Azure Sandbox Blueprint

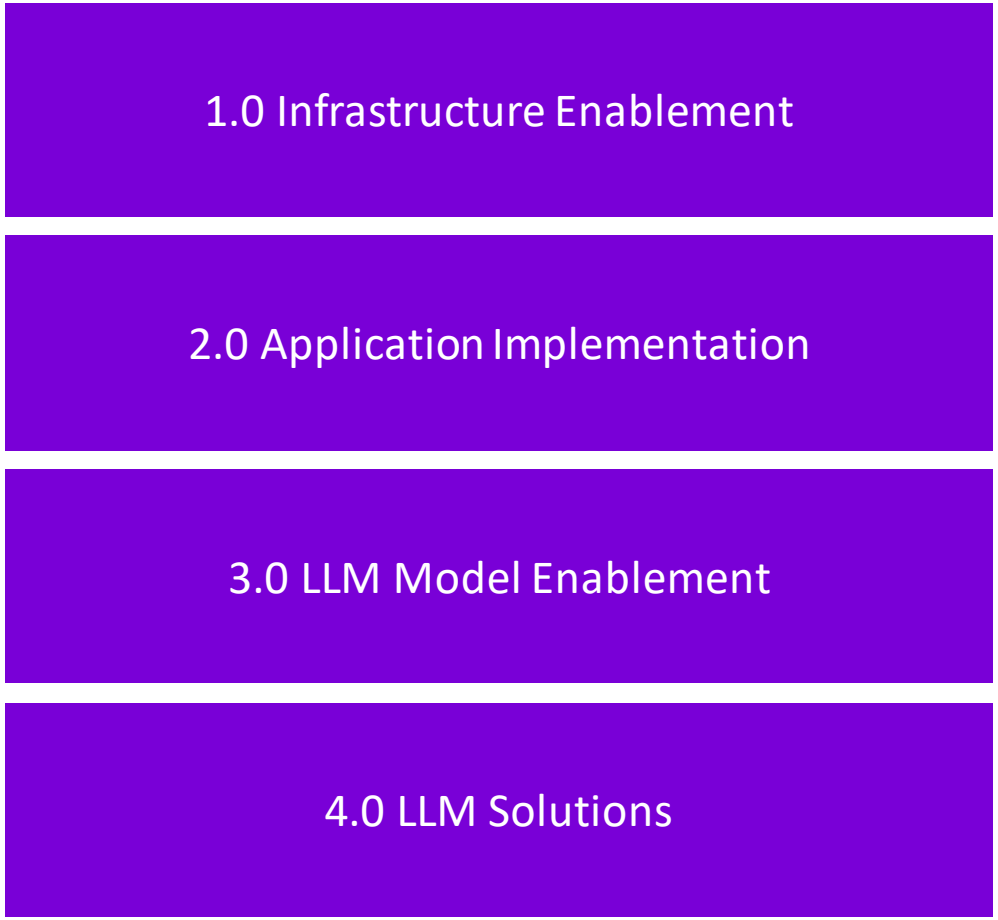
GenAI / LLM CoE

Table of Contents

Copyright © 2023 Accenture. All rights reserved.

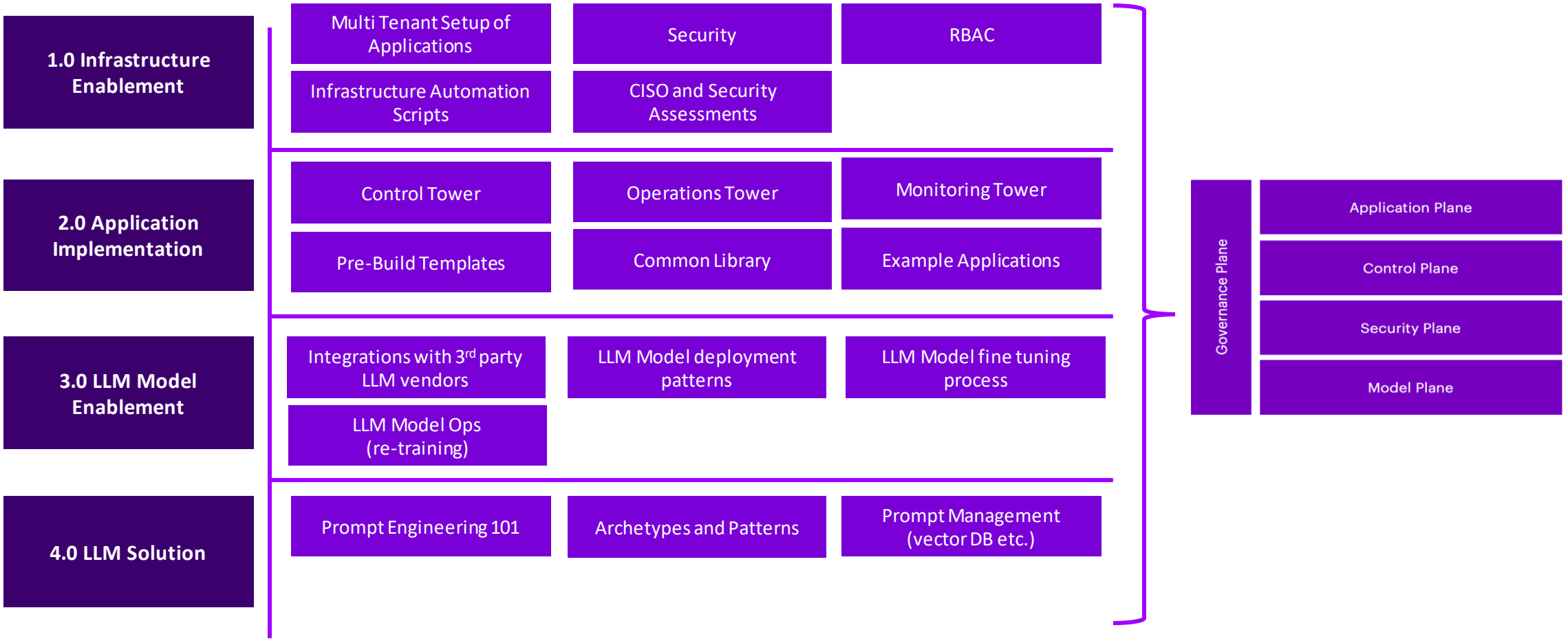
General Offerings

LLM CoE provides an **eco-system of processes, technology, security and governance** to **design, deploy and manage GAI** applications across domains. The LLM CoE helps our customers in **4 specific ways** :



Copyright © 2023 Accenture. All rights reserved.

General Offerings



Azure Sandbox provide jump start for use case implementation through general offerings in the area of Infrastructure Enablement, Application Implementation, LLM Model Enablement, LLM Solution

Copyright © 2023 Accenture. All rights reserved.

