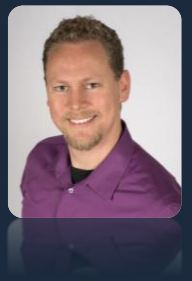
Arizona Cloud Innovation Summit

Microsoft

Get your Infrastructure AI Ready



Kyle Prawel
Sr. Azure Architect

042225

Infrastructure in the Age of AI – What are we here for?

Establish perimeters to foster collaboration through trust…

… resilient, performant design to encourage great user experience…

… cost, deployment, change controls with auditing…

… operationalize improvements, scale repeatable success…

Same as it always was…

… Infrastructure empowers development.

Learning From Other Projects…

- 95% of decision makers say AI is key but path toward implementation is unclear.[1]
- Taking 7+ months from AI pilot to production[3]
- 76% reported having trouble scaling AI efforts, 63% say repeatability is a key factor.[1]
- Only 54% of AI projects move to production[2]

Reference

What's the Commonality?

IDC research consistently shows that inadequate or lack of purpose-built infrastructure capabilities are often the cause of AI projects failing.

IDC Reference

… Most applications hold common context domains…

… Conway's Law and the merging of disparate teams…

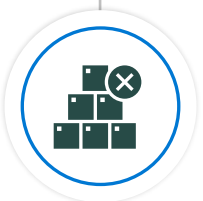
Infrastructure MUST establish Trust, which drives cross-context collaboration.

Centralized Governance Drives Velocity

AI Reference Architectures

Applying Governance FTW

Infra Governance Recommendations



Governance Checklists

When Done Correctly….

Teams trust data protection, friction reduced…

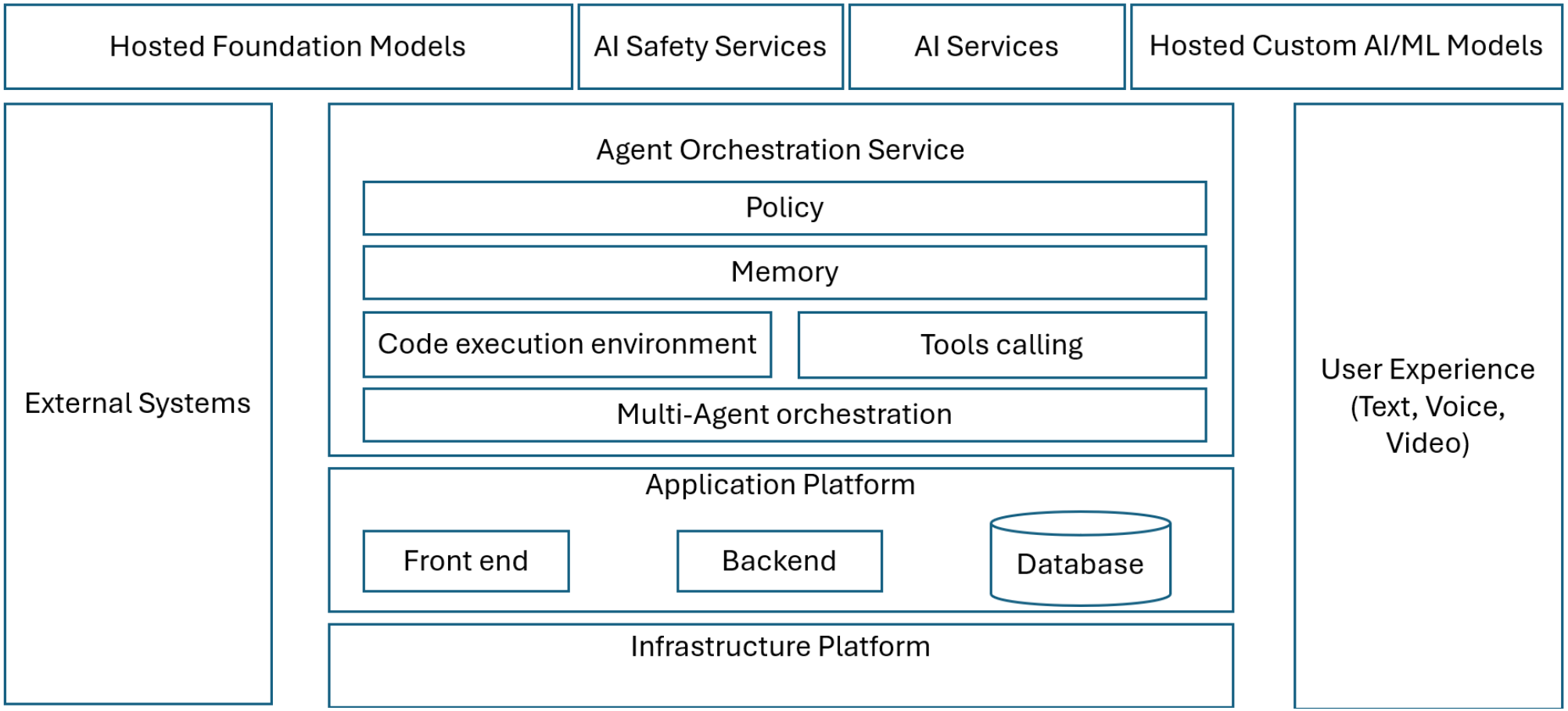
… Leadership knows costs are controlled, audited…

