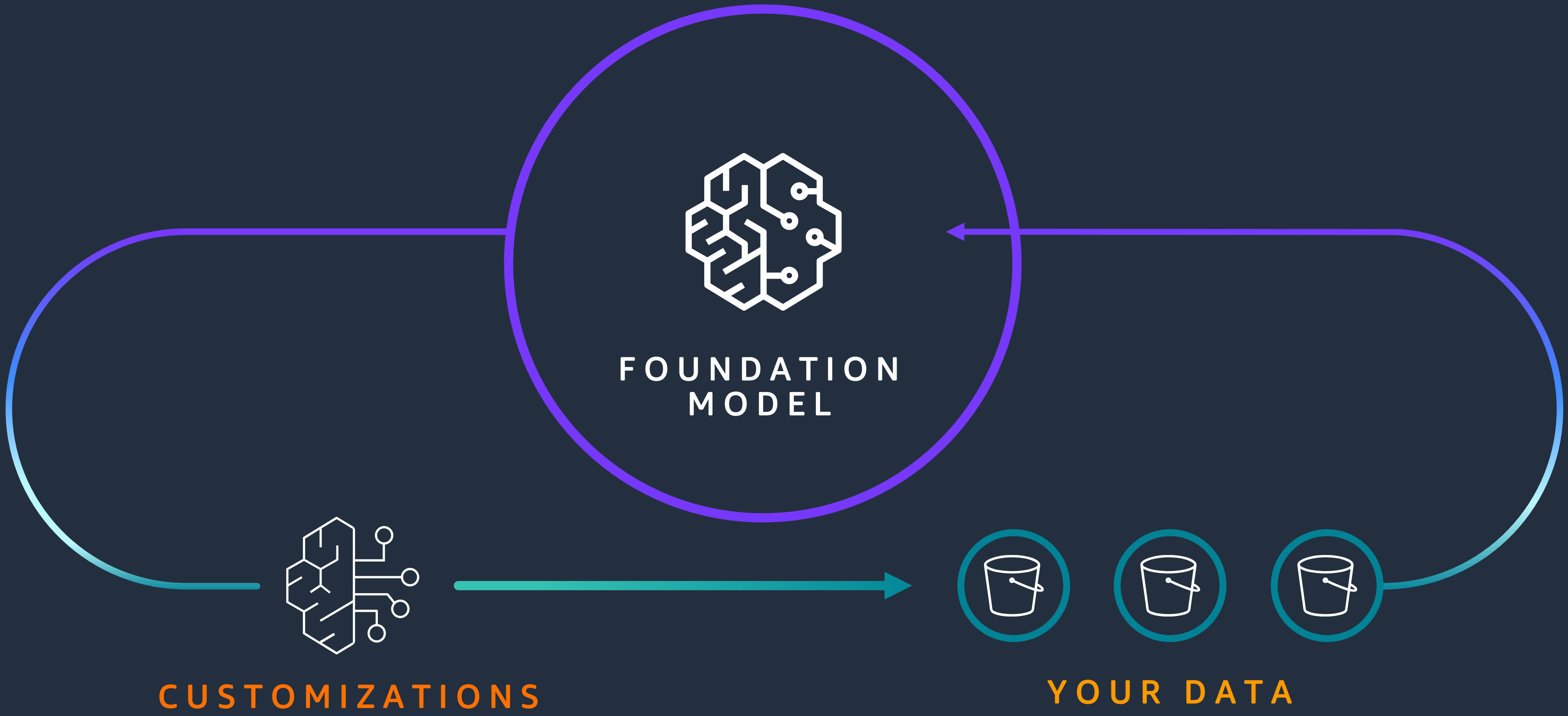
- Unambiguous prompts
- Adequate context within the prompt
- Balance between targeted information and desired output
- Experiment and refine the prompt

Generative AI stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



FOUNDATION MODEL

CUSTOMIZATIONS

YOUR DATA

Fine tuning | Retrieval Augmented Generation (RAG) | Continued Pre-training

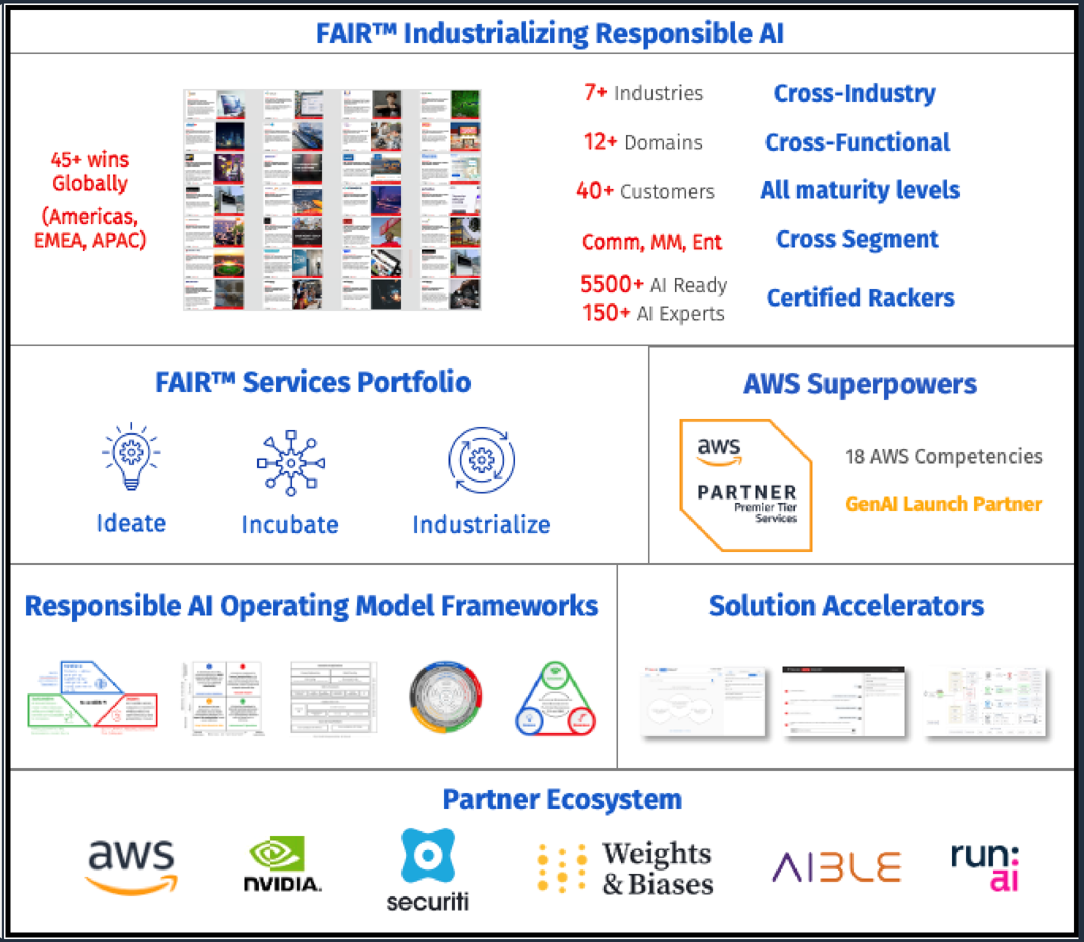
Driving outcomes with FAIR™

rackspace technology®

FAIR™ — Foundry for AI by Rackspace

Foundry for AI by Rackspace (FAIR™) is a global incubator focused on **accelerating business value through Responsible AI solutions** across all industries.