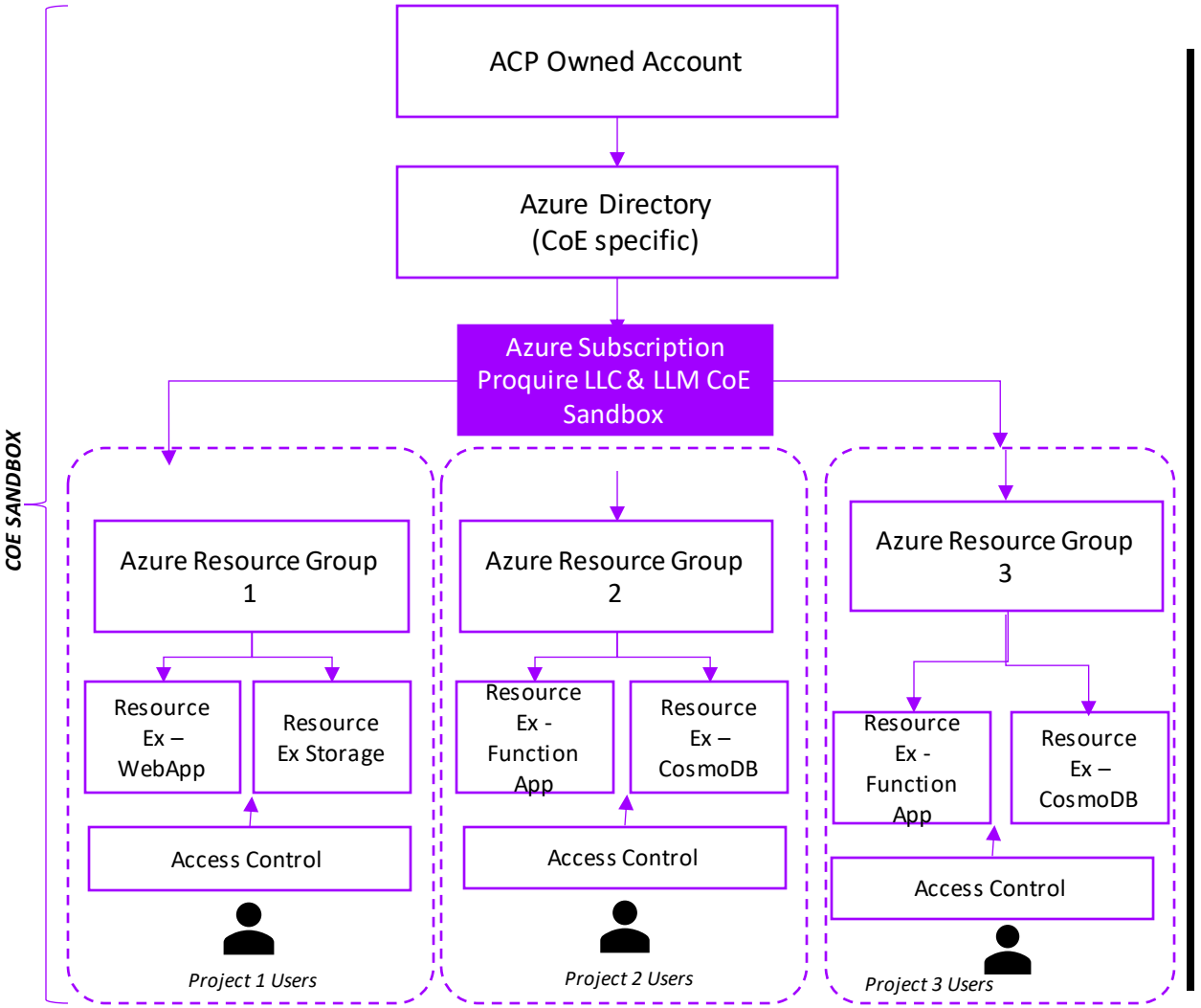
Azure Sandbox has to Offer

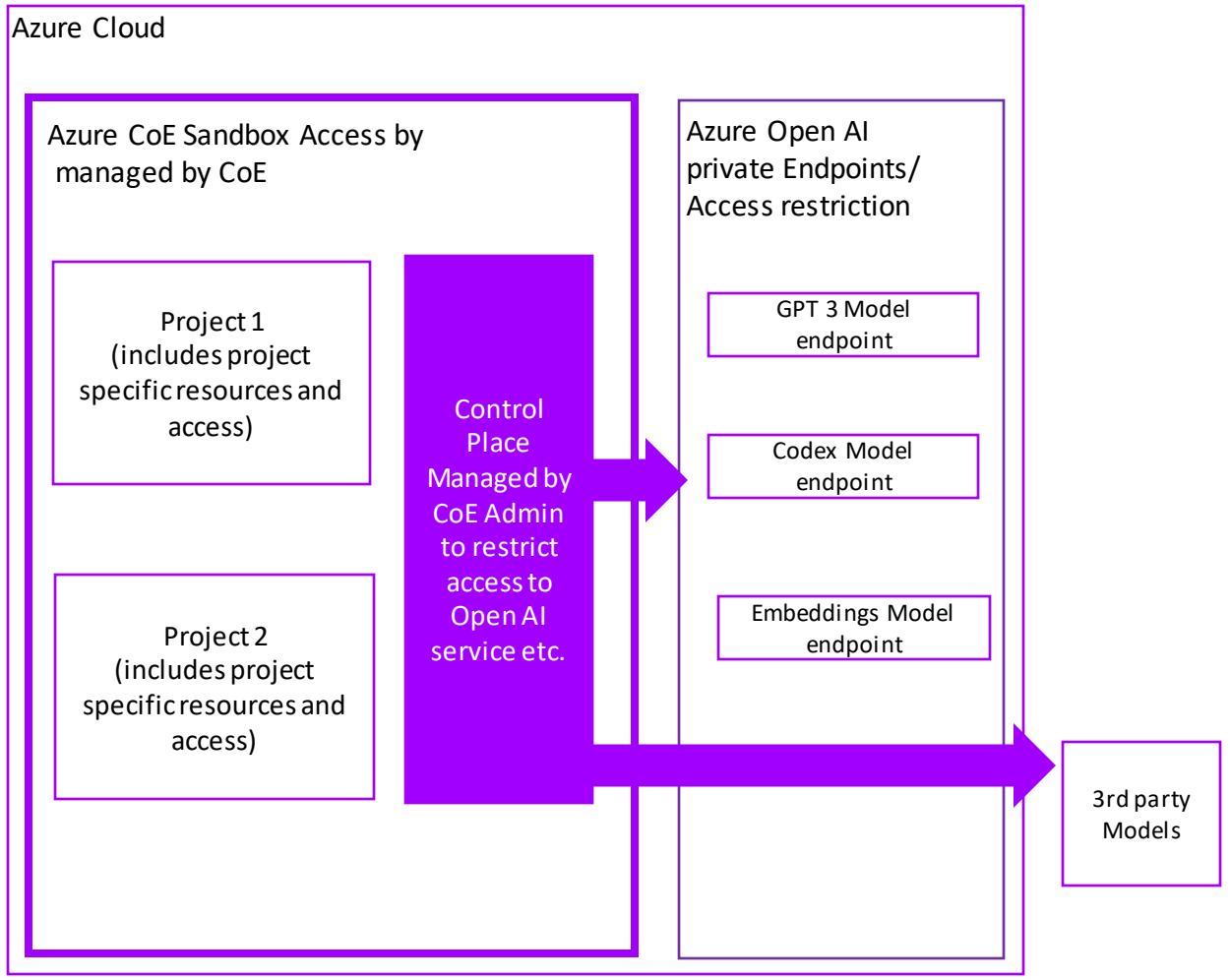
- We provide a dedicated Multi-tenant support in provisioning the environment
- Use case are onboarded based on their preference of sandbox usage being in compliant with Legal & CISO guidelines.
- All onboarding are isolated, and access controlled based on the requirement
- All azure infra services have networking and security guardrails
- Azure Sandbox access is provided via the controlled governed and managed COE wrapper services that have the security guardrails implemented. Each use case is isolated using a separate resource group and application key.

