Microsoft

Microsoft AI Tour

Microsoft

Securing AI applications on Azure

Matthew WONG
Senior Cloud Solution Architect - Microsoft

Jenius Shieh
Partner | DarkLab | Cyber Threat Operations - PWC

Agenda

1. Introduction on New Threats
2. AI safety
3. Microsoft and PWC Responsible AI Framework Model
4. Demo on AI Model Scan
5. Demo on Prompt Injection and AI Threat Protection
6. Quiz

Introduction

Top risks and concerns of generative AI

Data oversharing and data leaks

80%+

of leaders cited leakage of sensitive data as their main concern[1]

Emerging AI risks and threats

77%

of orgs are somewhat concerned about indirect prompt injection attacks and **11% are extremely concerned**[2]

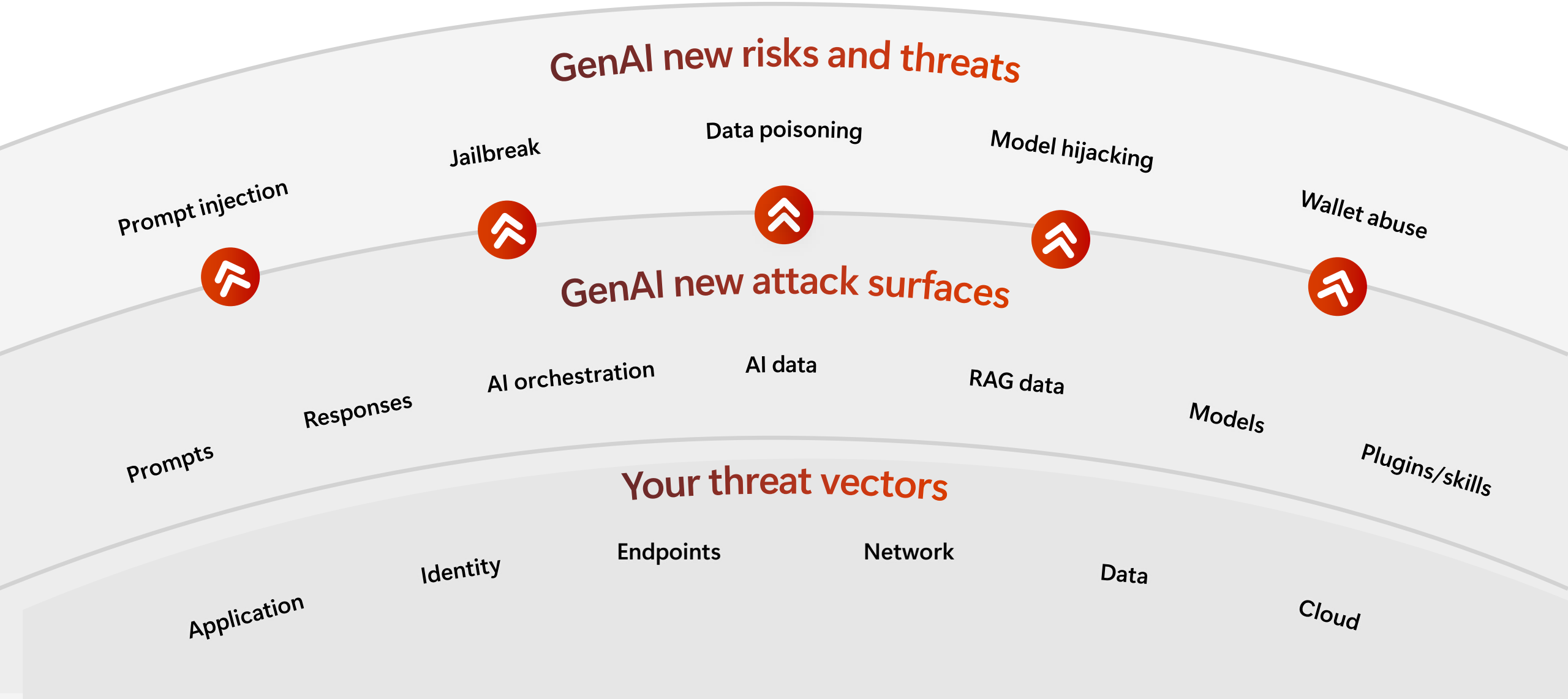
Model vulnerabilities

50%

of open-source generative AI models will underpin **more than 50% of enterprise generative use cases**[3]

1. First Annual Generative AI study: Business Rewards vs. Security Risks, , Q3 2023, ISMG, N=400
2. Gartner®, Gartner Peer Community Poll – If your org's using any virtual assistants with AI capabilities, are you concerned about indirect prompt injection attacks? GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.
3. Gartner®, Innovation Guide for Generative AI Models, 16th April 2024. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

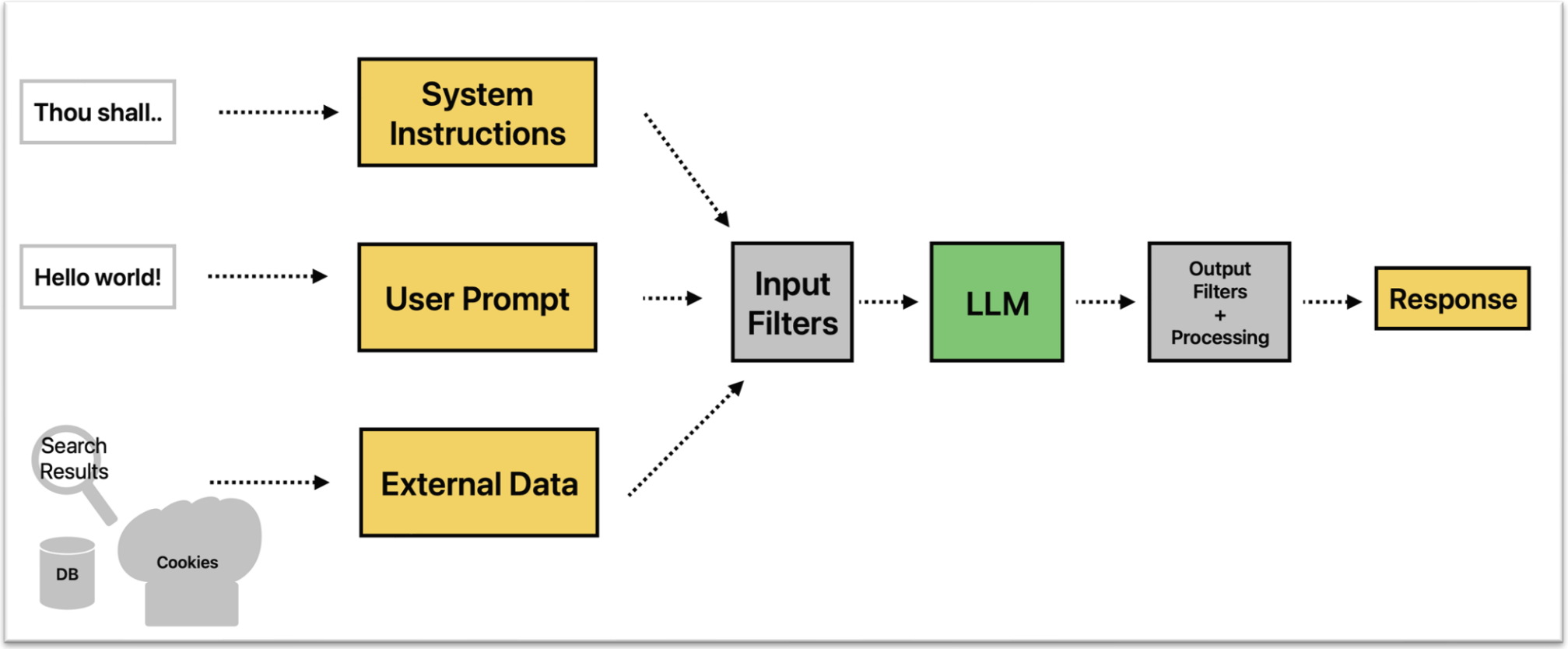
New attack surfaces introduce new risks and threats