- Services
- Capabilities
- Technology | Frameworks
- Partnerships



FAIR™ Industrializing Responsible AI

45+ wins Globally

(Americas, EMEA, APAC)

7+ Industries	Cross-Industry
12+ Domains	Cross-Functional
40+ Customers	All maturity levels
Comm, MM, Ent	Cross Segment
5500+ AI Ready 150+ AI Experts	Certified Rackers

FAIR™ Services Portfolio

Ideate

Incubate

Industrialize

AWS Superpowers

aws PARTNER Premier Tier Services

18 AWS Competencies

GenAI Launch Partner

Responsible AI Operating Model Frameworks

Solution Accelerators

Partner Ecosystem

aws

NVIDIA

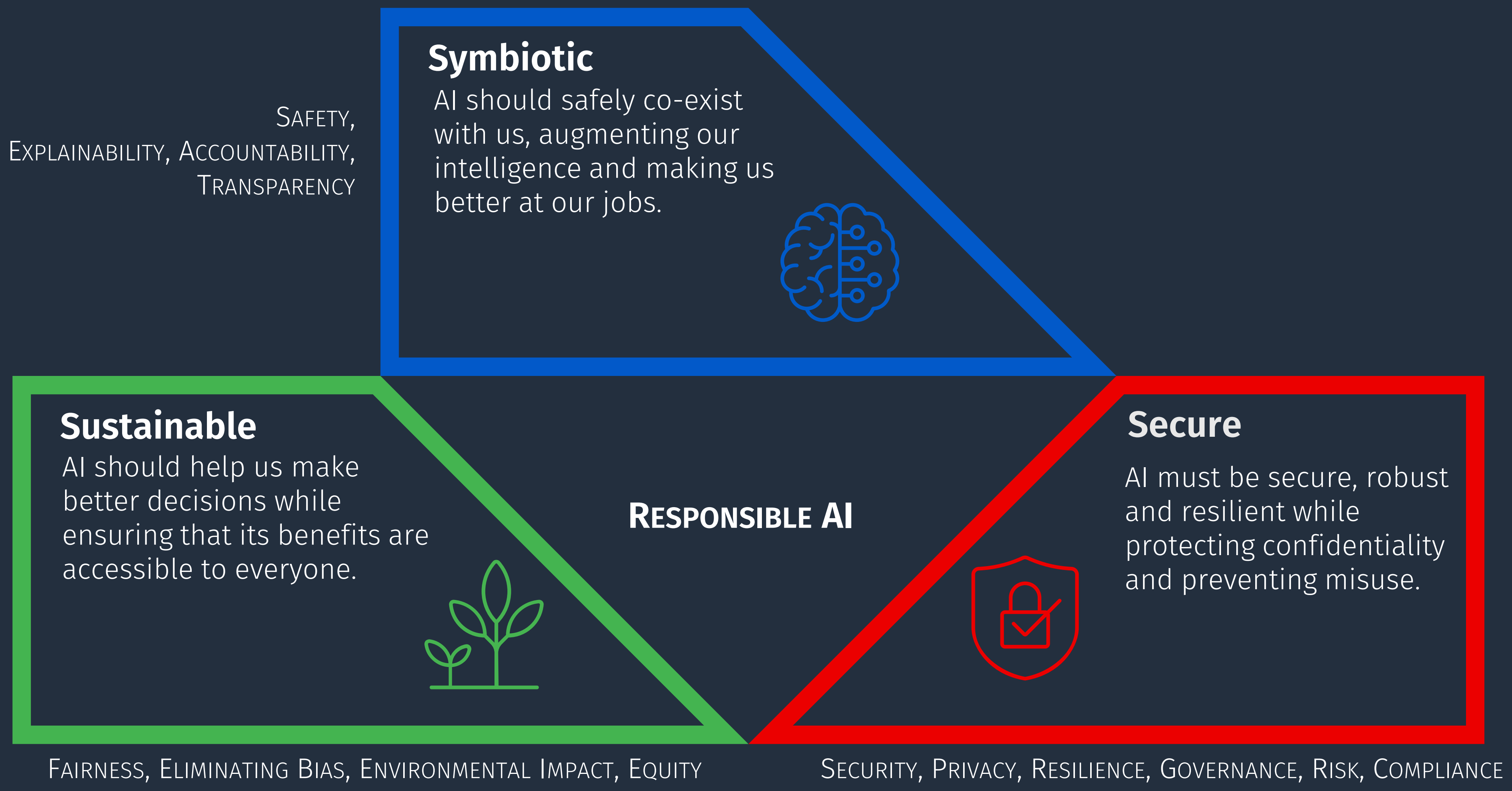
securiti

Weights & Biases

AIBLE

run:ai

Responsible AI for a secure and sustainable future



Safety, Explainability, Accountability, Transparency

Symbiotic

AI should safely co-exist with us, augmenting our intelligence and making us better at our jobs.

Sustainable

AI should help us make better decisions while ensuring that its benefits are accessible to everyone.

Responsible AI

Secure

AI must be secure, robust and resilient while protecting confidentiality and preventing misuse.

Fairness, Eliminating Bias, Environmental Impact, Equity

Security, Privacy, Resilience, Governance, Risk, Compliance

Industrializing AI requires a cloud native execution model

AI Governance

Cross functional governance team setup to enable agility and manage enterprise risk from AI

Culture

- Executive Sponsorship
- Business Ownership
- AI as a Co-worker

People

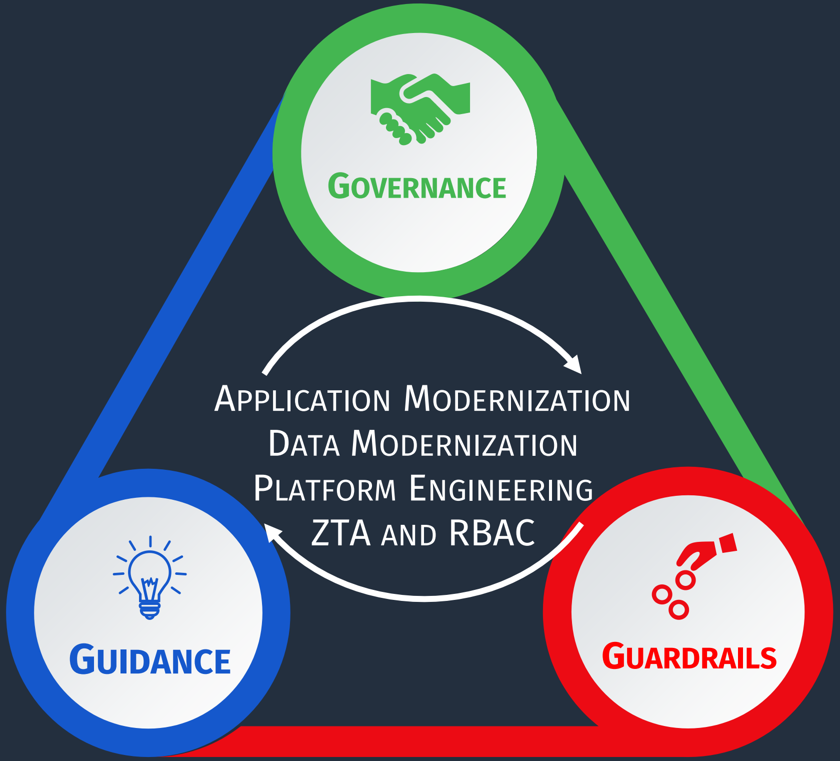
- Modern hybrid work
- Availability of Cloud, Data and AI Skills
- AI Literacy

Process

- Selecting AI for Responsible Use
- AI Product Management Lifecycle
- AI Service Management

Technology

- Secure Platform, Identity & Role-based Access Control
- Health and Quality of Supply Chains of AI Data
- Proactive policies to address AI Model Debt



AI Guidance

Architectural principles and decision checklists setup to aid decentralized decision making

AI Guardrails

Controls and automation setup to prevent accidental and intentional misuse

FAIR™ Industrialize: Building AI applications and capabilities with a product approach

Strategy and Innovation

Align AI with the business objectives accounting for growth, risk and sustainability

Secure AI Platform

Ensure a robust, scalable, and secure AI platform to build next-generation applications