… Security knows perimeter controls are in place, audited, alerted…

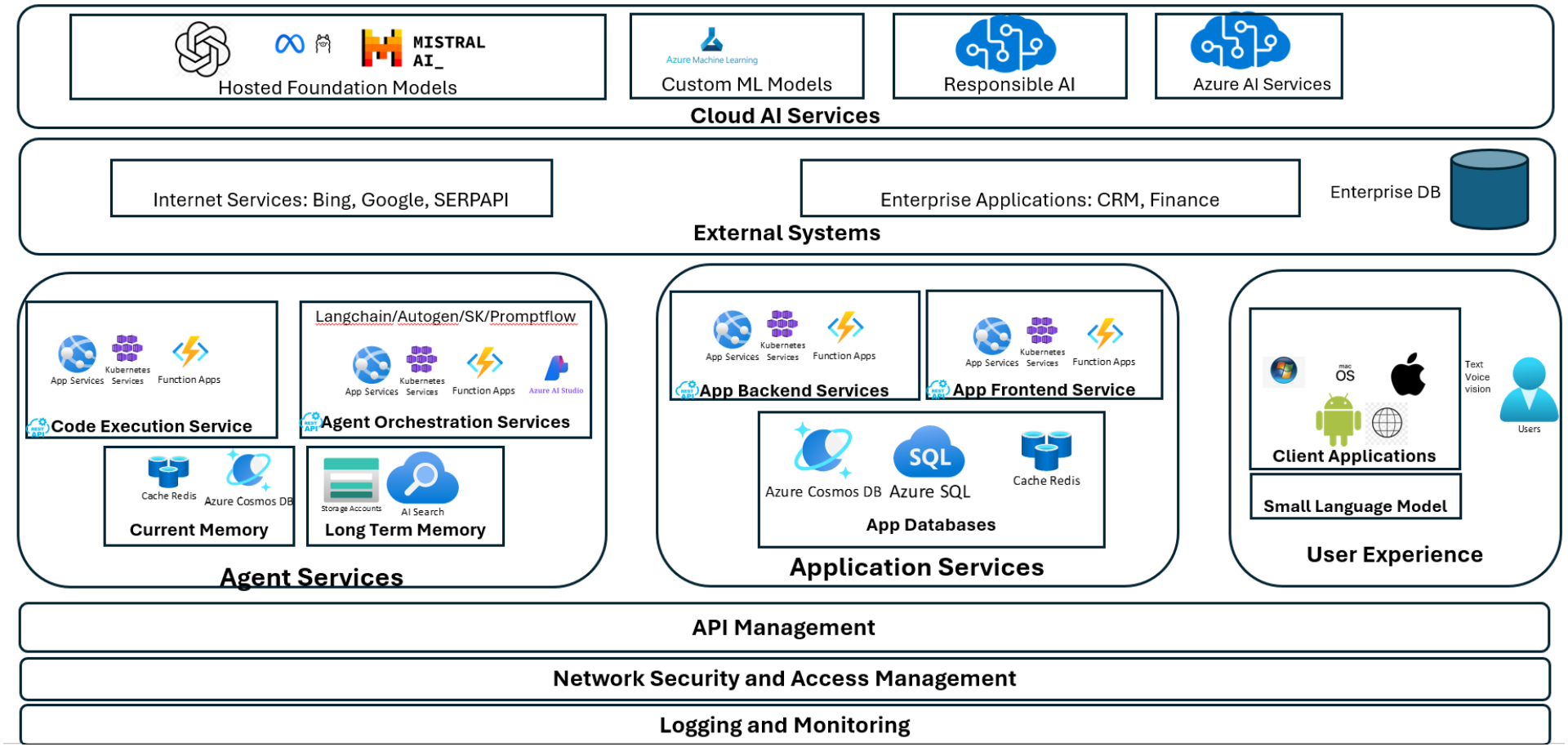
… developers have clear information on performance metrics, velocity targets…

… Infrastructure empowered development.

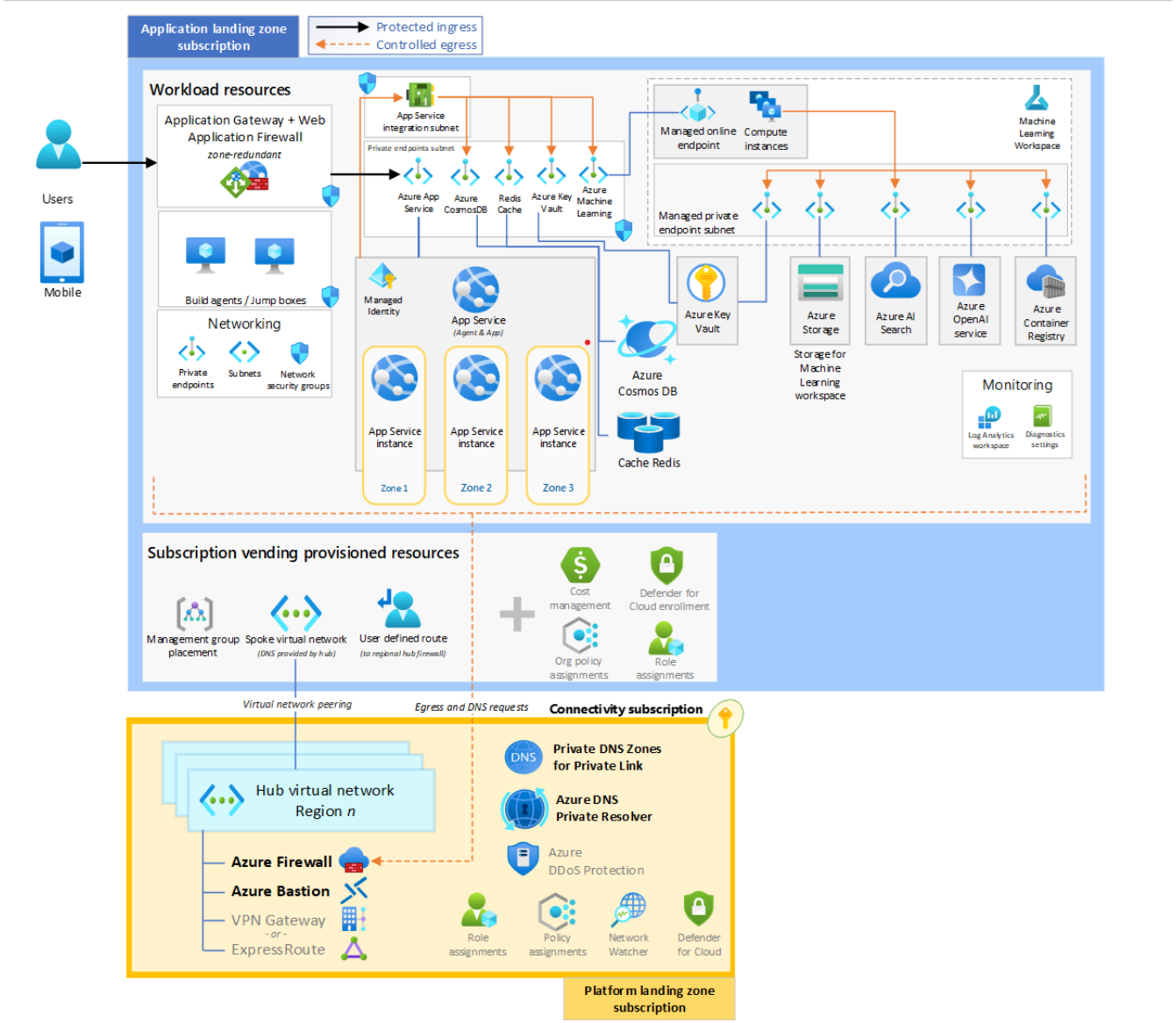
GenAI Reference Architecture – Logical View