Intrinsic and extrinsic risks arise from GenAI models, making the application of sufficient guardrails for responsible AI even more critical

Risks	Guardrails
INTRINSIC – risks posed by Generative AI models	
Writing generic and even bespoke code becomes automated, reducing the demand for human programmers	Put in specific processes to review code and ensure that outputs are consistent. E.g. for data science, MLOps can help to monitor outputs and maintain code quality
Administrative and even low level creative work becomes automated	Governments and organizations should ensure that educational policies are up to date and focus on value-adding skills in an AI-driven workforce
Academic work becomes prone to replacement with lack of original creativity and concerns around attributing credit to humans	Educational institutions should work closely with AI experts to create clear guidance on how AI should be used in academia
Model is prone to bias based on input data	Continue to review model outputs and provide oversight on what data the model is being exposed to, aligned with existing governing principles for AI model usage
EXTRINSIC – risks posed by human misuse of Generative AI	
Generative AI models could be mis-used for the following: • Deepfake • Phishing emails and social engineering • Malware • Copyright infringement • Abusive or harmful content • Disinformation + propaganda	**Developers/Creators** – apply responsible AI principles, such as those from Microsoft, Google and OpenAI. Prioritise nimble, iterative responses to events. Educate users on using generative AI as a tool. **Individual users** – apply a critical lens to generative AI. Maintain integrity of applications using generative AI – always validate an answer **Commercial** – apply generative AI with an understanding of the risks. Develop policies around intellectual property used for generative AI products, particularly for proprietary training data. Ensure policies apply to third-parties you are working with.

Typical LLM Inference Pipeline



Generative AI threat landscape

DARKREADING

Cybersecurity Topics | World | The Edge | DR Technology

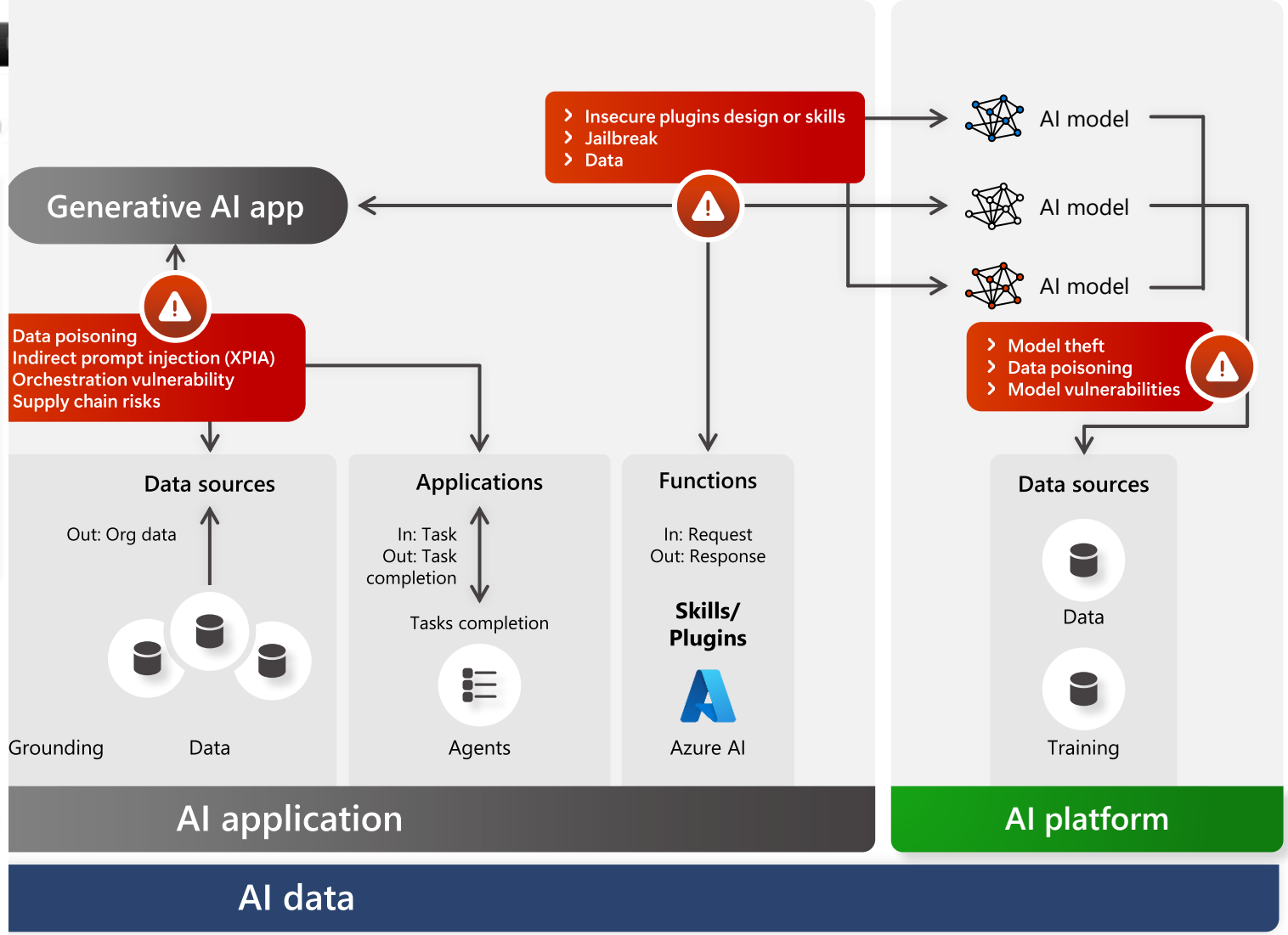
Elizabeth Montalbano, Contributing Writer
February 29, 2024

5 Min Read

SOURCE: WRIGHTSTUDIO VIA ALAMY STOCK PHOTO

Researchers have discovered about 100 machine learning (ML) models that have been uploaded to the Hugging Face artificial intelligence (AI) platform and potentially enable attackers to inject malicious code onto user machines. The findings further underscore the growing threat that lurks when attackers poison publicly available AI models for nefarious activity.