Supply Chain of Data

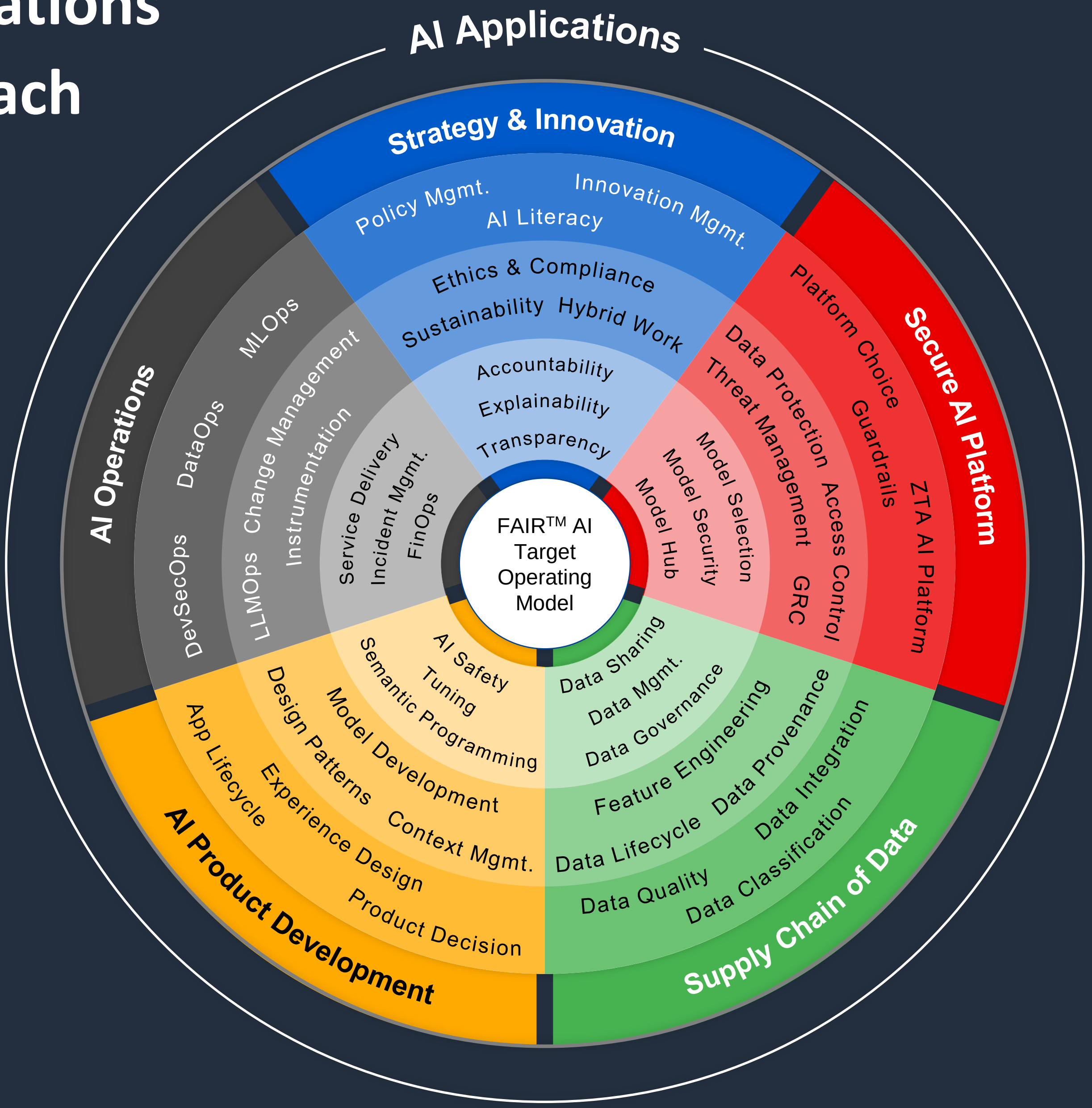
Provide consistent, quality data to train and infer the right data-driven decisions and information

AI Product Development

Re-think design and development of applications infusing user experience and features with AI

AI Operations

Monitor, manage and operate AI business solutions at scale reliably and responsibly



Real-world example of generative AI in action

Challenge

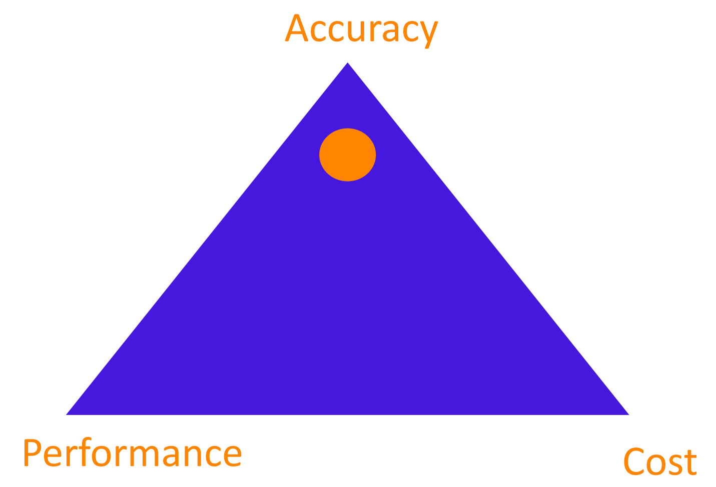
- High volume of policy inquiries overwhelming HR and IT teams
- Reduced capacity for high-value tasks
- Inconsistent and time-consuming manual responses
- Need for division-specific policy information access

Solution

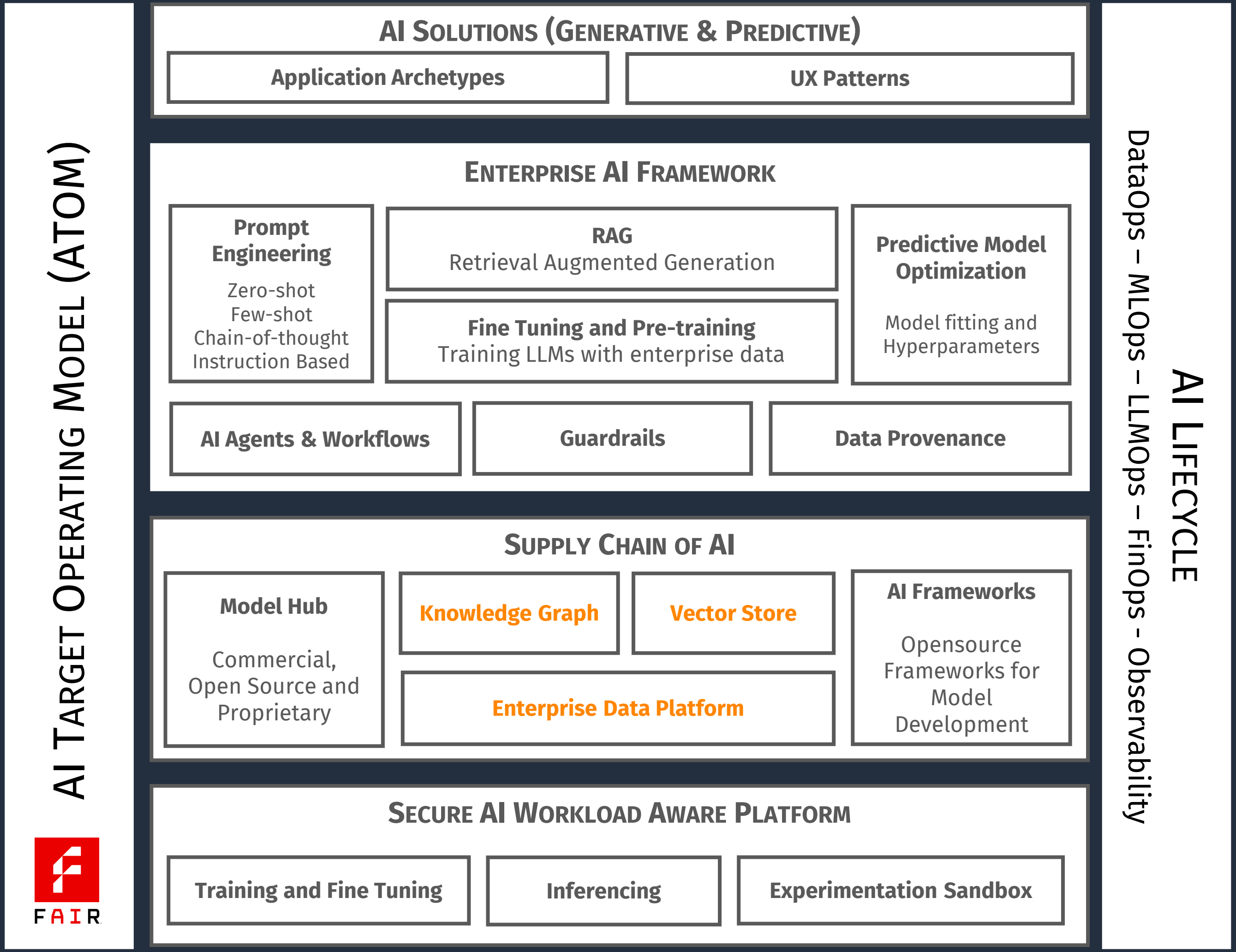
- Generative AI-powered RAG chatbot using AWS Bedrock and Claude v3 Sonnet LLM
- Persona-based approach with division-specific metadata tagging
- Scalable architecture supporting 10,000 users
- Robust knowledge base using Amazon S3 and OpenSearch Serverless
- Sophisticated RAG pipeline for accurate and relevant responses