Copyright © 2023 Accenture. All rights reserved.

Azure Sandbox Resource Groups and Project Setup

Below is a simplified view of how CoE sandbox will manage resource and allocation of LLM services within the account





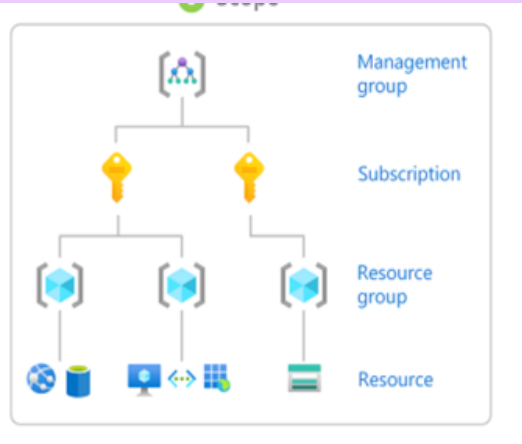
Copyright © 2023 Accenture. All rights reserved.

Access and User Controls

Each Azure Subscription is security managed & monitored by Accenture Cloud Platform team, and regularly scans and make sure the environment is secured as per Accenture Standards. Additionally Azure RBAC mechanism is followed while provisioning access to the onboarding users.

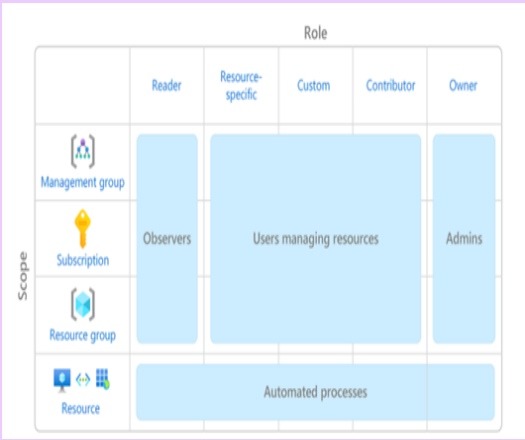
Project 1 – dedicated setup (all resource belongs to a specific subscription)

***1.Security Admin** – Accenture CMO team will have Global & Subscription Owner Permission.*



- Cloud Manager and Optimizer (CMO) uses the Prisma and Cloud Health tools to audit cloud accounts for security vulnerabilities.
- These tools scan the configuration of your discovered cloud estate against the Infosec Cloud Security Configuration Standard

***2. LLM CoE Admin team** will have Full Admin control over the Subscription*

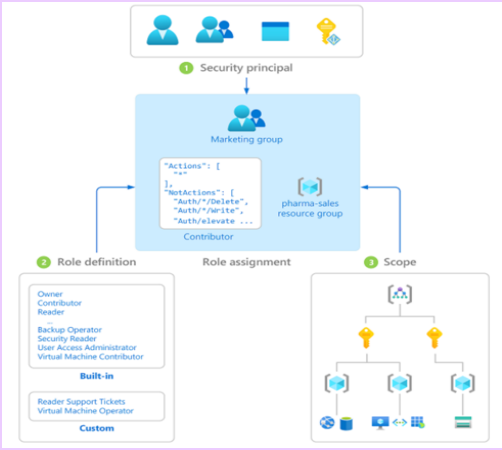


- LLM CoE Admin have full control over the access management.
- AD Group Management.
- Service Principal and App Register

***3.Project Use Case** AD Group Permission assignment*

- Project use case team will have Reader Access in RG
- Contributor access on its resources level for the respective Project RGs.
- Project team will not have any new resource creation access.

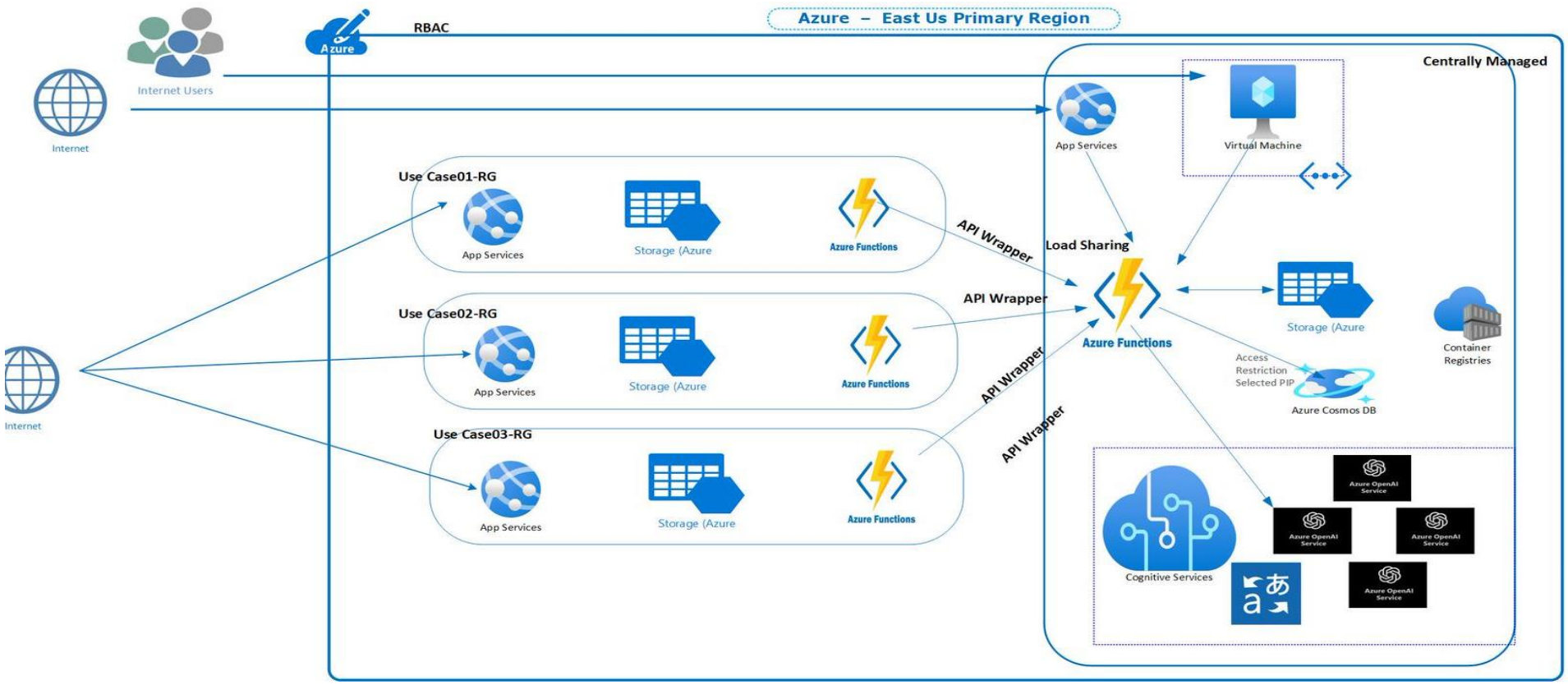
4.Service Principal – identity access provided.