GenAI Reference Architecture – Component View



GenAI Reference Architecture – Infrastructure View



Azure Landing Zone

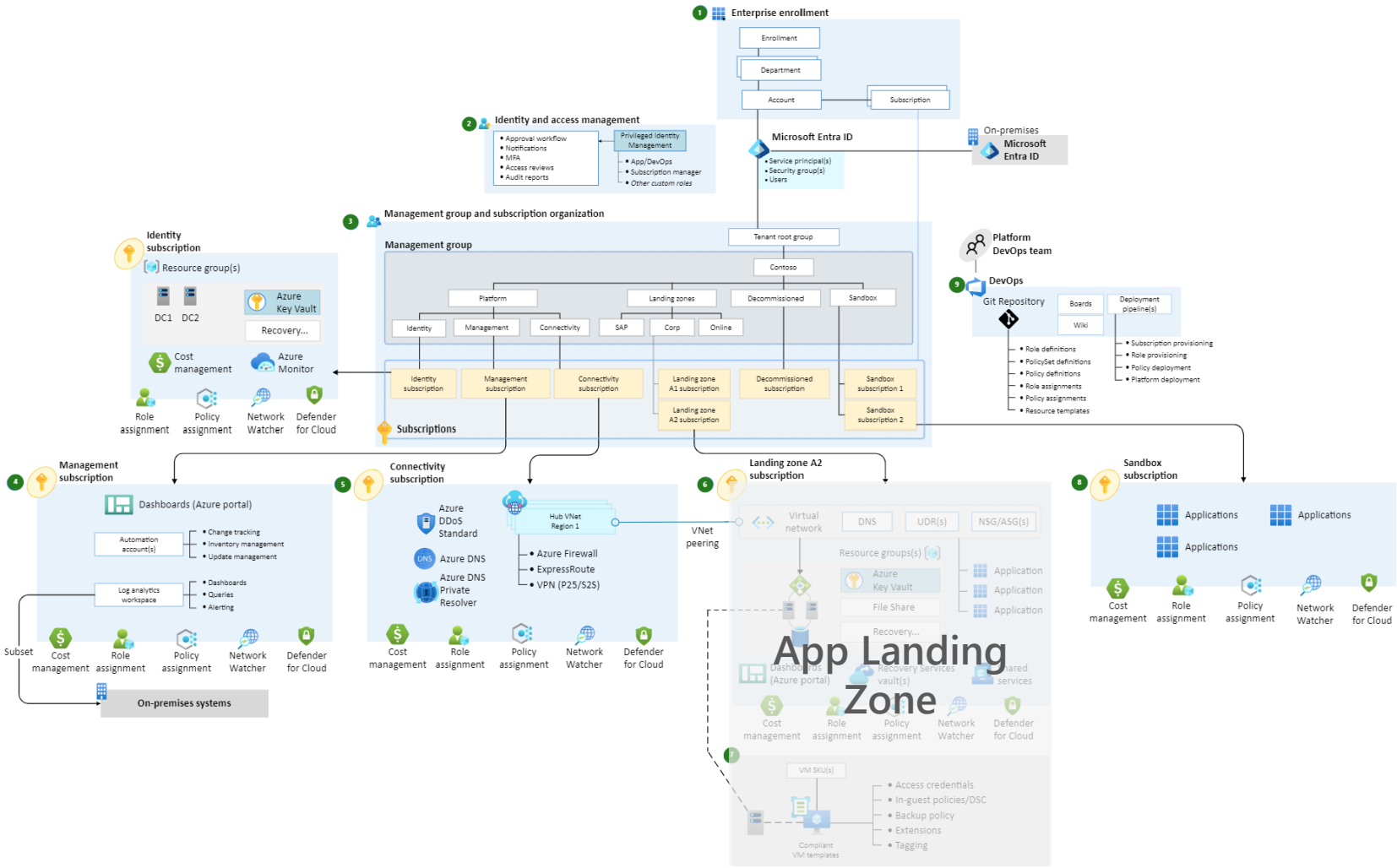
is a shared cloud foundation (Platform Landing Zone) to host your apps (Application Landing Zones)

Platform Landing Zone

Centralized and secured shared services (identity, connectivity and management) for all apps

App Landing Zone

App services, subscription, policy, networking and security



Common challenges in building green field AI Ready Infrastructure

- Security Principal Integration
- Secret Management
- Dev/Prod Segregation
- Automation/repeatability discipline using CI/CD

Key Azure Services


Azure Firewall


App Gateway


API Management


Azure Load Balancer


NSG


Traffic Manager

Azure DNS

Azure Private DNS Zone


Private Endpoints


Express Route


Entra ID Managed Identity

Getting a Brown Field to be an AI Ready Infrastructure

- Integrating with already well optimized CI/CD pipelines on AKS, Container Apps
- Building additional microservices using Container Apps, Function App
- Building LLM Apps using Azure AI Search, Prompt flow, Langchain
- LLM Ops for LLM evaluations, capturing feedback, Chat history
- Network configuration depending on PaaS Services