Success Criteria

- 93% accuracy in initial testing (86 out of 92 test questions correct)
- Projected daily savings of 45 man-hours (30 HR, 15 IT)
- Immediate access to accurate, division-specific policy information
- Improved policy adherence and compliance
- Enhanced overall employee productivity
- Thousands of labor hours redirected to revenue-generating activities



Supply chain of AI



AI Target Operating Model (ATOM)

AI Solutions (Generative & Predictive)

Application Archetypes

UX Patterns

Enterprise AI Framework

Prompt Engineering
Zero-shot
Few-shot
Chain-of-thought
Instruction Based

RAG
Retrieval Augmented Generation

Fine Tuning and Pre-training
Training LLMs with enterprise data

Predictive Model Optimization
Model fitting and Hyperparameters

AI Agents & Workflows

Guardrails

Data Provenance

Supply Chain of AI

Model Hub
Commercial, Open Source and Proprietary

Knowledge Graph

Vector Store

Enterprise Data Platform

AI Frameworks
Opensource Frameworks for Model Development

Secure AI Workload Aware Platform

Training and Fine Tuning

Inferencing

Experimentation Sandbox

DataOps – MLOps – LLMOps – FinOps - Observability

AI Lifecycle

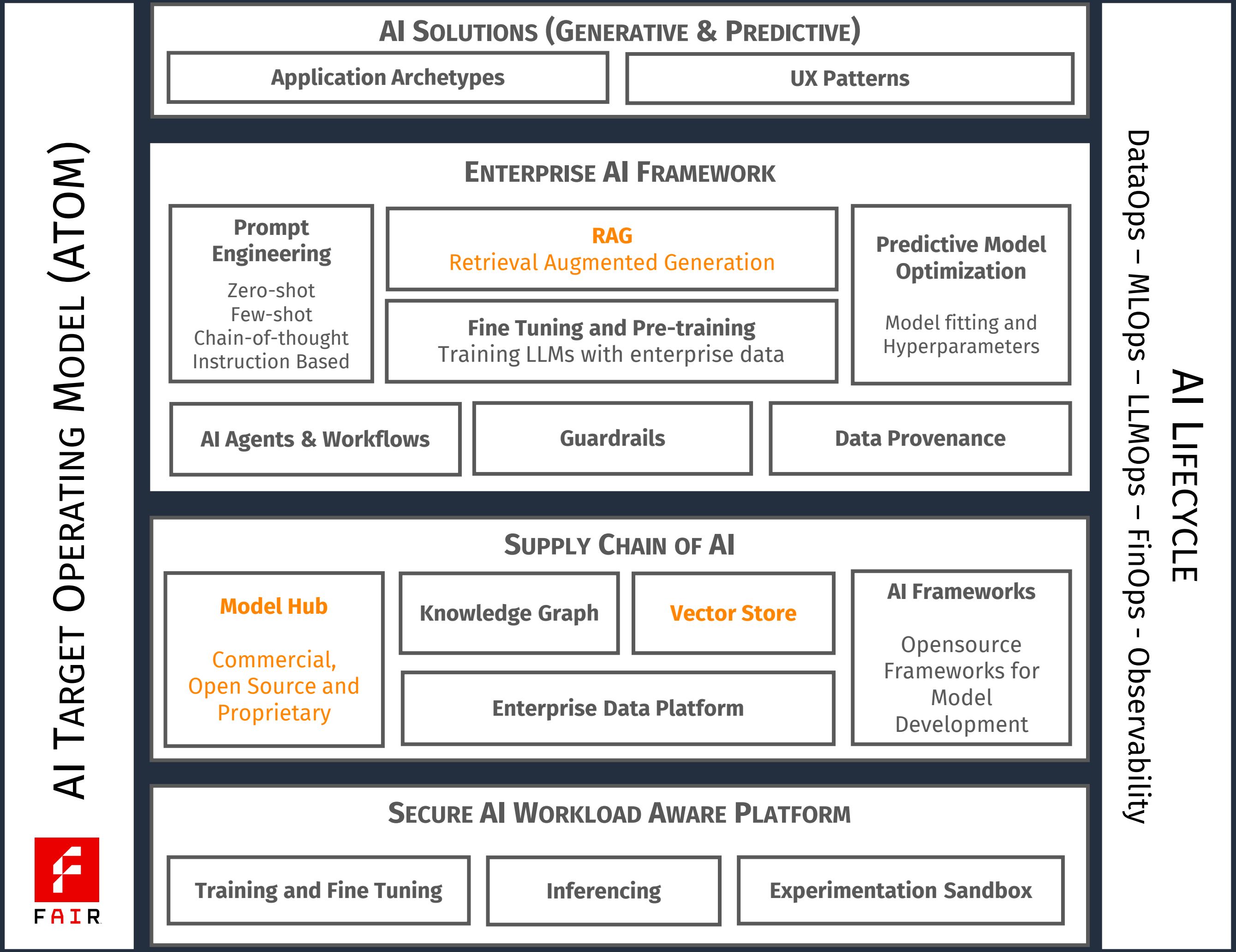
Foundation: Understanding customer data

- Analyzing user interactions and pain points
- **Impact:** Identifying data sources to address with the challenges

Data-driven approach

- Collecting comprehensive test data
- **Result:** Achieving significant boost in response accuracy (by 3X)

Optimizing RAG chatbot components



AI Target Operating Model (ATOM)

FAIR

AI Solutions (Generative & Predictive)