- *Identity access provided for App integration and DevOps deployment*

Copyright © 2023 Accenture. All rights reserved.

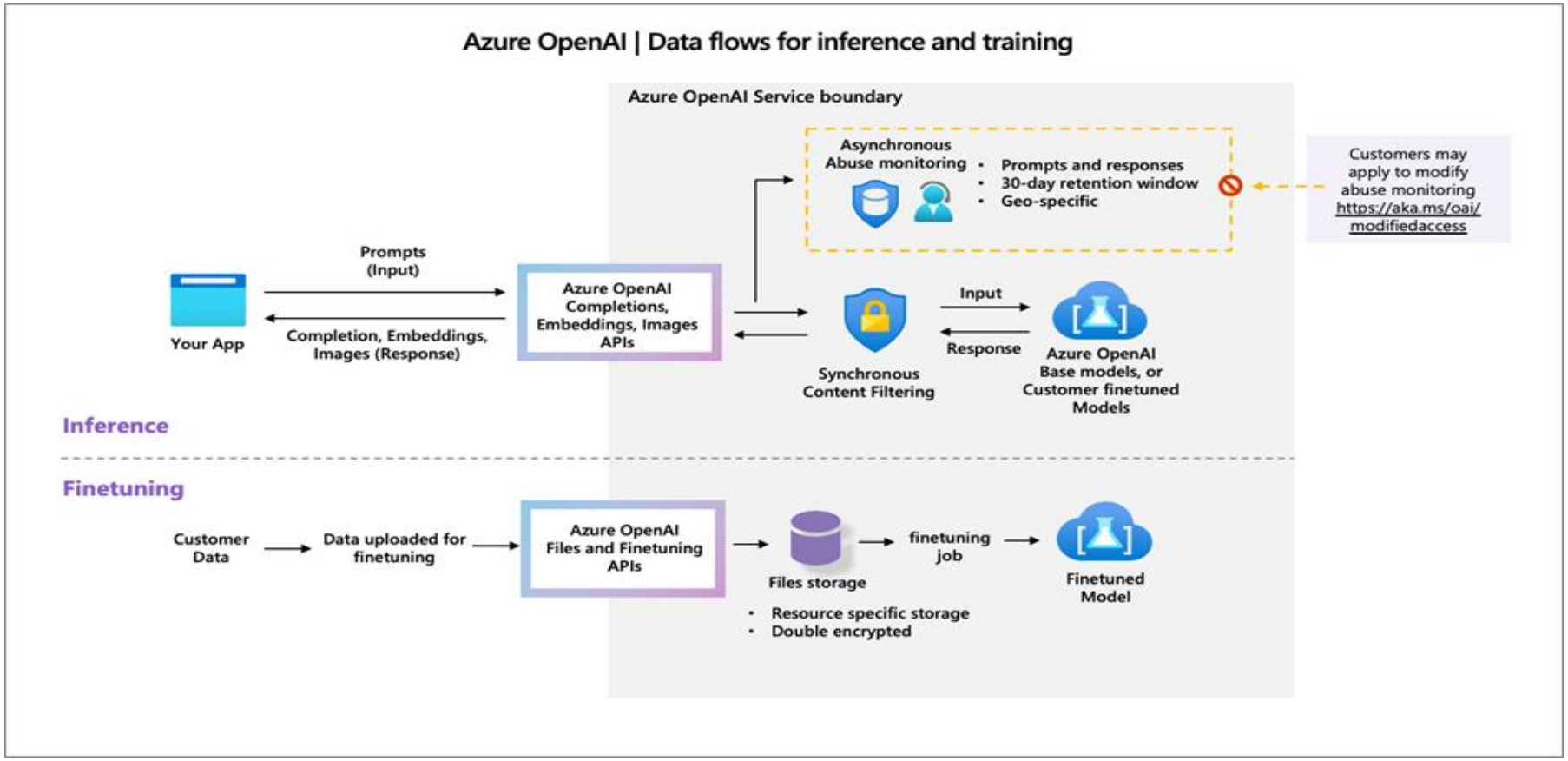
Azure Sandbox Infra Network Design



Azure Sandbox access is provided via the controlled governed and managed COE wrapper services that have the security guardrails implemented. Each use case is isolated using a separate resource group and application key.

Copyright © 2023 Accenture. All rights reserved.