Key Azure Services


Azure AI Studio


Prompt flow


Azure Container Apps


Azure Functions


AKS


Azure Event Hubs

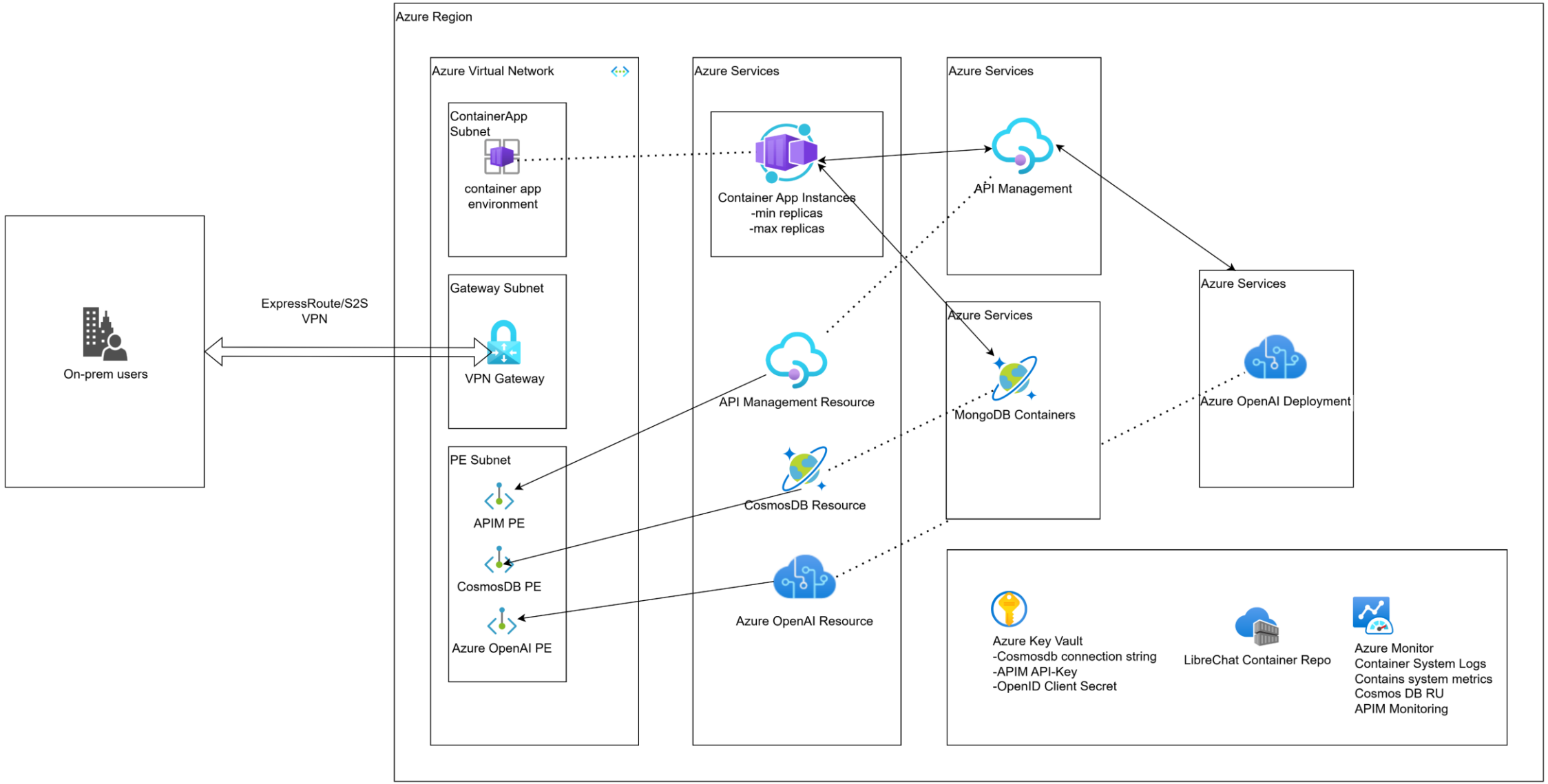

Azure Cosmos DB Mongo/SQL API


Azure ML Endpoints

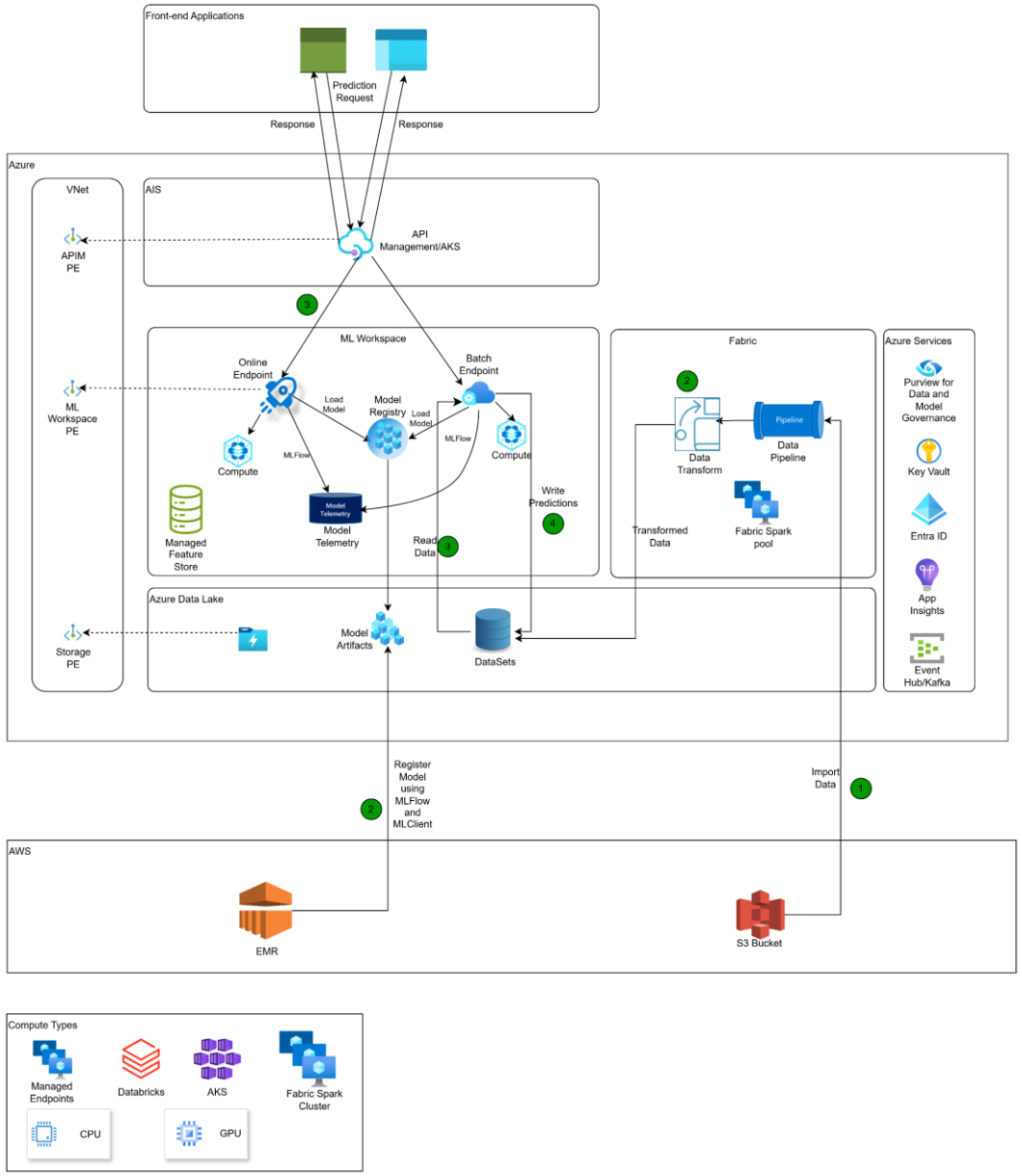

Azure AI Search


Azure Monitor

Brown Field simple architecture



Brown Field with PaaS Services



Centralized Resource Organization Recommendations

Recommendation	Benefit
Use management groups to enforce governance.	Management groups allow you to create and enforce governance policies on all resources within them. Create management groups that align with common workload needs. AI workloads often need to reach on-premises networks or require inbound connectivity from the internet.
Differentiate resources in terms of their connectivity and exposure to the Internet	The choice between using a "Corp" or "Online" management group within an Azure Landing Zone should be guided by the specific security requirements and the nature of cross-premises data consumption
Apply policies on the management group	Prefer audit policies (not deny) at the management group level. Audit policies allow application teams to manage their needs by creating sublevel policies or integrating audit requirements into their IaC deployment modules
Use subscriptions to separate application environments	Create separate subscriptions for development, test, and production environments to isolate and manage AI application resources. All the application environments for a single AI application should be in the same management group

Centralized Resource Organization Recommendations

Recommendation	Benefit
Use resource groups for resources that share a common lifecycle.	A shared AI resource, such as Azure AI Studio or Azure OpenAI, should have its own subscription and resource group. Resource groups also provide additional governance (policy), security (RBAC), and cost (budget) boundaries.
Standardize AI naming conventions.	Use a standard naming convention for Azure AI and machine learning resources so you can easily track AI resources. Understand the naming rules and restrictions for each Azure resource to develop resource names. Use the recommended AI and machine learning abbreviations since resources have name-length restrictions.
Enforce tag usage.	Use Azure policy to enforce tags for categorizing resources according to criteria such as ownership, environment, or application. Tags facilitate governance tasks like cost tracking, reporting, and automated policy application. Tags also support automation efforts in resource provisioning and deprovisioning.
Protect resources.	Apply resource locks to safeguard critical shared services from accidental deletion. Use deny policies in conjunction with Azure role assignments to prevent unauthorized resource deployments and configurations

Centralized Cost Governance Recommendations

Recommendation	Benefit
Use tags to allocate costs.	Configure an Azure Policy definition to enforce tagging on resources. Use tags to categorize resources by project, cost center, environment, and owner for better management and billing.
Use tag inheritance.	Use the tag inheritance setting in Cost Management. When enabled, tag inheritance applies billing, resource group, and subscription tags to child resource usage records
Manage billing accounts.	Use Microsoft Billing to manage your billing account and pay invoices. Ensure each AI project or team has a designated billing account to accurately track expenses.
Monitor costs.	Use Microsoft Cost Management to set budget alerts, cost anomaly alerts, and scheduled alerts for different AI projects. This helps in tracking spending against allocated budgets and maintaining financial discipline.
View spending patterns.	Use the Azure Cost analysis to tool to regularly review spending patterns to identify trends and areas for potential savings in VM usage.

Centralized Cost Governance Recommendations

Recommendation	Benefit