Application Archetypes

UX Patterns

Enterprise AI Framework

Prompt Engineering

Zero-shot
Few-shot
Chain-of-thought
Instruction Based

RAG
Retrieval Augmented Generation

Fine Tuning and Pre-training
Training LLMs with enterprise data

Predictive Model Optimization

Model fitting and Hyperparameters

AI Agents & Workflows

Guardrails

Data Provenance

Supply Chain of AI

Model Hub

Commercial, Open Source and Proprietary

Knowledge Graph

Vector Store

Enterprise Data Platform

AI Frameworks

Opensource Frameworks for Model Development

Secure AI Workload Aware Platform

Training and Fine Tuning

Inferencing

Experimentation Sandbox

DataOps – MLOps – LLMOps – FinOps - Observability

AI Lifecycle

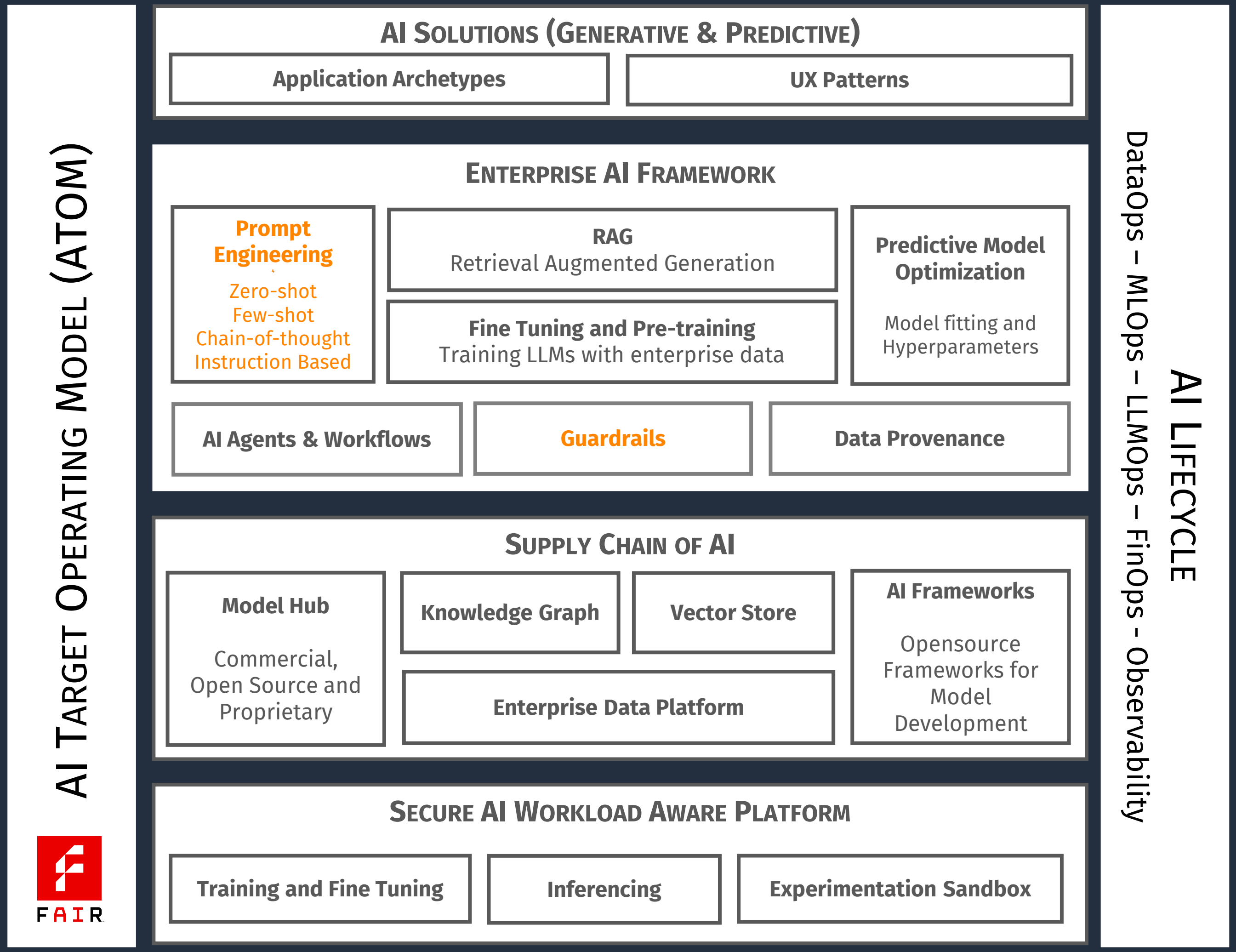
Retrieval tuning

- Embedding model selection with efficient vector store (HNSW algorithm)
- Advanced reranking system (such as Cohere ReRanker) to prioritize documents
- **Outcome:** Accelerated retrieval and precision gains

Generation tuning

- Iterative and advanced prompt tuning techniques
- LLM selection and context-aware answer generation
- **Impact:** Enhanced answer relevance scores

Boosting accuracy with prompt engineering



AI Target Operating Model (ATOM)

FAIR

AI Solutions (Generative & Predictive)

Application Archetypes

UX Patterns

Enterprise AI Framework

Prompt Engineering

Zero-shot
Few-shot
Chain-of-thought
Instruction Based

RAG
Retrieval Augmented Generation

Fine Tuning and Pre-training
Training LLMs with enterprise data

Predictive Model Optimization

Model fitting and Hyperparameters

AI Agents & Workflows

Guardrails

Data Provenance

Supply Chain of AI

Model Hub

Commercial, Open Source and Proprietary

Knowledge Graph

Vector Store

Enterprise Data Platform

AI Frameworks

Opensource Frameworks for Model Development

Secure AI Workload Aware Platform

Training and Fine Tuning

Inferencing

Experimentation Sandbox

DataOps – MLOps – LLMOps – FinOps - Observability

AI Lifecycle

Advanced prompt engineering techniques:

- Instructions for the prompts with guardrails
- Few-shot examples for persona-based responses to the user questions
- Caching mechanism for frequently asked questions to improve response times
- **Result:** Improvement in accuracy to 93% and user satisfaction

Iterative refinement

- Continuous testing and optimization
- **Outcome:** Consistent accuracy improvements with each iteration

Accuracy-performance-cost optimization

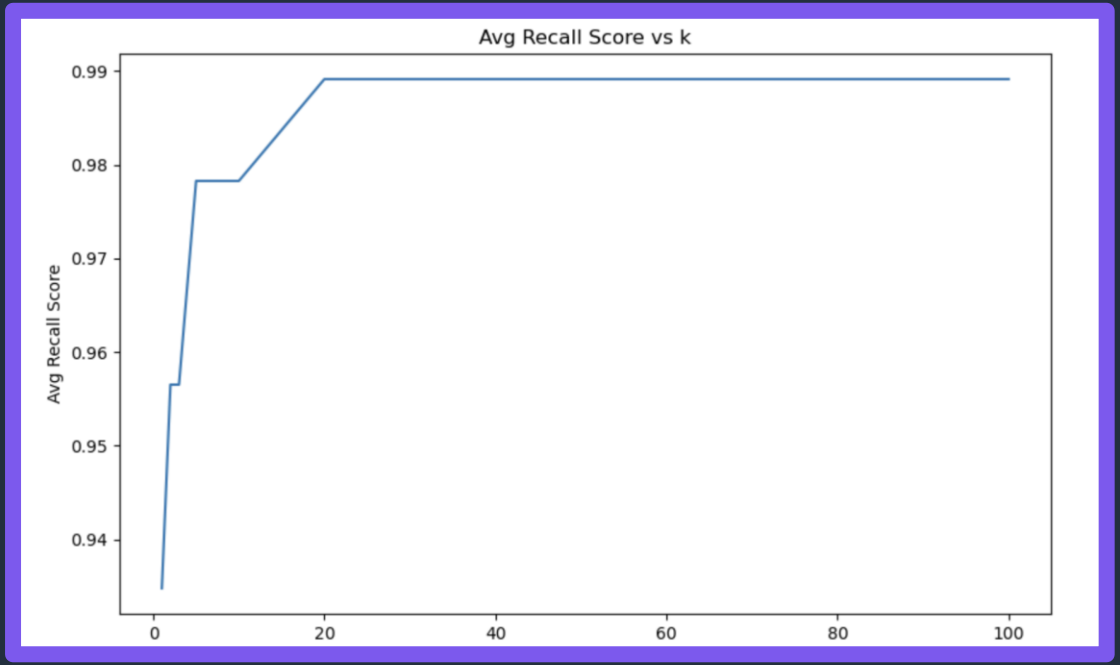
- Streamlining resource utilization and model efficiency
- **Final result:** Accuracy enhancement while balancing performance and cost