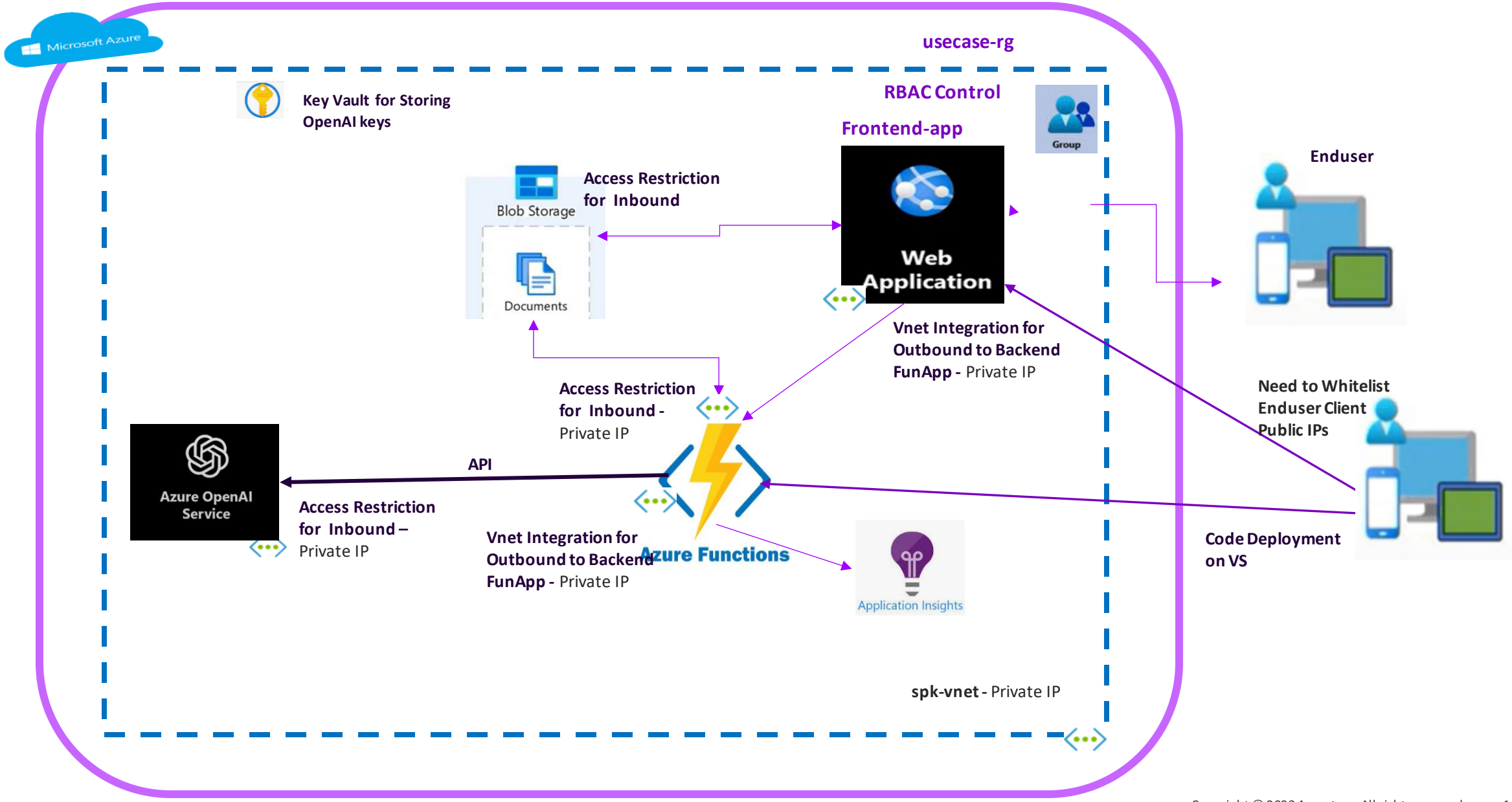
Azure Open AI Data Security



**The models are stateless: no prompts or generations are stored in the model. Additionally, prompts and generations are not used to train, retrain, or improve the base models.

Copyright © 2023 Accenture. All rights reserved.

Network Architecture (Use Case Illustrative)

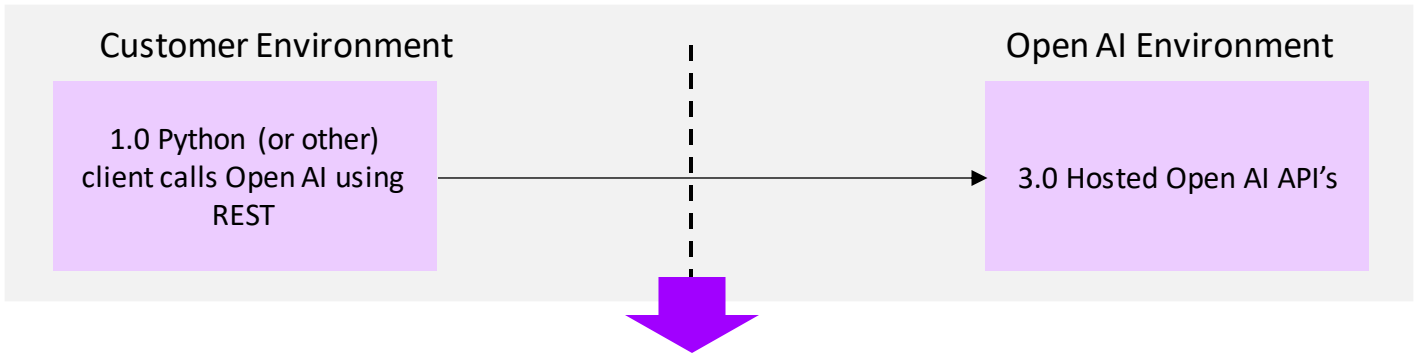


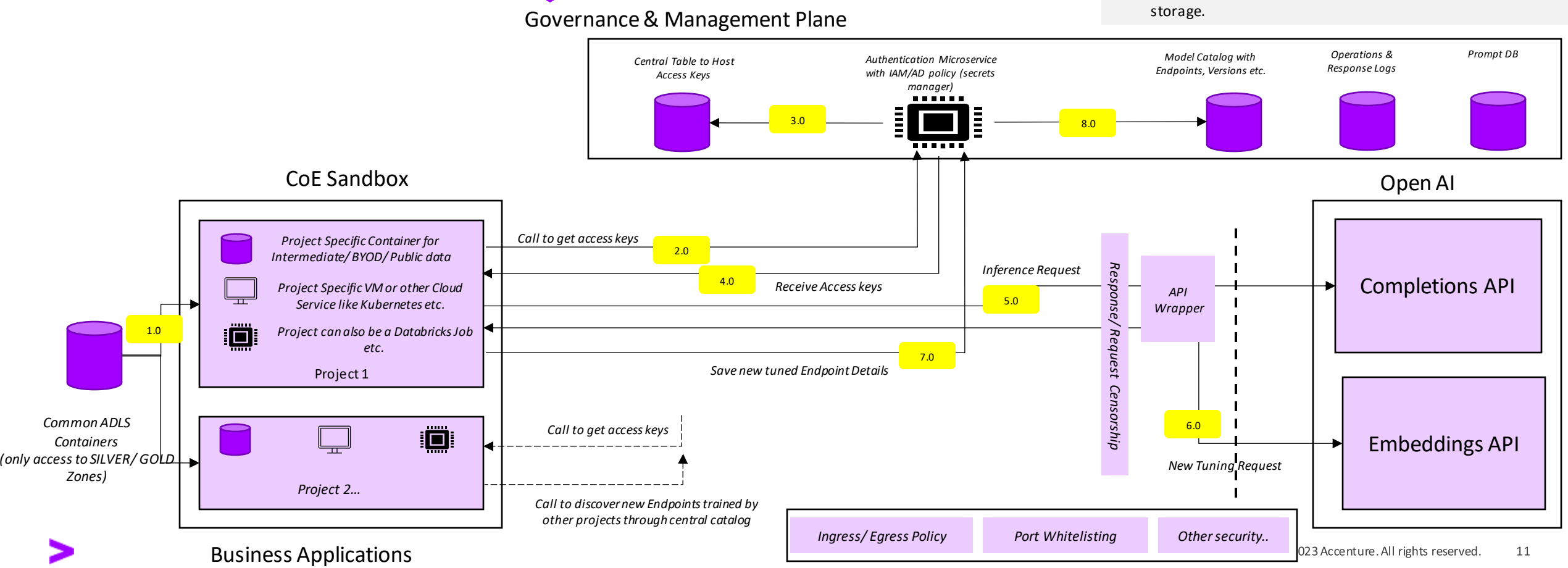
Copyright © 2023 Accenture. All rights reserved.

Azure Open AI – Model Integration Blueprint

Key Takeaways

- This setup defines hot to stand up a secure and governed Sandbox with customers own VPC
- Create a central Key Management/ access management repository
- Create a central catalog of applications, customized endpoints and versions
- Manage a microservice/ API to authenticate and share common keys
- Manage a microservice/ API to save/ discover trained endpoints
- Dedicated project with segmentation for each application with access to project specific tools/ services and dedicated storage.





023 Accenture. All rights reserved.

Data Security in Sandbox

Ask	Security Measures taken

How does information get to your sandbox?	• Data is shared by the project/client teams via
	  • MS Teams Private Channel
	  • SFTP
	  • SharePoint link
	• Azure blob storage account is created with no public access enabled.
	• It is securely uploaded to the azure blob storage for further access within the RG resources

How do you secure the information on your sandbox? What tools are you using for data monitoring, access control, etc.?	• Data is stored into azure blob storage and access is restricted via Azure RBAC controls • All operations are done via secure transfer that is enabled with TLS 1.2 • The data at rest is protected by storage service encryption using the Microsoft managed keys. • Storage blob/container level Soft delete is enabled with 7days of retention for data recovery. • Azure monitor, activity logs provides the monitoring analytics.
What users can access?	• Users are provided with the web application UI link that is SSO enabled to access. • Client users are not provided access directly to the azure resources in Accenture owned sandbox. • Client owned environment can be enabled with the same level of monitoring and access control for them to manage.
Physical infrastructure security	• All the azure resources for the usecase are Vnet integrated and public access restriction. • All the FTP access is disabled, and the connections are only via TLS1.2 protocol. • Users can access the web application UI link that is SSO enabled to access.