Allow specific virtual machine SKUs.	Use Azure policy to allow only the virtual machines SKUs that align with your AI budget. The built-in policy definition Allowed virtual machine SKUs in Azure policy to allow the deployment of only the SKUs you need.
Block resources.	Use policies to prevent the deployment of certain expensive resources.
Block expensive models.	Understand the costs of different models vary depending on which model series you choose. For example, there might be charges per 1,000 tokens for both input and output tokens. Being aware of this can help you choose the most cost-effective model series for your needs.
Auto shutdown compute instances.	Configure the schedule and idle shutdown of compute instances. This can significantly reduce costs by ensuring you don't pay for unused compute time.
Monitor Costs and Set Up Alerts.	After you start using Azure AI services, use Cost Management features to set budgets, monitor costs, and set up cost alerts. You can review forecasted costs and identify spending trends to find areas where you might want to act. Additionally, set up cost alerts to notify you when you exceed your budgeted amount. This comprehensive approach will enable you to effectively manage and control your costs.

Centralized Operational Governance Recommendations

Recommendation	Information
Configure alerts.	Enable recommended alert rules to receive notifications of metric deviations.
Use CI/CD pipelines.	Use CI/CD pipelines to automate the testing and deployment phases of your AI workloads. Automate the deployment of AI models and related infrastructure through GitHub to reduce manual errors and improve efficiency. Leverage release gates and approval workflows to maintain control over deployment processes.
Use infrastructure as code	Use IaC with tools like Azure Resource Manager templates or Terraform allows for the declarative and consistent provisioning of required infrastructure, adhering to best practices and compliance requirements.
Version-control infrastructure changes.	Track and version-control IaC configurations in a source repository. This practice provides transparency in infrastructure changes and simplifies rollbacks in case of issues.

Centralized Regulatory Compliance Recommendations

Recommendation	Information
Develop industry-specific compliance checklists.	Compile thorough checklists reflecting the regulatory demands relevant to your industry, such as CJIS, HIPAA, or financial services regulations. This step helps ensure that your AI initiatives remain compliant from the onset.
Enforce data residency and handling policies.	Data residency requirements stipulate that data must be stored within specific geographic boundaries and handled following local laws. This is critical for multinational or multi-regional deployments of AI solutions on Azure. Utilize Azure's region-specific services and configurations to adhere to these regulations.
Azure Policy.	Enforce compliance with built-in policy definitions. Azure Policy Documentation.
Microsoft Purview Compliance Manager.	Manage compliance with a dashboard that offers insights and recommendations. Azure Compliance Manager Documentation.
Azure Machine Learning.	Features supporting responsible AI such as model interpretability and transparency. Azure Machine Learning documentation
Azure Data Residency and Sovereignty.	Select region-specific services to adhere to data protection laws. Azure Geography and Regional Services Azure Geography and Regional Services

Centralized Regulatory Compliance Recommendations

Recommendation	Information
Use built-in regulatory compliance policies.	Use the applicable regulatory compliance initiatives for your industry. Apply additional policies based on the AI services you're using. •Azure Studio and AI services, see Azure Policy Regulatory Compliance controls. For •Azure Machine Learning, see Azure Policy built-in policy definitions for Azure Machine Learning. •Azure Virtual Machines, see Built-in policy definitions for Azure Virtual Machines
Consult common standards.	Use standards, such as ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), to audit your policies applied to AI workloads.