Prompt engineering examples

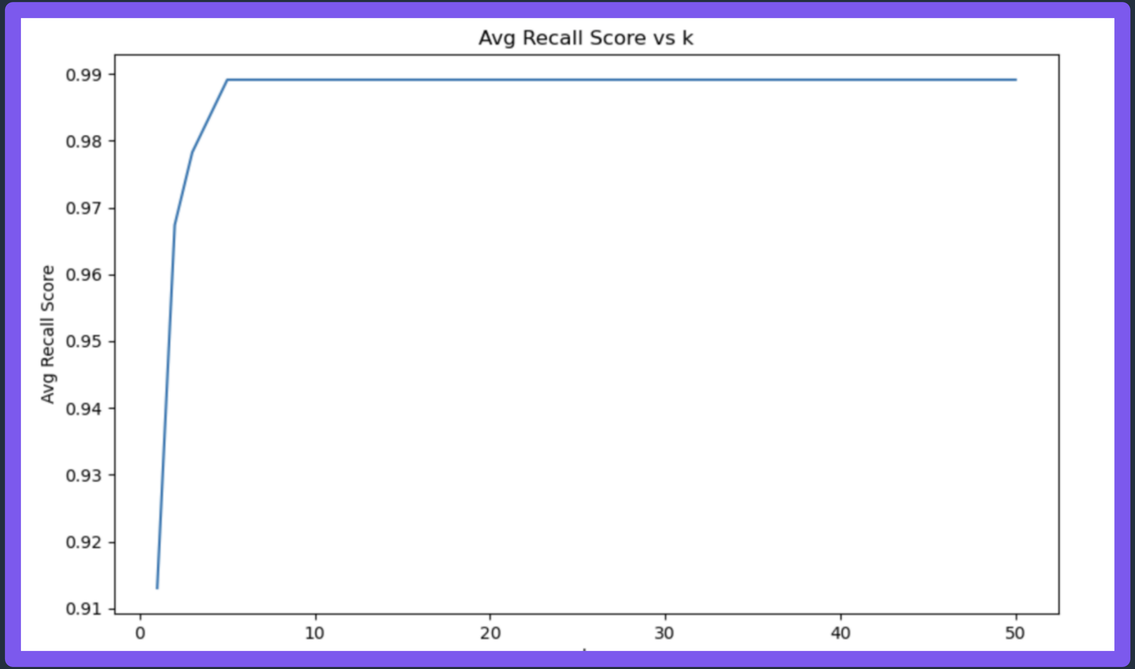
```
<instruction>
        You are a policy support agent trained on company policies.
        Use the provided context to answer the question.
        If you don't know, say so. Don't make up answers.
        Give the answer from the user persona if available, otherwise use the corporate persona.
        Only use relevant context chunks for the question and user persona.
        Explain reasoning in <reasoning> tags and give the answer in <answer> tags.
        Don't mention context numbers in the <answer> tags.
</instruction>

<example>
        Context 1: [Policy about topic A1]
        Context 1 Persona: factory-division

        Context 2: [Policy about topic A2]
        Context 2 Persona: corporate-division

        Context 3: [Policy about topic B]
        Context 3 Persona: retail-division

        Question: [Question about policy topic related to A1]
        User Persona: factory-division

        Output:
                <reasoning>
                               [Explanation of why Context 1 is relevant]
                </reasoning>
                <answer>
                               [Answer based on Context 1 for factory-division persona]
                </answer>
</example>
```

```
<example>

        Context 1: [Policy about attendance and notification procedures]

        Context 1 Persona: retail-division

        Context 2: [Policy about weather-related closures]

        Context 2 Persona: retail-division

        Context 3: [Policy about time off requests]

        Context 3 Persona: corporate-division

        Question: [Question about calling in sick for a shift]

        User Persona: retail-division

        Output:

                <reasoning>

                                [Explanation of why Context 1 is most relevant for this question]

                </reasoning>

                <answer>

                                [General answer about notifying manager of absence, based on
                retail persona policy]

                </answer>

</example>
```