The discovery of the malicious models by JFrog Security Research is part of ongoing research by the firm into how attackers can use ML models to compromise user environments, according to a blog post published this week.



LLM01: 2025 Prompt Injection

LLM01:2025 Prompt Injection

A Prompt Injection Vulnerability occurs when user prompts alter the...

Read More

LLM02: 2025 Sensitive Information Disclosure

LLM02:2025 Sensitive Information Disclosure

Sensitive information can affect both the LLM and its application...

Read More

LLM03: 2025 Supply Chain

LLM03:2025 Supply Chain

LLM supply chains are susceptible to various vulnerabilities, which can...

Read More

LLM04: 2025 Data and Model Poisoning

LLM04:2025 Data and Model Poisoning

Data poisoning occurs when pre-training, fine-tuning, or embedding data is...

Read More

LLM05: 2025 Improper Output Handling

LLM05:2025 Improper Output Handling

Improper Output Handling refers specifically to insufficient validation, sanitization, and...

Read More

LLM06: 2025 Excessive Agency

LLM06:2025 Excessive Agency

An LLM-based system is often granted a degree of agency...

Read More

LLM07: 2025 System Prompt Leakage

LLM07:2025 System Prompt Leakage

The system prompt leakage vulnerability in LLMs refers to the...

Read More

LLM08: 2025 Vector and Embedding Weaknesses

LLM08:2025 Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities present significant security risks in systems...

Read More

LLM09: 2025 Misinformation

LLM09:2025 Misinformation

Misinformation from LLMs poses a core vulnerability for applications relying...

Read More

LLM10: 2025 Unbounded Consumption

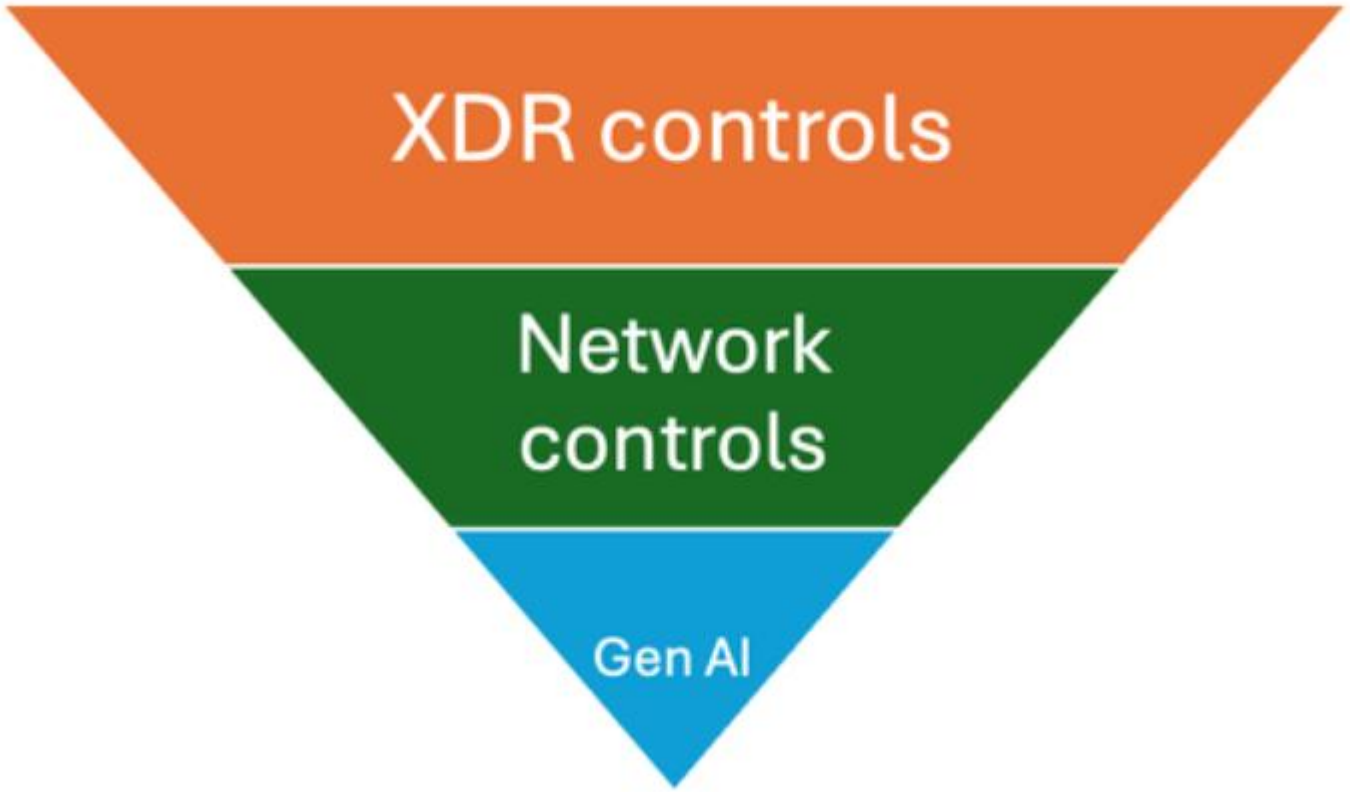
LLM10:2025 Unbounded Consumption

Unbounded Consumption refers to the process where a Large Language...

Read More

https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-v2025.pdf

Layered Defense using Native Azure Security Services

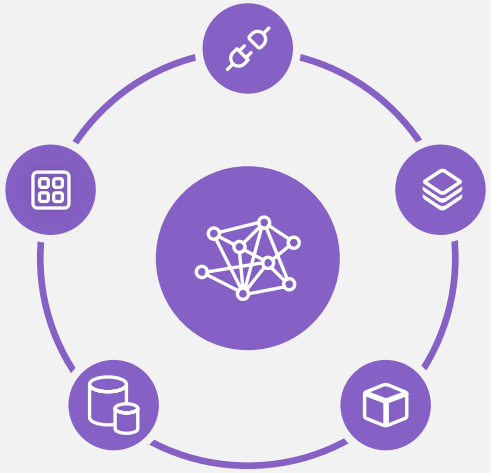


https://techcommunity.microsoft.com/blog/microsoftdefendercloudblog/securing-multi-cloud-gen-ai-workloads-using-azure-native-solutions/4222728

Protect AI apps from code to runtime

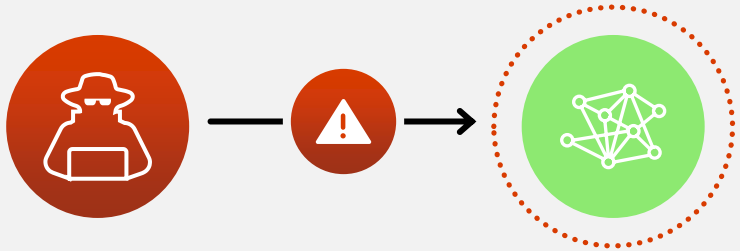
Start secure

AI security posture management (AI-SPM)



Stay secure

Threat protection for AI workloads



Microsoft Defender for Cloud

AI Safety

Microsoft's Responsible AI Principles

Fairness
AI systems should treat all people fairly.