Copyright © 2023 Accenture. All rights reserved.

Sandbox Patterns

- Azure LLM CoE Sandbox is an end-to-end offerings for generative AI based applications that provides templates, blueprints, guidelines and guardrails for enabling LLM projects.
- The CoE Sandbox provides comprehensive capabilities across model, security, control, application and governance
- As an offering from LLM CoE sandbox capability, there are 3 ways to setup the Azure Open AI Sandbox access. We will be covering details in next few slides on
  - What are different patterns to access LLM COE Sandbox environment
  - Who is this for and what will they get as per part of sandbox onboarding
  - Responsibilities of Sandbox and Account team while implementing usecases
- Accenture Owned COE - **Shared COE** Sandbox instance(Pattern #1) - There is a shared multi-tenant Sandbox instance for several COE projects.
- Accenture Owned Project Env - **Dedicated** Sandbox instance(pattern #2) - A dedicated Sandbox instance for a team can be setup by CoE Sandbox team.
- Client Owned Environment – Sandbox owned by client but leveraging CoE expertise for defining blueprints, project template, infrastructure and platform scripts, control tower to fast track their platform build out

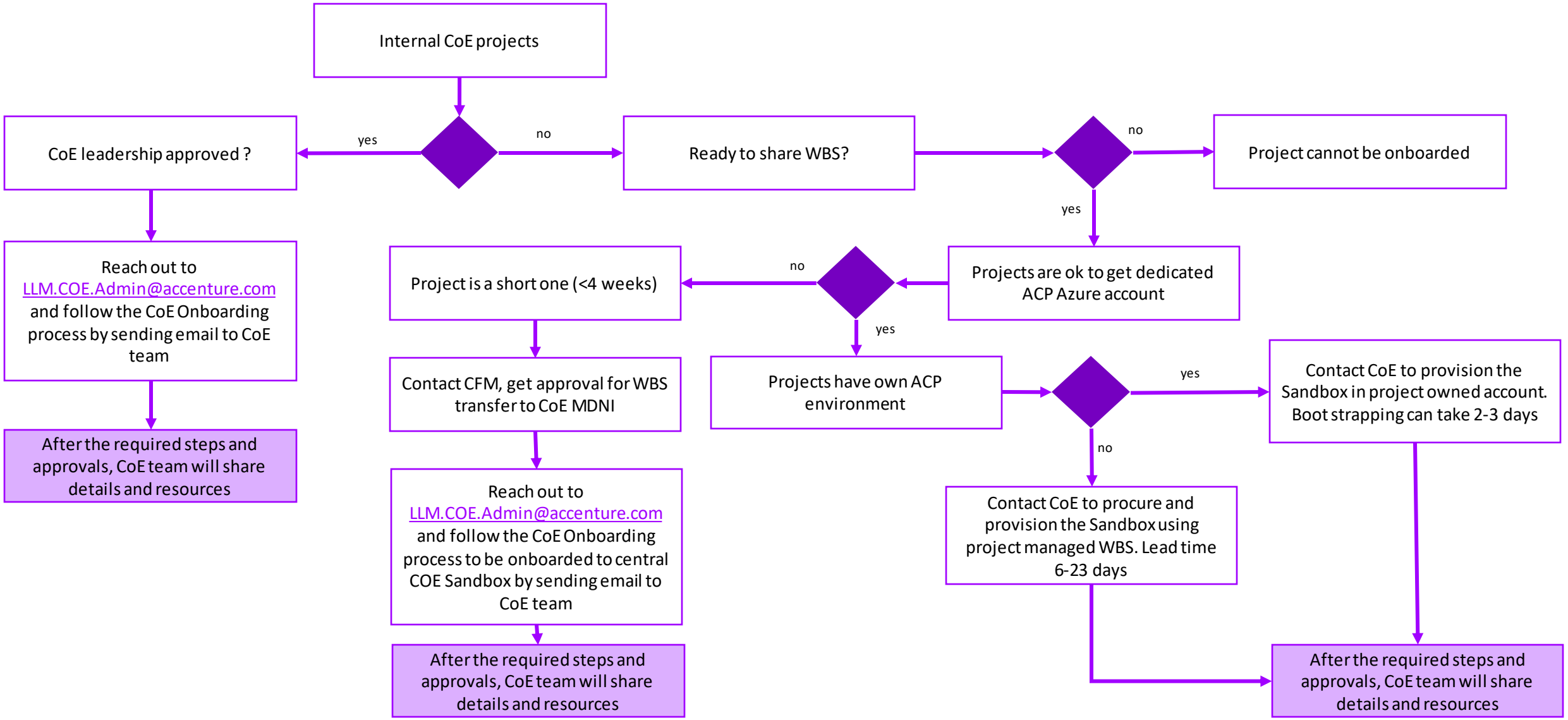
Copyright © 2023 Accenture. All rights reserved.

CoE Project Infrastructure Options

Environment Type	Shared Accenture Environment	Dedicated Accenture Environment	Client Provided Environment
Usage Scenario	Demos, offering development, Short/Med terms PoCs(Internal, Client)	Longer term Internal PoCs and Pilots Medium term client PoCs as applicable	Longer term client PoCs and Pilots
Env Status	Pre-created and shared services	Created on Demand and shared nothing	Created on Demand and shared nothing
Performance	Multi-Tenant, Throughput limitations during peak period imposed by quota	Single-Tenant, No limitations in throughout other than account quota	Single-Tenant, No limitations in throughout other than account quota
Cloud Account Owner	Accenture	Accenture	Client
Cloud Account WBS	Generative AI MDI WBS	Project provided WBS	Project provided WBS
Project Duration	Client project < 8 weeks	Client project 8+ weeks	Client projects any duration
Lead Time	1-2 days	2-4 weeks (new cloud account creation through ACP, CSP LLM access)	~4-6 weeks after client provided access to client cloud account, also depends on the client's infra provisioning process (naming conventions, n/w design/security, RBAC, readiness testing, customization).
Data Types	Public Data, Standard App Archetypes, Accenture Internal Data Client Data (CDP, Contracts approved)	Public Data Accenture Internal Data Client Data (CDP, Contracts approval)	Public Data Client Data (CDP, Contract approval)
Data Protection	Access Controlled in Cloud Data Store Project specific API Key for Clients	Access Controlled in Cloud Store Environment specific API Key	Access Controlled in Cloud Store Client specific LLM Key
Foundation Model Zero/Few Shot/Prompting	Yes	Yes	Yes
Foundation Model Customization	No	Yes	Yes
Env Customization	Limited	Full	Full
Platform Ops PoC Design/Build	CoE SME through Solution Support	CoE SME through Solution Support	CoE SME through Solution Support

Copyright © 2023 Accenture. All rights reserved.