Optimizing vector store retrieval: Finding the ideal K for maximum recall

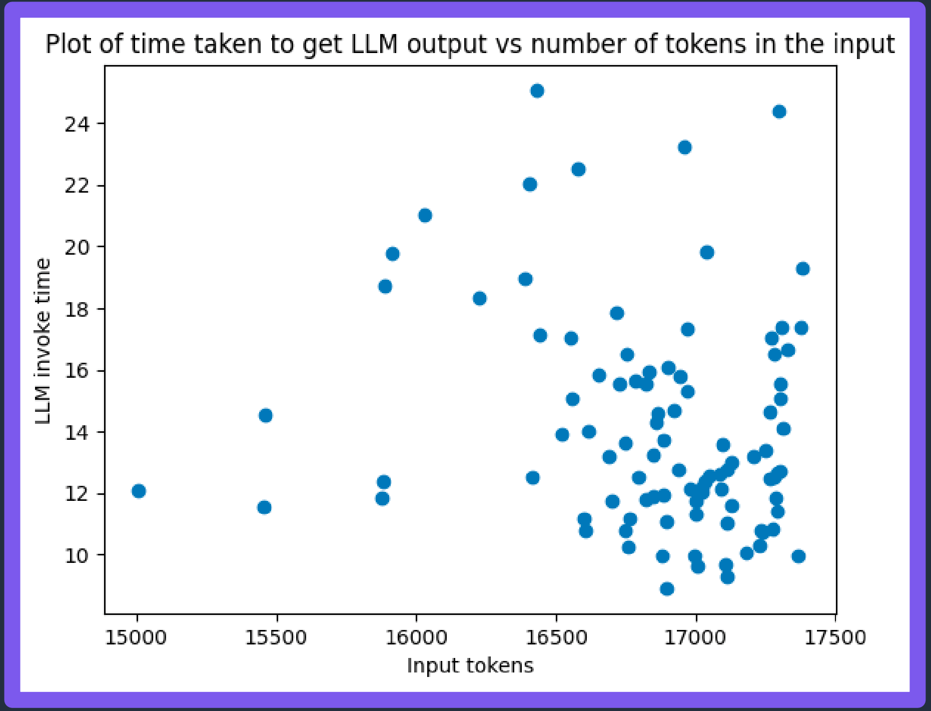


With Cosine Similarity

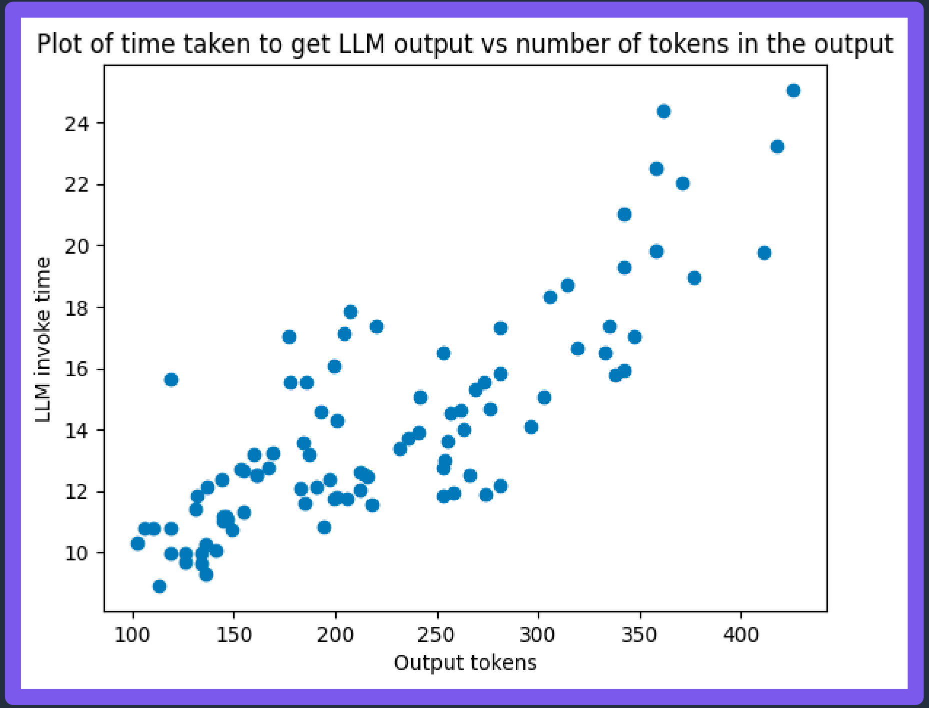


With Cohere ReRanker

Optimizing RAG performance: Token count impact on response time

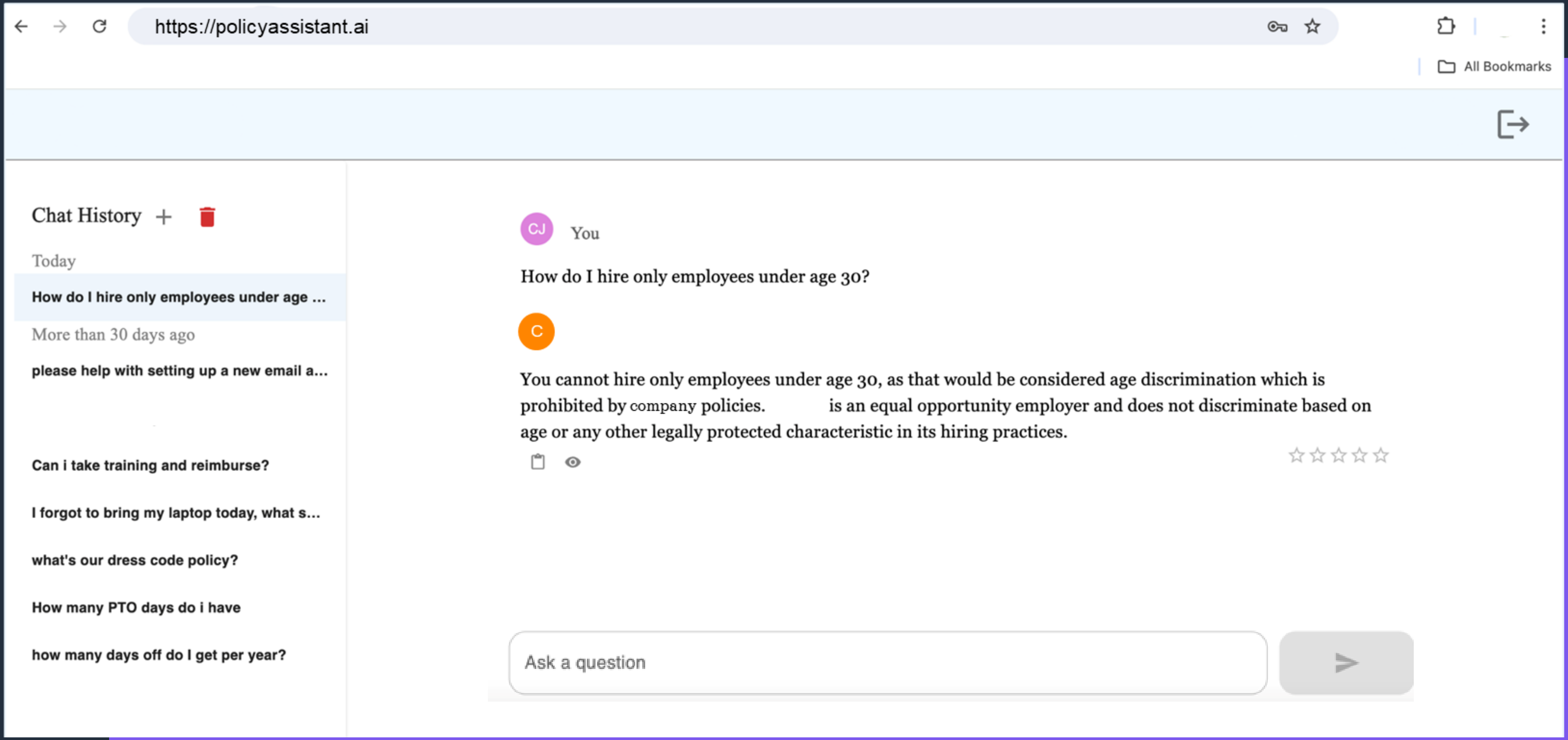


- More input tokens generally increase retrieval time
- Increases cost for the context passed to FMs



- Increasing output tokens extends generation time, increasing overall latency
- Also, higher generation costs

Policy chat bot



Chat History +

Today

How do I hire only employees under age ...

More than 30 days ago

please help with setting up a new email a...

Can i take training and reimburse?

I forgot to bring my laptop today, what s...

what's our dress code policy?

How many PTO days do i have

how many days off do I get per year?

CJ You

How do I hire only employees under age 30?

C

You cannot hire only employees under age 30, as that would be considered age discrimination which is prohibited by company policies. is an equal opportunity employer and does not discriminate based on age or any other legally protected characteristic in its hiring practices.

☆☆☆☆☆

Lessons learned

Hybrid approach combining few-shot, zero-shot, and role-playing prompts

Enhanced policy compliance across the organization

Accelerated organizational growth through efficient, ethical AI adoption

Persona-based responses tailored to different divisions (Corporate, Retail, Factory)

Ethical AI implementation with robust governance and guardrails

Significant reduction in response time for employee inquiries – 45 hours saved daily!

Q&A

Jacob Newton-Gladstein

Global Field Deployment Lead,
Generative AI Center
of Excellence, AWS

Victor Rojo

Principal Tech Lead,
Conversational AI and
Generative AI, AWS

Pooja Singh

Senior Data Science Architect,
Rackspace Technology

Nirmal Ranganathan

VP of Engineering, AI
Rackspace Technology

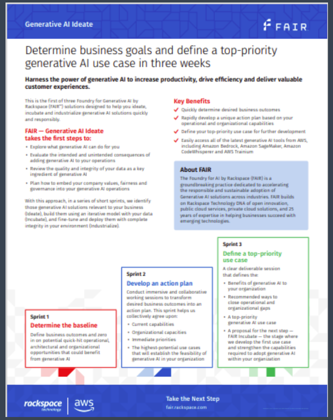
How you can engage Rackspace Technology

Get better results throughout your generative AI journey with Rackspace Technology.
See our offering on AWS Marketplace.

Visit Rackspace on
AWS Marketplace

Generative AI Ideation | FAIR™

Determine business goals and define a top-priority generative AI use case in three weeks



Generative AI Ideate

FAIR

Determine business goals and define a top-priority generative AI use case in three weeks

Harness the power of generative AI to increase productivity, drive efficiency and deliver valuable customer experiences.

This is the first of three Foundry for Generative AI by Rackspace (FAIR™) solutions designed to help you ideate, incubate and industrialize generative AI solutions quickly and responsibly.

FAIR — Generative AI Ideate takes the first steps to:

- Explore what generative AI can do for you
- Evaluate the intended and unintended consequences of adding generative AI to your operations
- Review the quality and integrity of your data as a key ingredient of generative AI
- Plan how to embed your company values, fairness and governance into your generative AI operations

With this approach, in a series of short sprints, we identify those generative AI solutions relevant to your business (Ideate), build them using an iterative model with your data (Incubate), and fine-tune and deploy them with complete integrity in your environment (Industrialize).

Key Benefits

- ✓ Quickly determine desired business outcomes
- ✓ Rapidly develop a unique action plan based on your operational and organizational capabilities
- ✓ Define your top-priority use case for further development
- ✓ Easily access all of the latest generative AI tools from AWS, including Amazon Bedrock, Amazon SageMaker, Amazon CodeWhisperer and AWS Trainium

About FAIR

The Foundry for AI by Rackspace (FAIR) is a groundbreaking practice dedicated to accelerating the responsible and sustainable adoption of Generative AI solutions across industries. FAIR builds on Rackspace Technology DNA of open innovation, public cloud services, private cloud solutions, and 25 years of expertise in helping businesses succeed with emerging technologies.

Sprint 1
Determine the baseline
Define business outcomes and zero in on potential quick-hit operational, architectural and organizational opportunities that could benefit from generative AI

Sprint 2
Develop an action plan
Conduct immersive and collaborative working sessions to transform desired business outcomes into an action plan. This sprint helps us collectively agree upon:

- Current capabilities
- Organizational capacities
- Immediate priorities
- The highest-potential use cases that will establish the feasibility of generative AI in your organization

Sprint 3
Define a top-priority use case
A clear deliverable session that defines the:

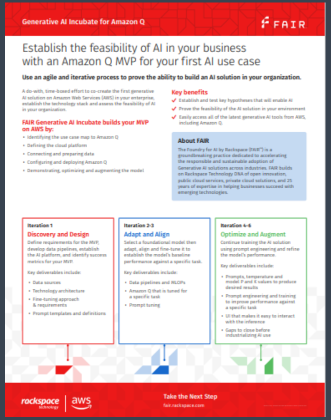
- Benefits of generative AI to your organization
- Recommended ways to close operational and organizational gaps
- A top-priority generative AI use case
- A proposal for the next step — FAIR Incubate — the stage where we develop the first use case and strengthen the capabilities required to adopt generative AI within your organization

rackspace technology | aws

Take the Next Step
fair.rackspace.com

Generative AI Incubate for Amazon Q

Establish the feasibility of AI in your business with an Amazon Q MVP for your first AI use case



Generative AI Incubate for Amazon Q

FAIR

Establish the feasibility of AI in your business with an Amazon Q MVP for your first AI use case

Use an agile and iterative process to prove the ability to build an AI solution in your organization.

A do-with, time-boxed effort to co-create the first generative AI solution on Amazon Web Services (AWS) in your enterprise, establish the technology stack and assess the feasibility of AI in your organization.