Reliability and safety
AI systems should perform reliably and safely.

Privacy and security
AI systems should be secure and respect privacy.

Inclusiveness
AI systems should empower everyone and engage people.

Transparency
AI systems should be understandable.

Accountability
People should be accountable for AI systems.

Microsoft's Responsible AI Principles

Fairness
AI systems should treat all people fairly.

Reliability and safety
AI systems should perform reliably and safely.

Privacy and security
AI systems should be secure and respect privacy.

Inclusiveness
AI systems should empower everyone and engage people.

Transparency
AI systems should be understandable.

Accountability
People should be accountable for AI systems.

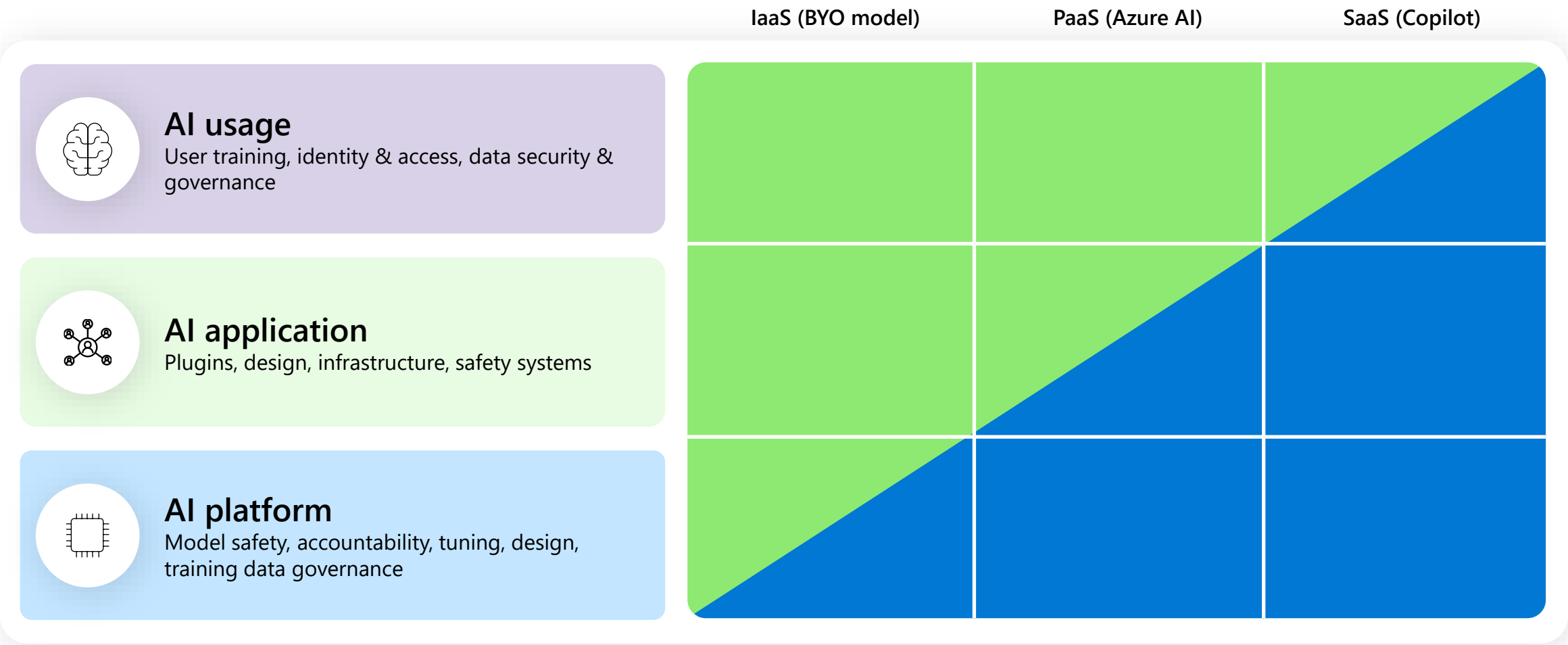
PwC's Responsible AI Framework

Core Elements of a Responsible AI Framework

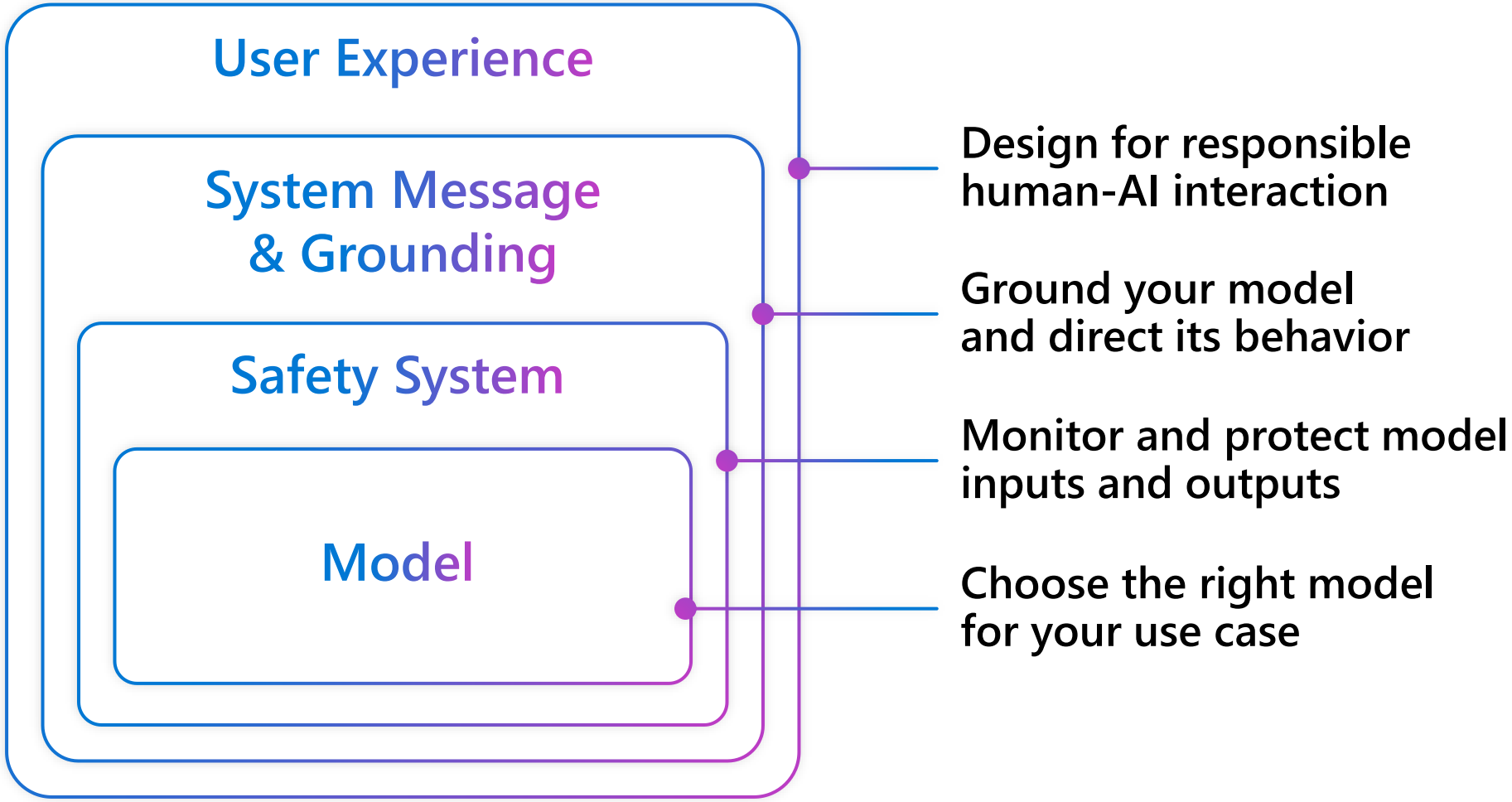
Foundational Capabilities	Operating Model and Governance	Application Lifecycle
Responsible AI Principles	Operating Model - Roles & Responsibilities	AI Development and Deployment Standards
AI Use Case Inventory	Governance Committee and Escalations	AI Testing and Monitoring
AI Risk Taxonomy	AI Risk and Control Matrix	Risk Mitigation Tracking and Reporting
AI Risk Intake and Tiering	Training and Communication	