Centralized Security Governance Recommendations

Recommendation	Benefit
Centralize identity management.	Use Microsoft Entra ID as the central identity provider for all users, applications, and services. Have one team that manages identity for all AI workloads. This ensures a unified and consistent approach to identity management.
Have a central team create subscriptions for AI application resources.	Configure each subscription with the least privilege access. But allow the application team to manage their own resources using Azure roles.
Configure least privilege access to centralized AI resources.	For centralized AI resources shared by multiple applications and application teams, start by granting minimum access, such as the Reader (Azure role) and elevating to Contributor (Azure role) if the limited permissions slow development too much.
Activate Azure Defender.	Enable Azure Defender to provide advanced threat protection for your workloads, including virtual machines, storage accounts, and databases.
Implement distinct access controls for each environment.	Protect resources from accidental deployments by limiting each deployment pipeline's identity to its specific environment.
Govern infrastructure as code.	Govern IaC templates by using Microsoft Defender for Cloud to Detect IaC misconfigurations and enforce secure deployments.

Centralized Security Access Recommendations

Recommendation	Benefit
Use Microsoft Entra ID for authentication and authorization by using RBAC and MFA using conditional access policies.	Enhances security by ensuring only authorized users can access the system using the least privilege access. Use Microsoft Entra ID
Use Managed identities for authentication between Azure AI services. Example: Machine learning and Azure AI search.	Managed identities improve security by eliminating the need to store credentials and manually manage and rotate service principals. Use Managed Identities for authentication.
Disable public access to AI services unless your workload requires it. Enable private endpoints for your Azure AI services to keep the traffic private.	Controlling access to Azure AI services helps prevent attacks from unauthorized users. Using private endpoints ensures network traffic remains private between the application and the platform. Disable Public access & Create Private endpoints
Don't provision public IP addresses for ML compute. Always set enableNodePublicIp to false when provisioning ML compute clusters	Refrain from provisioning public IP addresses to enhance security by limiting the potential unauthorized access to your compute clusters. Disable public IP

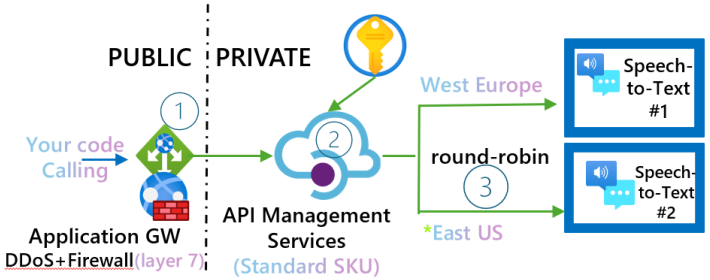
Centralized Data Security Recommendations

Recommendation	Benefit
Encrypt data at rest using customer-managed keys when possible.	• Encrypting data at rest enhances data security by ensuring that sensitive data is encrypted by using keys directly managed by you. • If you have a regulatory requirement to manage your own encryption keys, use this feature to comply with the requirement. Azure OpenAI Service encryption of data at rest
Secure keys using Azure key vault when using Azure AI services.	• Separating secrets from code by storing them in Key Vault reduces the chance of leaking secrets. • Separation also facilitates central management of secrets, easing responsibilities like key rotation. Azure OpenAI Service encryption of data at rest
Enable defender for cloud to get recommendations on your AI application.	• Defender for cloud can discover vulnerabilities within your Gen AI libraries by scanning your source code and container images. • Scanning your data sources to uncover risks that might lead to data breaches. Enable threat protection for AI workloads

Centralized Network Security Recommendations

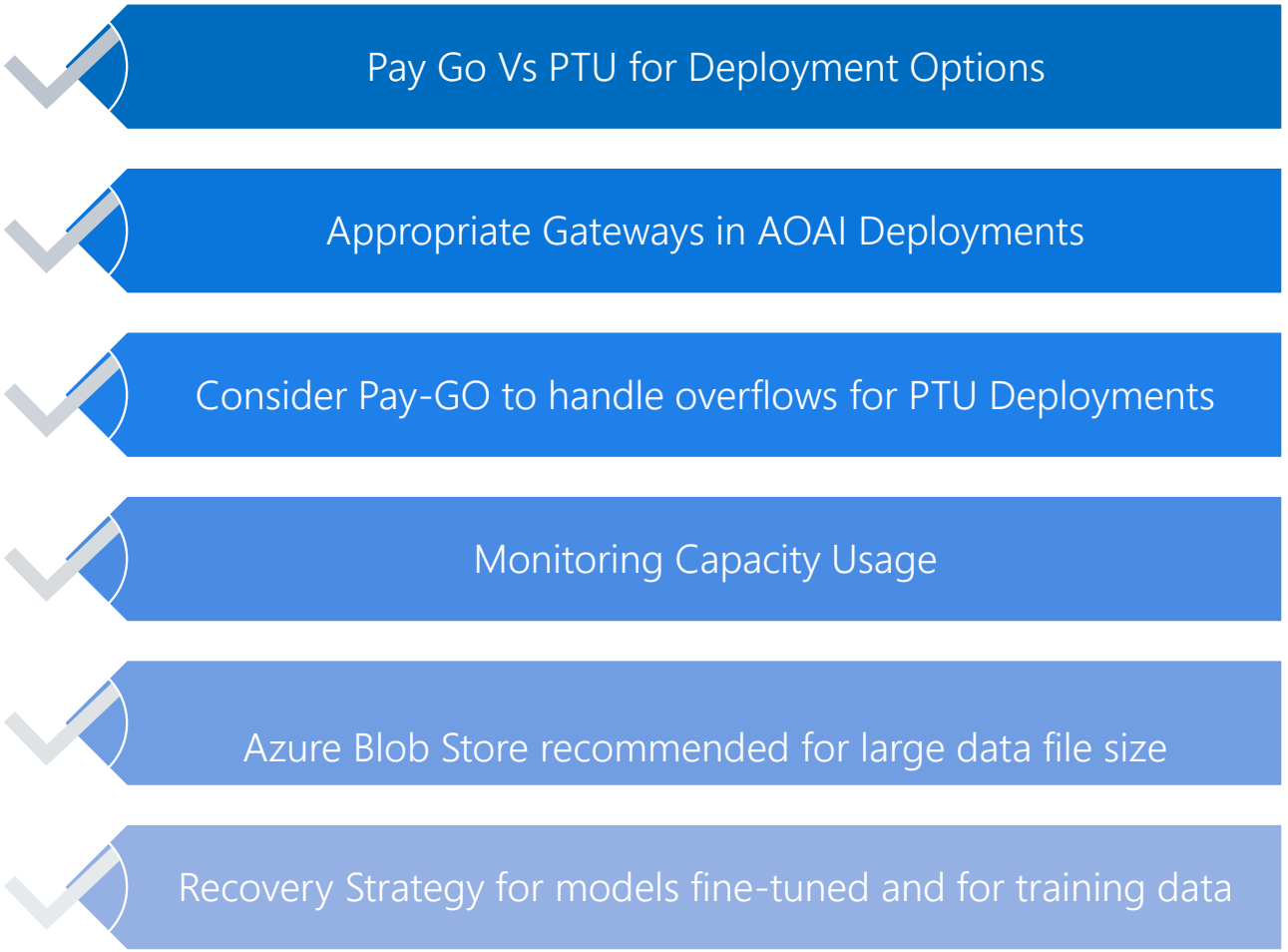
Recommendation	Benefit
Apply Network Isolation to control the traffic flow between the Azure AI dependent services.	Network Isolation reduces the risk of lateral movement in case of a breach. Managed virtual network isolation for Azue machine learning
Implement NSGs for your AI workloads VNETs to ensure only allowed traffic is granted.	NSGs help with following the least privilege access control and restrict unauthorized access to your environment Configure inbound and outbound network traffic - Azure Machine Learning | Microsoft Learn
Disable public ports where applicable. Example; ensure SSH port is closed on your ML clusters.	Disabling SSH access helps prevent unauthorized individuals from gaining access and potentially causing harm to your system and protects you against brute force attacks. Disable the public SSH port
Use the network firewall in the hub to inspect and filter network traffic flowing in/out of your AI workload spoke.	A network firewall like an Azure firewall will help with traffic inspection and control for inbound and outbound traffic. Use Azure Firewall
Deploy private endpoints for all Azure resources that are dependent on your AI workload. Example; Azure key vault, Azure storage, container registry, etc..	Network isolation bolsters security by restricting access to Azure platform as a service (PaaS) solutions to private IP addresses only. Deploy Private endpoints