CoE Sandbox – Onboarding Decisions Process



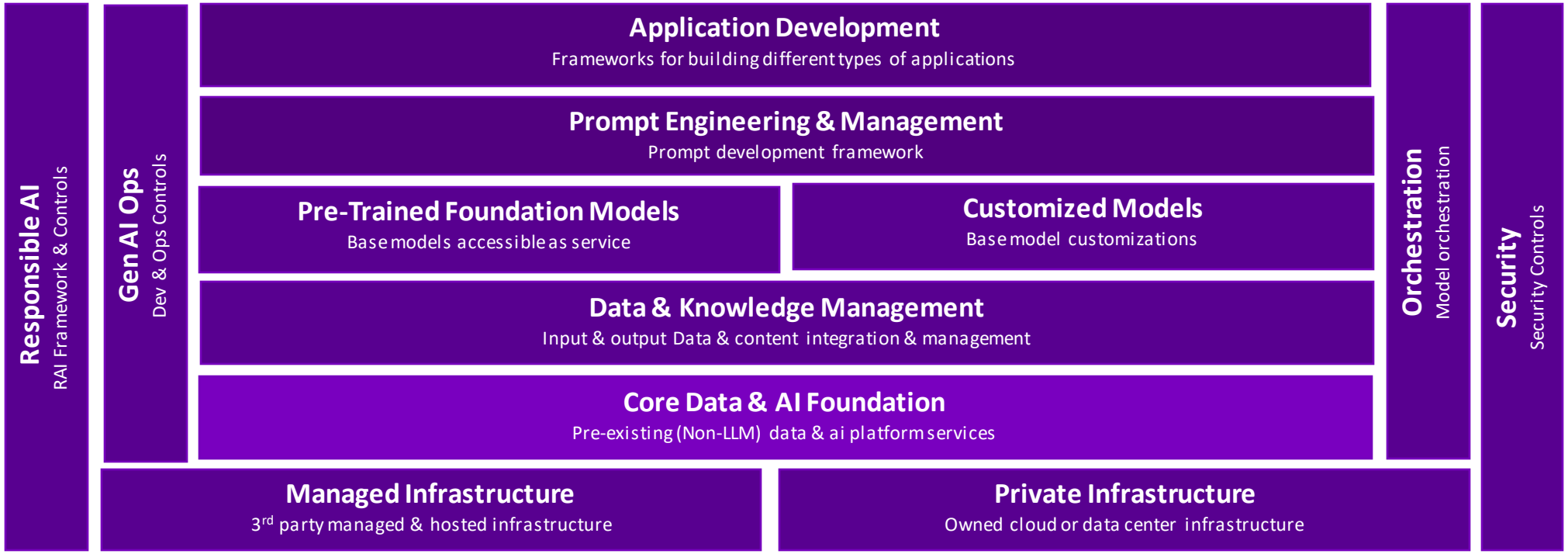
Copyright © 2023 Accenture. All rights reserved.

Azure LLM Models Available (as of today)

Azure LLM Model	Version	Used For	Costing	Rate Limitations	Quota -	Regions available	Fine Tuning***
gpt4	0314	Text/Code	Gpt4(prompts) – 0.03 USD (per 1000 tokens) Gpt4(completions) – 0.06 USD (per 1000 tokens)	18 RPM, 20K tokens per minute, 8192 max input tokens per request	3 instances per subscription	East US. France Central	N/A
gpt432k	0314	Text/Code	Gpt4(prompts) – 0.06 USD (per 1000 tokens) Gpt4(completions) – 0.12 USD (per 1000 tokens)	60K tokens per minute, 32768 max input tokens per request		East US. France Central	N/A
Dall-E		Image	2 USD per 100 images	2 Concurrent Requests, 1000 characters per request		East US	N/A
gpt-35-turbo	0301	Text	0.002 USD per 1000 tokens	240K tokens per minute, 4096 max input tokens per request		East US, France Central, South-Central US, UK South, West Europe	N/A
text-davinci-003		Text	0.02 USD per 1000 tokens	120K tokens per minute, 4097 max input tokens per request		East US, West Europe	N/A
text-embedding-ada-002	2	Text Embeddings	0.0001 per 1000 tokens	240K tokens per minute, 8191 max input tokens per request		East US, South Central US	N/A
text-embedding-ada-002	1	Text Similarity	0.0001 per 1000 tokens	240K tokens per minute, 2046 max input tokens per request		East US, South Central US, West Europe	N/A

GEN AI PLATFORM: LEVEL 1 REFERENCE ARCHITECTURE

Below is a quick view of the building blocks needed for implementing LLM powered solutions. This reference architecture divides the enablement into 12 components as below:

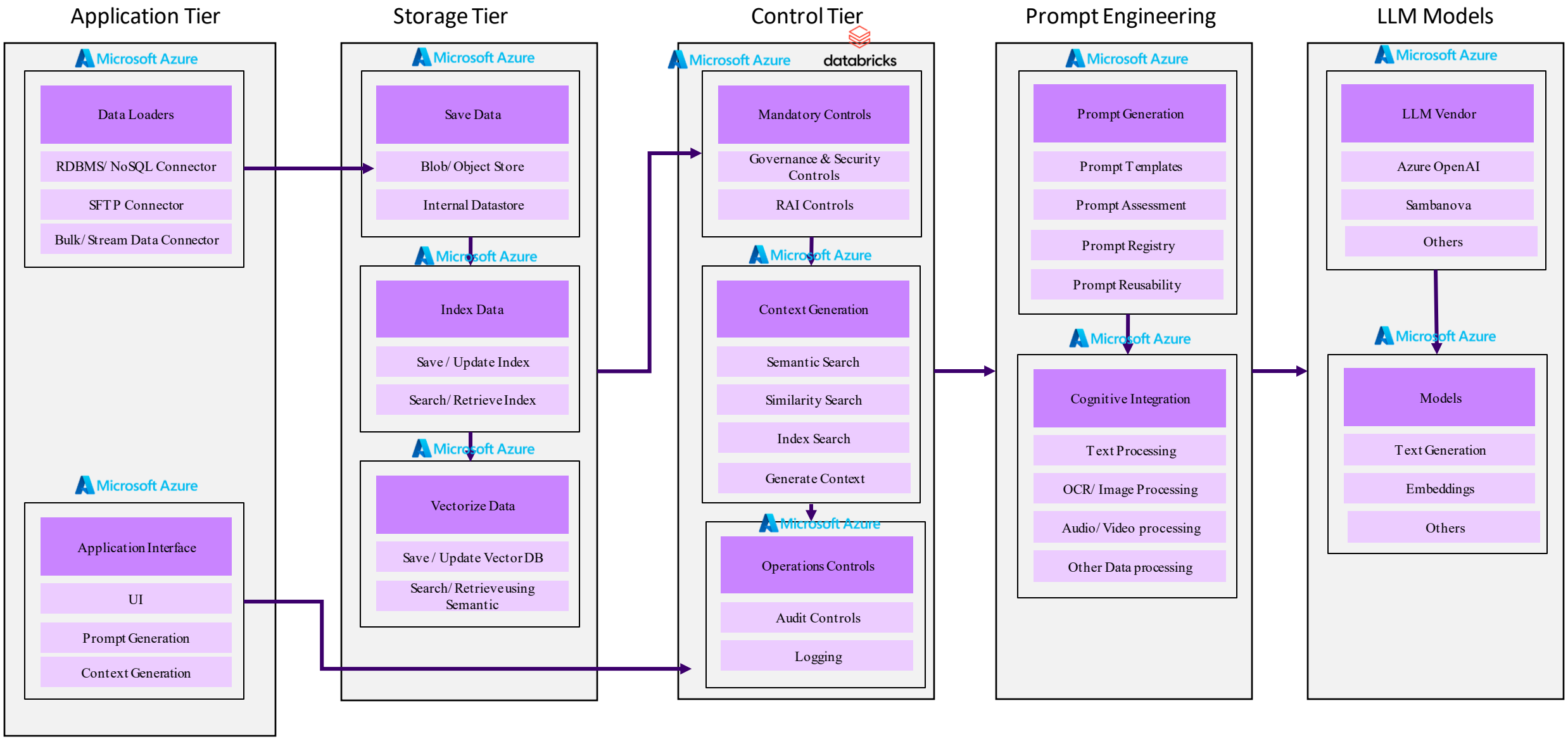


Copyright © 2023 Accenture. All rights reserved.