Policies and Procedures Across Risk Domains (e.g., cyber, privacy, legal, model risk)

AI security shared responsibility model



Risk mitigation layers



Safety Models

Update content filter

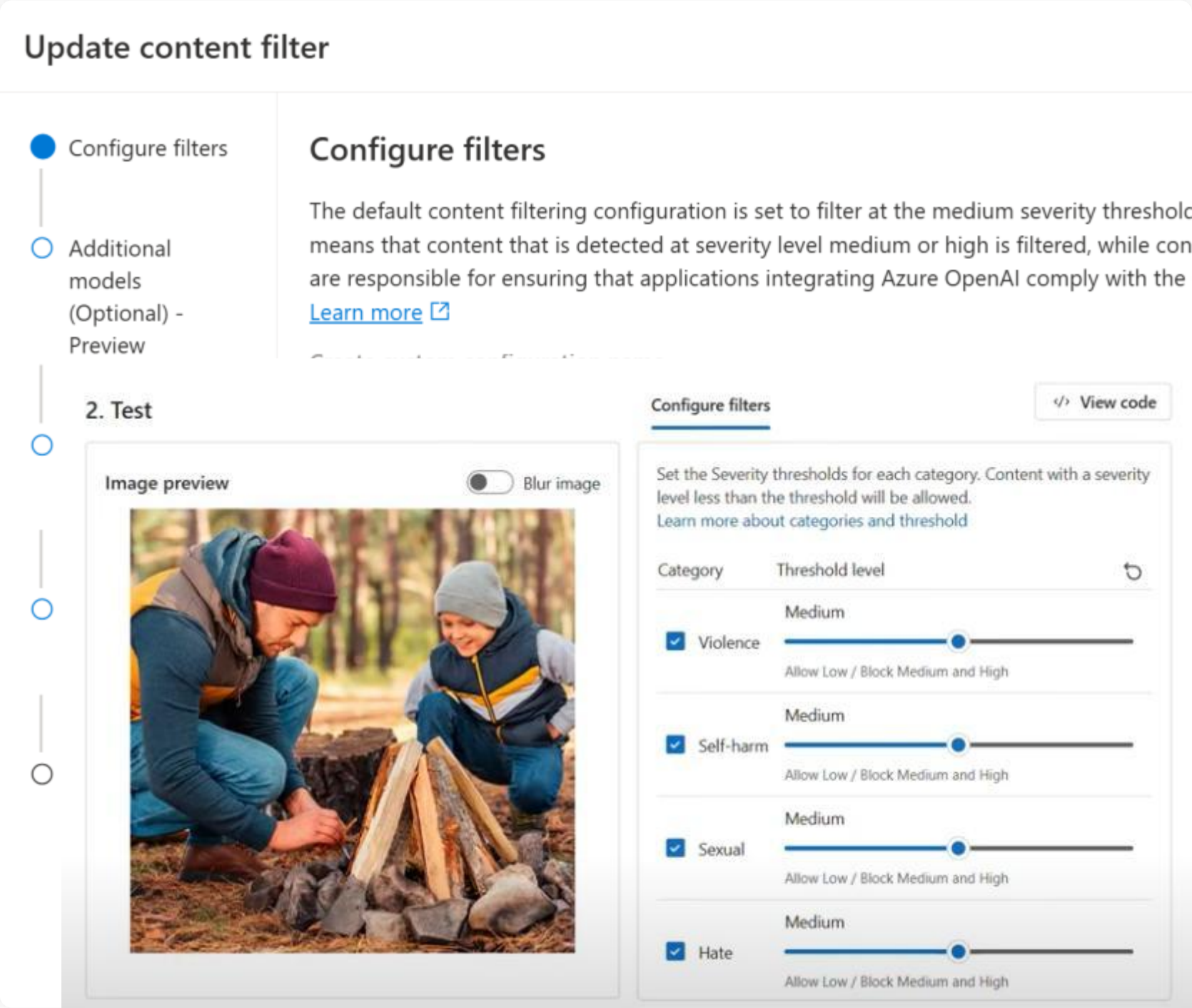
- Configure filters
- Additional models (Optional) - Preview
- Add blocklist (Optional) - Preview
- Streaming mode (Optional) - Preview
- Review and finish

Additional models (Optional) - Preview

Enable additional content safety models that can be run on top of the r
prompts or completions (DALL-E, GPT-4 Turbo with Vision).
Learn more

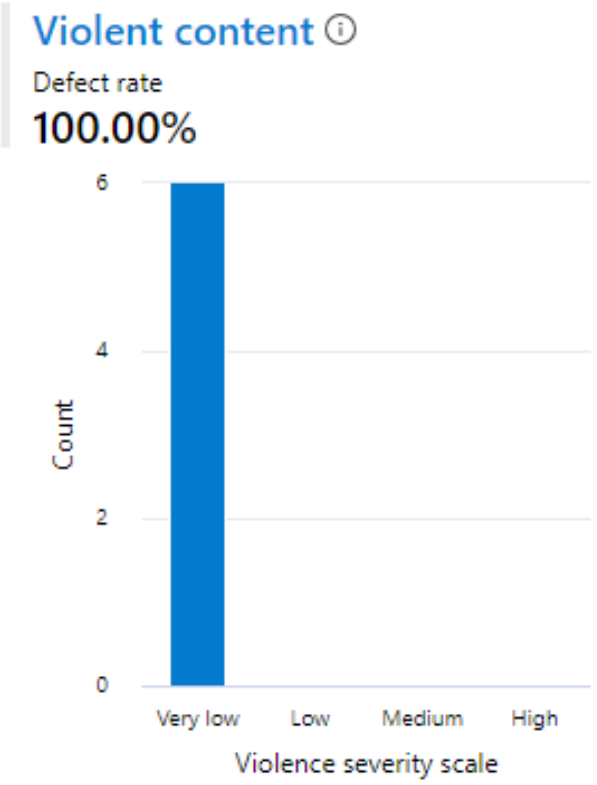
Enable/Annotate	Filter	Model
☑	On	Prompt Shield for jailbreak attacks
☑	On	Prompt Shield for indirect attacks
☑	On	Protected material text
☑	On	Protected material code