Networking Recommendations

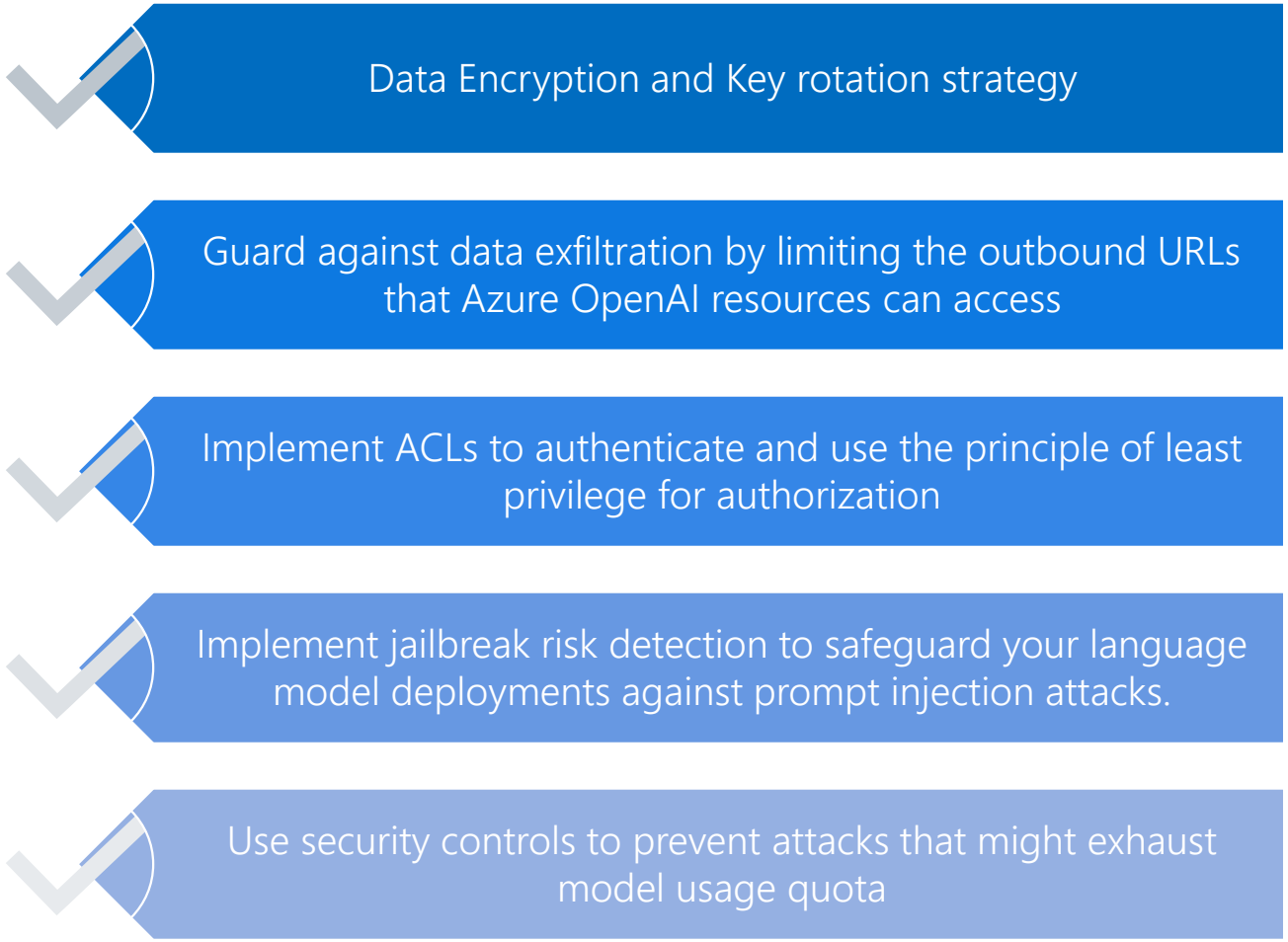


Recommendation	Benefit
Private networking: "Corp" type LZ vNets (no public IP policy) for internal AI solutions, which default with private link and no-public IP policy. --Corp & Online for storage problem	• Optimal routing and minimal latency for branch and Azure connectivity. Most AI services support private endpoints. • Connects seamlessly to on-premises ESLZ - Corp VS Online Application LZ's
Azure Virtual WAN & Express route to on-premises, and Azure Firewall in hub. Avoid Bastion. Control via Policys.	• Optimal routing and minimal latency for branch-to-branch and branch-to-Azure connectivity • ExpressRoute offers fast, reliable, private connections, with Hub firewall for centralized security and management Private networking & Firewall - Azure
AI powered Apps: APW+API mgmt. with load balancing with network segmentations	• Build highly secure, scalable web front ends using Azure Load Balancer or Application Gateway for high availability and performance. • Segment your network into isolated zones with security controls at each boundary. Loadbalancing - best practices
Leverage AI services with global endpoint. AI services, such as Speech API, with built in security. Load-balance: round-robin/weighted multiple endpoints to load-balance in regions	• Already have AGW and API mgmt as frontend, managed by Azure. • Note that a LZ with no-public IP access will not allow provisioning, without private endpoint. Enterprise Scale AIFactory on Azure - Networking PrivateDNS & PaaS