17

OUR LLM ARCHITECTURE BLUEPRINT

** Internal use only



Copyright © 2023 Accenture. All rights reserved.

LLMOps Overview

LLMOps allows for the efficient deployment, monitoring and maintenance of large language models. LLMOps, like traditional Machine Learning Ops (MLOps), requires a collaboration of data scientists, DevOps engineers and IT professionals. There are various dimensions of LLMOps and each of them is explained two sections

LLMOps Components

Prompt Engineering, RAI and Drift detection Dimension of LLMOps

Security and PII Dimensions

Copyright © 2023 Accenture. All rights reserved.

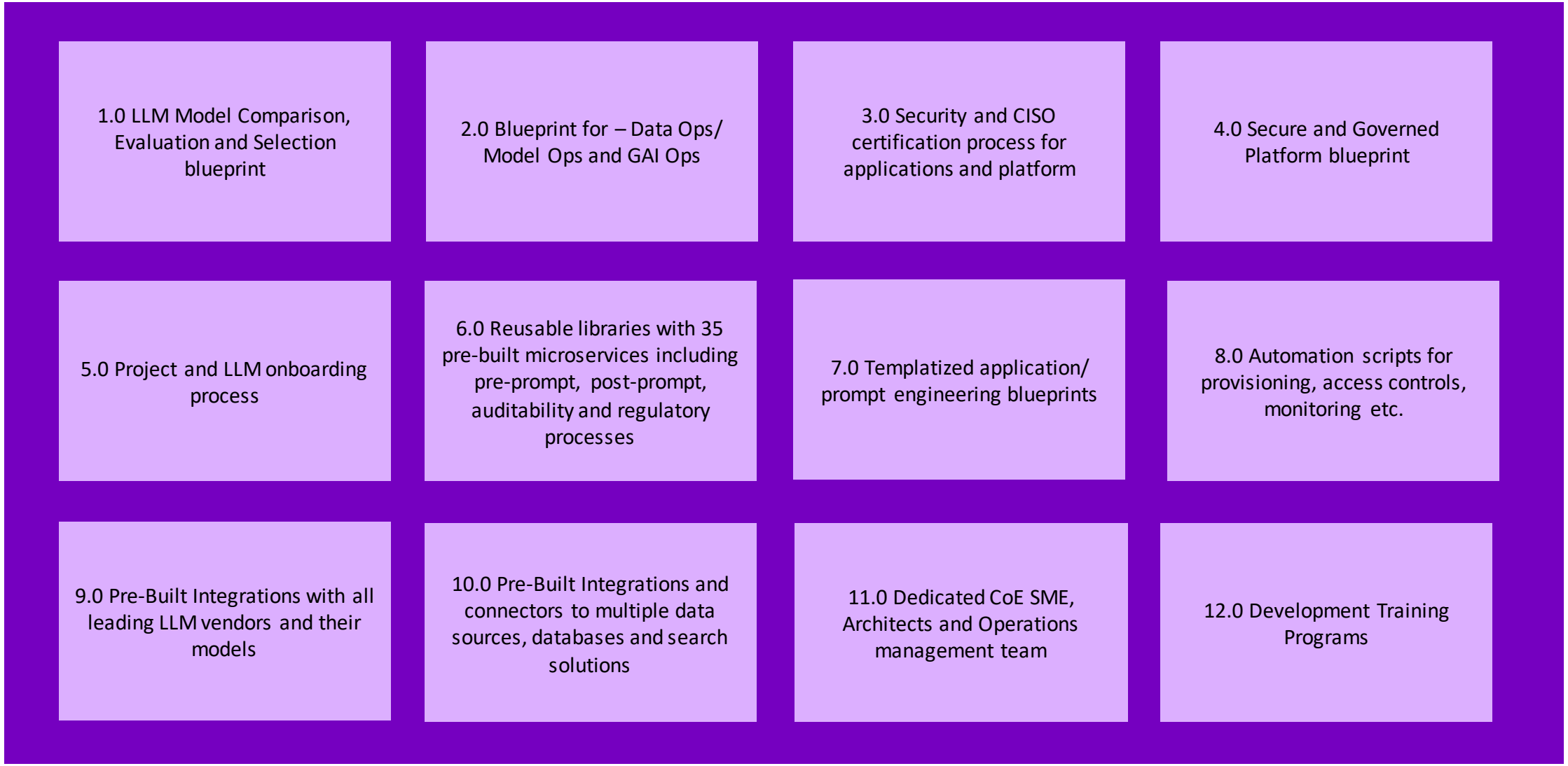
LLM Azure Assets and Offerings

- COE has developed and common microservices based offerings that help in re-usability and jump-start the implementation.
- Following slides will provide the list of the offerings that are available in COE-in-a-box packaging solution

Copyright © 2023 Accenture. All rights reserved.

CoE Sandbox – Enterprise LLM Adoption in a Box

** Internal use only



Copyright © 2023 Accenture. All rights reserved.

CoE Sandbox – Highlights (July 2023)

- ✓ LLM Sandbox Platform available in Azure cloud for provisioning.
  - ✓ *GCP and AWS in Progress
- ✓ LLM Sandbox available for
  - ✓ Azure Open AI
  - ✓ Sambanova
  - ✓ Nvidia Cloud Version
  - ✓ Scale AI
  - ✓ Co:Here
- ✓ 7 Client projects running in CoE –Hartford, Marriott, BMW, Savola, GMC, Corebridge, ADM and 2 in pipeline
- ✓ Overall Email request received from team for Onboarding into CoE = 177
- ✓ Shared Sandbox Costing ~ 25K USD till date

Copyright © 2023 Accenture. All rights reserved.

COE Managed Delivery Projects

Project	Project Description	Project POC
Hartford GenAI	Create a Q&A Bot for the underwriter's team at Hartford to improve their process/policies	revathi.kottaru@accenture.com
ADM	Create a bot based on the ADM specific data	kamakshi.subramaniam@accenture.com
GMC	Develop an FAQ document generation and Q&A bot based on the Accenture M&A data	v.a.visweswaraiah@accenture.com
CoreBridge	Create a chat interface based on the user guide and installation documents	v.a.visweswaraiah@accenture.com

Copyright © 2023 Accenture. All rights reserved.

Thank You!

Copyright © 2022 Accenture. All rights reserved.