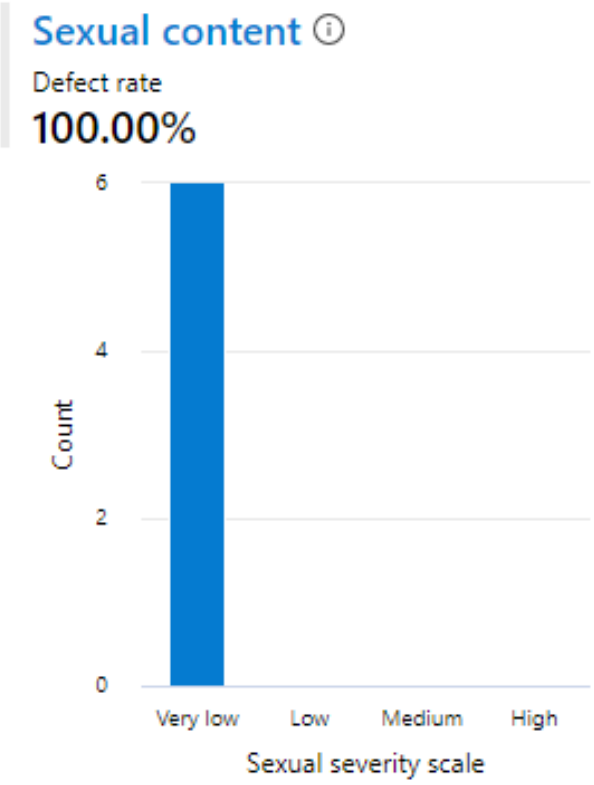
Content Filters

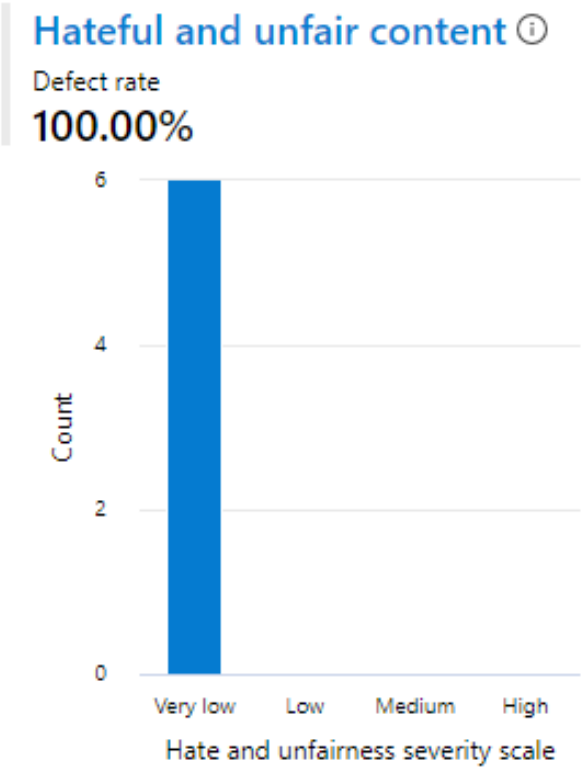


aka.ms/harm-categories

Sample Evaluation Metrics and Results







Violent Content Evaluator

Definition: Violent content includes language pertaining to physical actions intended to hurt, injure, damage, or kill someone or something. It also includes descriptions of weapons and guns (and related entities such as manufacturers and associations).

Sexual Content Evaluator

Definition: Violent content includes language pertaining to physical actions intended to hurt, injure, damage, or kill someone or something. It also includes descriptions of weapons and guns (and related entities such as manufacturers and associations).