Reliability Design Checklist



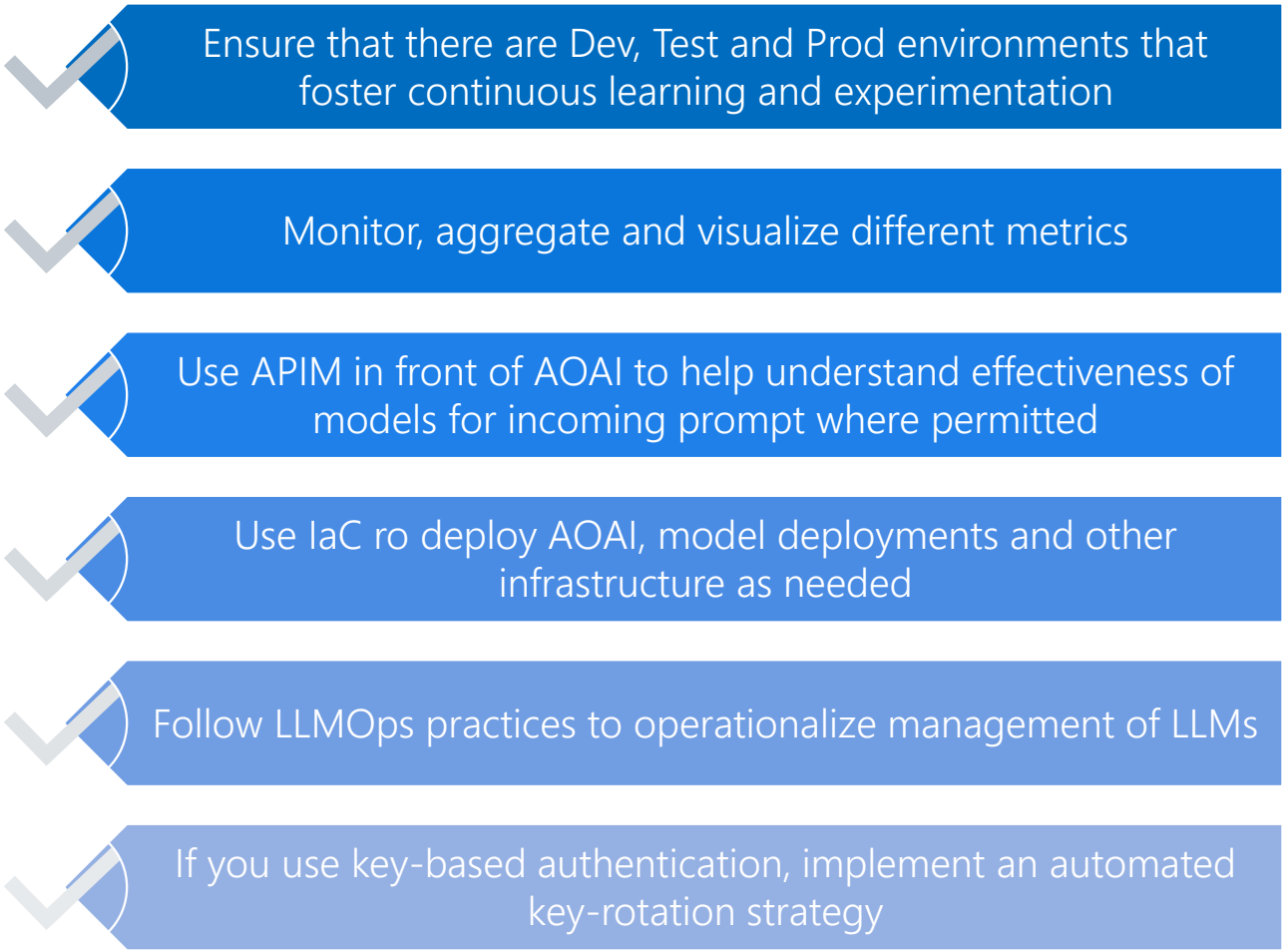
Security Design Checklist



Cost Optimization Checklist

- Develop cost model considering prompt input and response sizes
- Optimize usage, use Pay-As-you-Go until token usage is predictable
- Use PTU when token usage is predictable and sufficiently high over a period of time
- Maximize Azure OpenAI price breakpoints
- Consider model pricing and capabilities when you choose models.
- Set up cost-tracking system and monitor model usage
- Batch requests where possible to minimize the per-call overhead
- Consider costs of different models if your solution requires fine-tuning

Operational Excellence Checklist



Performance Efficiency Checklist

- Choose between high and low priority traffic for synchronous responses vs responses that can be batched
- Leverage AOAI benchmarking tool to validate PTU deployments
- Use PTU for production workloads
- Utilize gateways to route traffic in same or different regions
- Model selection based on performance requirements, with tradeoff between speed and output complexity
- For conversational interfaces consider streaming to enhance perceived performance
- Consider when to use fine-tuning carefully
- Consider using dedicated model deployments per consumer group to provide per-model usage isolation

Microsoft

Thank you!

For post-summit feedback, follow-ups, or to connect with one of our Microsoft resources, please email – Arizona Cloud Innovation Summit at acis@microsoft.com