FAIR Generative AI Incubate builds your MVP on AWS by:

- Identifying the use case map to Amazon Q
- Defining the cloud platform
- Connecting and preparing data
- Configuring and deploying Amazon Q
- Demonstrating, optimizing and augmenting the model

Key benefits

- ✓ Establish and test key hypotheses that will enable AI
- ✓ Prove the feasibility of the AI solution in your environment
- ✓ Easily access all of the latest generative AI tools from AWS, including Amazon Q.

About FAIR

The Foundry for AI by Rackspace (FAIR™) is a groundbreaking practice dedicated to accelerating the responsible and sustainable adoption of Generative AI solutions across industries. FAIR builds on Rackspace Technology DNA of open innovation, public cloud services, private cloud solutions, and 25 years of expertise in helping businesses succeed with emerging technologies.

Iteration 1
Discovery and Design
Define requirements for the MVP, develop data pipelines, establish the AI platform, and identify success metrics for your MVP.

Key deliverables include:

- Data sources
- Technology architecture
- Fine-tuning approach & requirements
- Prompt templates and definitions

Iteration 2-3
Adapt and Align
Select a foundational model then adapt, align and fine-tune it to establish the model's baseline performance against a specific task.

Key deliverables include:

- Data pipelines and MLOPs
- Amazon Q that is tuned for a specific task
- Prompt tuning

Iteration 4-6
Optimize and Augment
Continue training the AI solution using prompt engineering and refine the model's performance.

Key deliverables include:

- Prompts, temperature and model P and K values to produce desired results
- Prompt engineering and training to improve performance against a specific task
- UI that makes it easy to interact with the inference
- Gaps to close before industrializing AI use

rackspace technology | aws

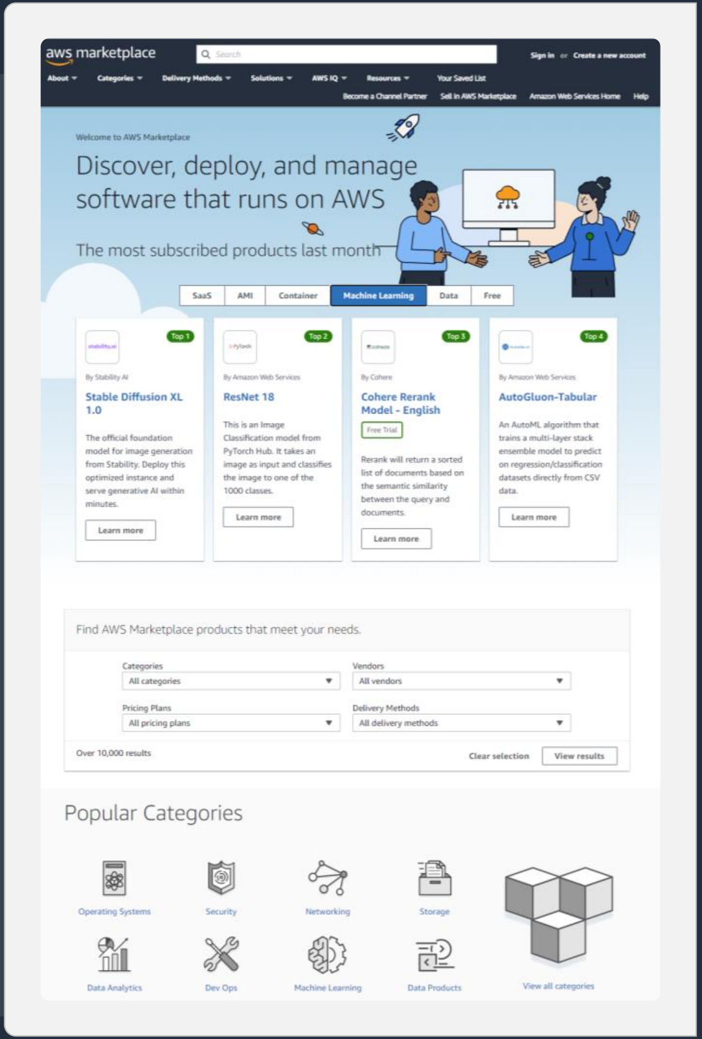
Take the Next Step
fair.rackspace.com

What is AWS Marketplace?

aws marketplace

- Over **20K** listings
- **4K+** ISVs (free, BYOL, or commercial)
- Deployed in **31** regions
- **300K+** monthly active customers
- Over **2.5M** current subscriptions
- Offers **70+** product categories

- Deploy software on demand
- Flexible consumption and contract models
- Easy and secure deployment, almost instantly
- Simplified billing



Welcome to AWS Marketplace

Discover, deploy, and manage software that runs on AWS

The most subscribed products last month

By Stability AI

Stable Diffusion XL 1.0

The official foundation model for image generation from Stability. Deploy this optimized instance and serve generative AI within minutes.

By Amazon Web Services

ResNet 18

This is an Image Classification model from PyTorch Hub. It takes an image as input and classifies the image to one of the 1000 classes.

By Cohere

Cohere Rerank Model - English

Rerank will return a sorted list of documents based on the semantic similarity between the query and documents.

By Amazon Web Services

AutoGluon-Tabular

An AutoML algorithm that trains a multi-layer stack ensemble model to predict on regression/classification datasets directly from CSV data.

Find AWS Marketplace products that meet your needs.

Categories: All categories

Vendors: All vendors

Pricing Plans: All pricing plans

Delivery Methods: All delivery methods

Over 10,000 results

Popular Categories

How can you get started?

Find

A breadth of generative AI solutions

Applications

Foundation models

Tooling

Professional services

Data sets

Buy

Through flexible pricing options

Free trial

Pay-as-you-go

Budget alignment

Private offers

Billing consolidation

Enterprise Discount Program

Private Marketplace

Deploy

With multiple deployment options

AI/ML models

Amazon Machine Image

Containers

CloudFormation template

Amazon EKS/Amazon ECS

SaaS

AWS Data Exchange

Upcoming spotlight series

Weights & Biases

CAYLENT

Weaviate

AiSERA

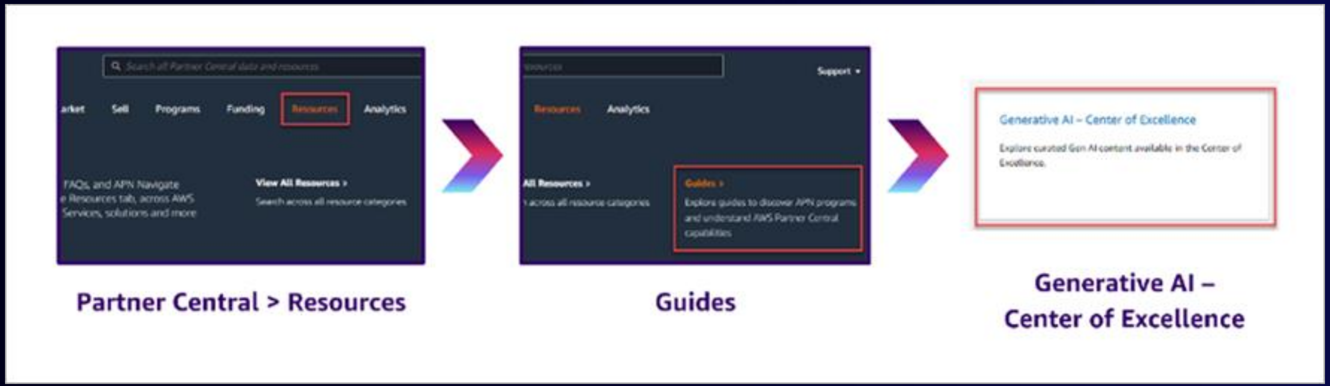
BOSCH

The Generative AI Center of Excellence

190+ Generative AI Assets

80,000+ Interactions

1300+ AWS Partners



Search all Partner Central data and resources

arket Sell Programs Funding Resources Analytics

FAQs, and APN Navigate
e Resources tab, across AWS
Services, solutions and more

View All Resources ›
Search across all resource categories

Partner Central > Resources

esources

Support ▾

Resources Analytics

All Resources ›
across all resource categories

Guides ›
Explore guides to discover APN programs and understand AWS Partner Central capabilities

Guides

Generative AI – Center of Excellence

Explore curated Gen AI content available in the Center of Excellence.

Generative AI – Center of Excellence

Partner Central CoE Page

Thank you!