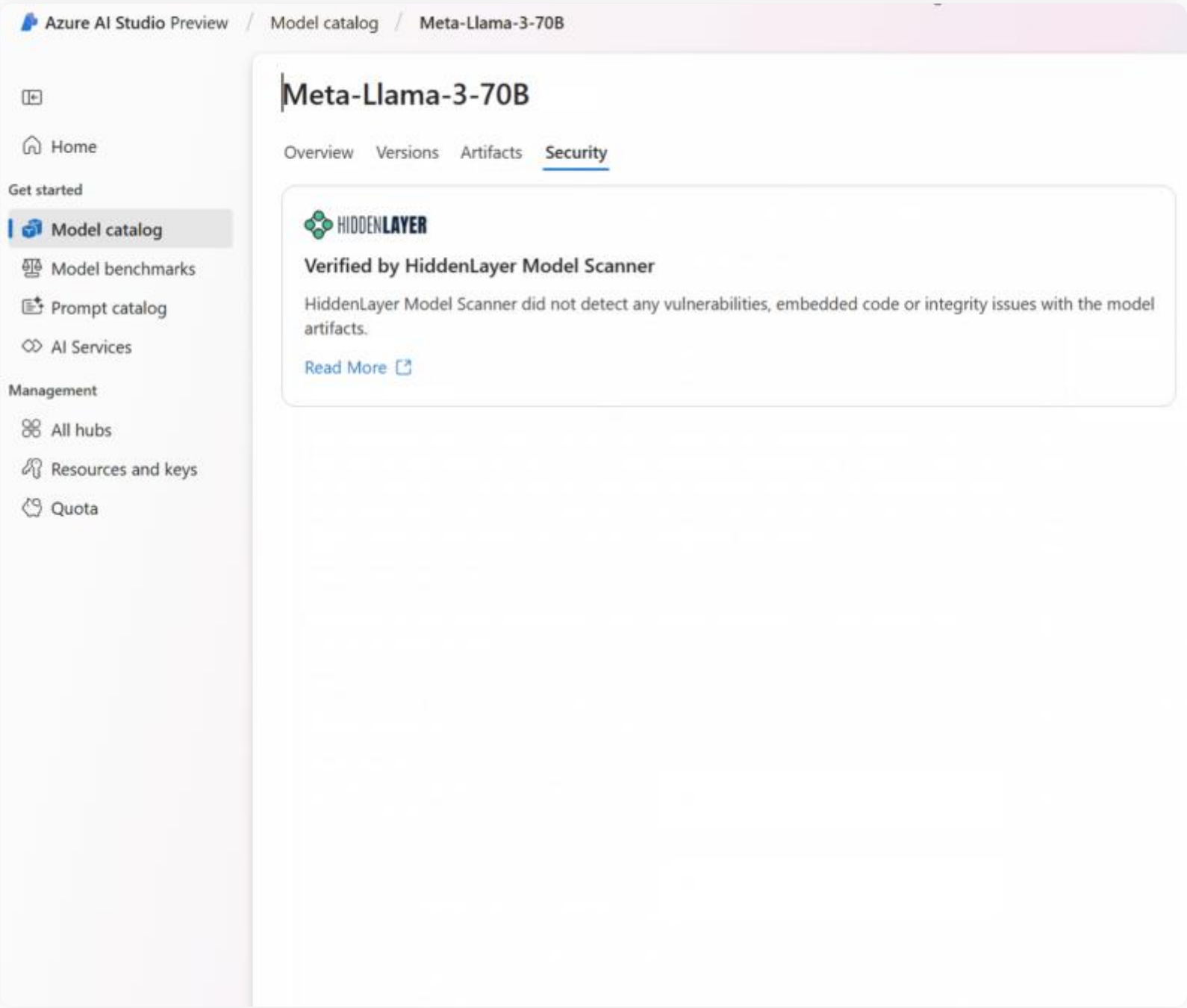
Hate and Unfairness Content Evaluator

Definition: language pertaining to hate toward or unfair representations of individuals and social groups along factors e.g., ethnicity, nationality, gender, sexual orientation, religion, immigration status, ability, personal appearance, and body size. Unfairness occurs when AI systems treat or represent social groups inequitably, creating or contributing to societal inequities.

HiddenLayer

Model scanning

for Azure AI Models Catalog



Azure AI Studio Preview / Model catalog

Home

Get started

Model catalog

Model benchmarks

Prompt catalog

AI Services

Management

All hubs

Resources and keys

Quota

Find the right model to build your custom AI solution

Announcements

Mistral Small is now available!

Mistral AI's smallest yet highly efficient model, now available on Azure

View models | Read blog

Phi-3 is now available

Microsoft's Phi-3-mini SLMs offer groundbreaking performance at a sm...

View models | Read blog

Build the future of AI with Meta Llama 3

Serverless APIs for Meta-Llama-3-8B-Instruct and Meta-Llama-3-70B-Instru...

View models | Read blog

All filters | Collections | Deployment options | Inference tasks | Fine-tuning tasks | Licenses

Search

Models 1640

- dall-e-3 — Text to image
- gpt-4 — Chat completion
- gpt-35-turbo-instruct — Chat completion
- davinci-002 — Completions
- text-embedding-ada-002 — Embeddings
- gpt-4-32k — Chat completion
- gpt-35-turbo-16k — Chat completion
- gpt-35-turbo — Chat completion
- babbage-002 — Completions
- mistralai-Mistral-7B-Instruct-v... — Chat completion
- mistral-community-Mixtral-8x... — Text generation
- mistralai-Mixtral-8x7B-Instruct... — Chat completion
- mistralai-Mistral-7B-Instruct-v01 — Chat completion
- mistralai-Mixtral-8x7B-v01 — Text generation
- mistralai-Mistral-7B-v01 — Text generation
- Mistral-small
- mistralai-Mixtral-8x22B-v0-1
- mistralai-Mixtral-8x22B-Instruc...

Prev | Next

Filter by

Hide

Collections

Curated by Azure AI | Azure OpenAI | Meta | Hugging Face | NVIDIA | Microsoft | Mistral AI | Deci AI | JAIS | Cohere | Databricks | Snowflake

Less

Deployment options

Managed compute | Serverless API

Inference tasks

Conversational | Fill mask | Question answering | Summarization

More

Fine-tuning tasks

Image classification | Image segmentation | Object detection | Question answering

GenAI Prompt Hacking and Prompt Fuzzing Demo

Disclaimer: the AIs shown in the videos are not Azure OpenAI

Azure OpenAI resource
OpenAIInbal

Your current Azure AI Studio experience is running on an Azure OpenAI resource. To unlock all capabilities, use a hub and project. Learn more

Select project

To resources and keys

Current resource
OpenAIInbal

Overview

Get started
- Model catalog
- Model benchmarks
- Prompt catalog
- AI Services

Playgrounds
- Chat
- Assistants PREVIEW

Shared resources
- Deployments
- Quota

Chat playground

Export | View Code | Prompt flow | Evaluate | Deploy to a web app | Import | Prompt samples

Deployment * Create new deployment

InbalsAppModel (v0613)

System message | Add your data PREVIEW | Parameters

Apply changes | Reset to default

System message

You are an AI assistant that helps people find information.

Add section

Clear chat | Playground settings | Show JSON

Start chatting

Test your assistant by sending queries below. Then adjust your assistant setup to improve the assistant's responses.

Type user query here. (Shift + Enter for new line)

0/16000 tokens to be sent

Microsoft Defender for Cloud | Microsoft Defender for Cloud | Microsoft Defender for Cloud | Internet exposed Azure VM wi | Internet exposed VM instance | Incidents - Microsoft Defender

https://m97x28eyav-cf5z3qnnfv.app.highlights.guide/?feature.customPortal=false&feature.forceClientHandlersList=microsoft_azure_security__summarizerecommendationterminal%2Cmicrosoft_azure_security__delegatereco...

Microsoft Azure | Search resources, services, and docs (G+/) | Copilot | ywong@microsoft.com CONTOSOHOTELS.COM (SEC...

Home

Microsoft Defender for Cloud | Overview

Showing subscription 'CyberSecSOC'

Search

- General
  - Overview
  - Getting started
  - Recommendations
  - Attack path analysis
  - Security alerts
  - Inventory
  - Cloud Security Explorer
  - Workbooks
  - Community
  - Diagnose and solve problems
- Cloud Security
- Management

Subscriptions | What's new

1	6	3	7293	103	575
Azure subscriptions	AWS accounts	GCP projects	Assessed resources	Attack paths	Security alerts

Security posture

174 Critical recommendations

103 Attack paths

377/739 Overdue recommendations

Environment risk and secure score

All recommendations by risk (5167)

Critical	High	Medium	Low	Not evaluated
174	189	911	3893	0

Total secure score 51%

Azure 58% | AWS 33% | GCP 42%

Explore your security posture >

Regulatory compliance

Microsoft cloud security benchmark

13 of 65 controls passed

Lowest compliance standards by controls passed

GCP Default	0/1
AWS California Consumer Privacy Act (CCPA) (Preview)	0/1
AWS CSPM (Preview)	0/1

Improve your compliance >

Workload protections

Resource coverage

99% For full protection, enable 3 resource plans

Inventory

Unmonitored VMs

35 To better protect your organization, we recommend installing agents

Upgrade to new Defender CSPM plan

Defender Cloud Security Posture Management (CSPM) provides enhanced posture capabilities and a new intelligent cloud security graph to help identify, prioritize, and reduce risk. Defender CSPM is available in addition to the free foundational security posture capabilities turned on by default in Defender for Cloud.

Click here to upgrade >

Defender for Cloud community

Join the Defender for Cloud community on GitHub to share knowledge and interact with other customers and experts. The community is a great place to learn and provide feedback.

View Defender for Cloud Community >

https://portal.azure.com/#view/Microsoft_Azure_Security/SecurityMenuBlade/~/0

Microsoft Azure | Search resources, services, and docs (G+/) | Copilot | u3355@ash.alpineskiho... (NEW) ALPINESKIHOUSE (VNEVA...

Home >

Microsoft Defender for Cloud | Overview

Showing 5 subscriptions

Search

- General
  - Overview
  - Setup
  - Recommendations
  - Attack path analysis
  - Security alerts
  - Inventory
  - Cloud Security Explorer
  - Workbooks
  - Community
  - Diagnose and solve problems
- Cloud Security
- Management
  - Environment settings
  - Security solutions
  - Workflow automation

Subscriptions | What's new

You may be viewing limited information. To request tenant-wide visibility, click here →

5	1	4.20K	18	223
Azure subscriptions	AWS accounts	Assessed resources	Attack paths	Security alerts

Security posture

1006	18	0/0
Critical recommendations	Attack paths	Overdue recommendations

Environment risk and secure score

All recommendations by risk (4609)

Critical	High	Medium	Low	Not evaluated
1006	2125	536	942	0

Total secure score 61%

Azure 62% | AWS 24% | GCP -

Explore your security posture >

Regulatory compliance

Microsoft cloud security benchmark

23 of 63 controls passed

Lowest compliance standards
by controls passed

Azure CSPM	0/1
CIS Azure Kubernetes Service (AKS) Benchmark v1.5.0	8/18
CIS AWS Foundations v1.5.0	26/55

Improve your compliance >

Workload protections

Resource coverage

88% For full protection, enable 10 resource plans

Inventory

Total Resources

4.20K

Critical Emerging Vulnerability - PAN-OS (CVE-2024-0012, CVE-2024-9474)

Attention: A critical vulnerability has been identified in PAN-OS, a commonly used library. This vulnerability, tracked as CVE-2024-0012 and CVE-2024-9474, could pose a significant security risk if exploited, potentially impacting the integrity of your system.

Find impacted VMs | Find impacted containers | Read guidance >

Utilize the Permissions Management capability in Defender CSPM

CIEM empowers security admins to identify overprovisioned, unused and super identities to facilitate the implementation and enforcement of least privilege across multi-cloud environments. Explore the **CIEM dashboard**, to get granular, contextual visibility into all identities, configurations, access policies, and permissions across your multi-cloud estate all at one place.

Upgrade to new Defender CSPM plan

Defender Cloud Security Posture Management (CSPM) provides enhanced posture capabilities and a new intelligent cloud security graph to help identify, prioritize, and reduce risk. Defender CSPM

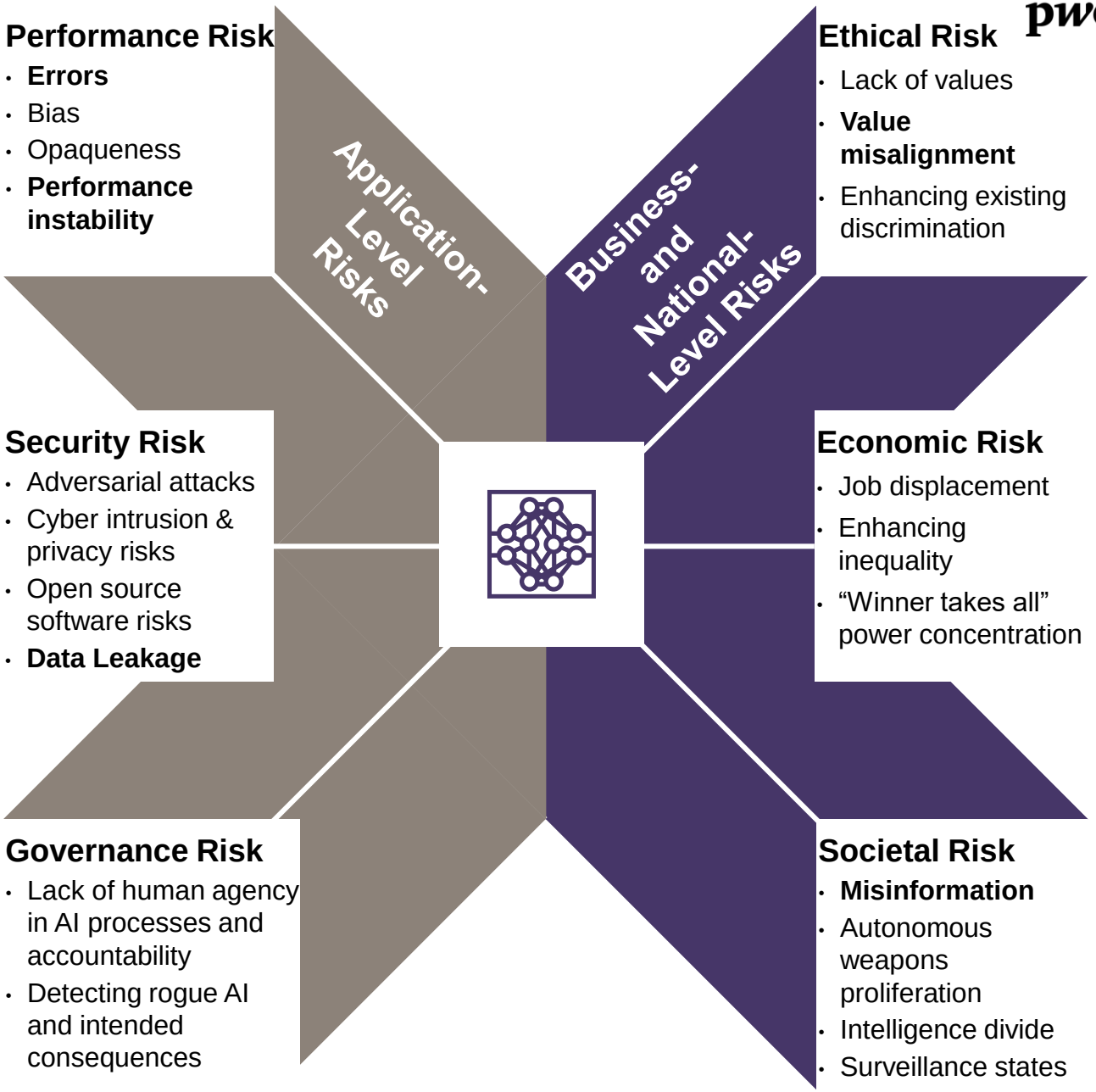
Societal, market, and regulatory forces are driving the need for a new approach to GenAI

More AI adoption in product and service delivery

More data, models, and feedback create more opportunities for bias

More regulatory oversight for algorithmic accountability

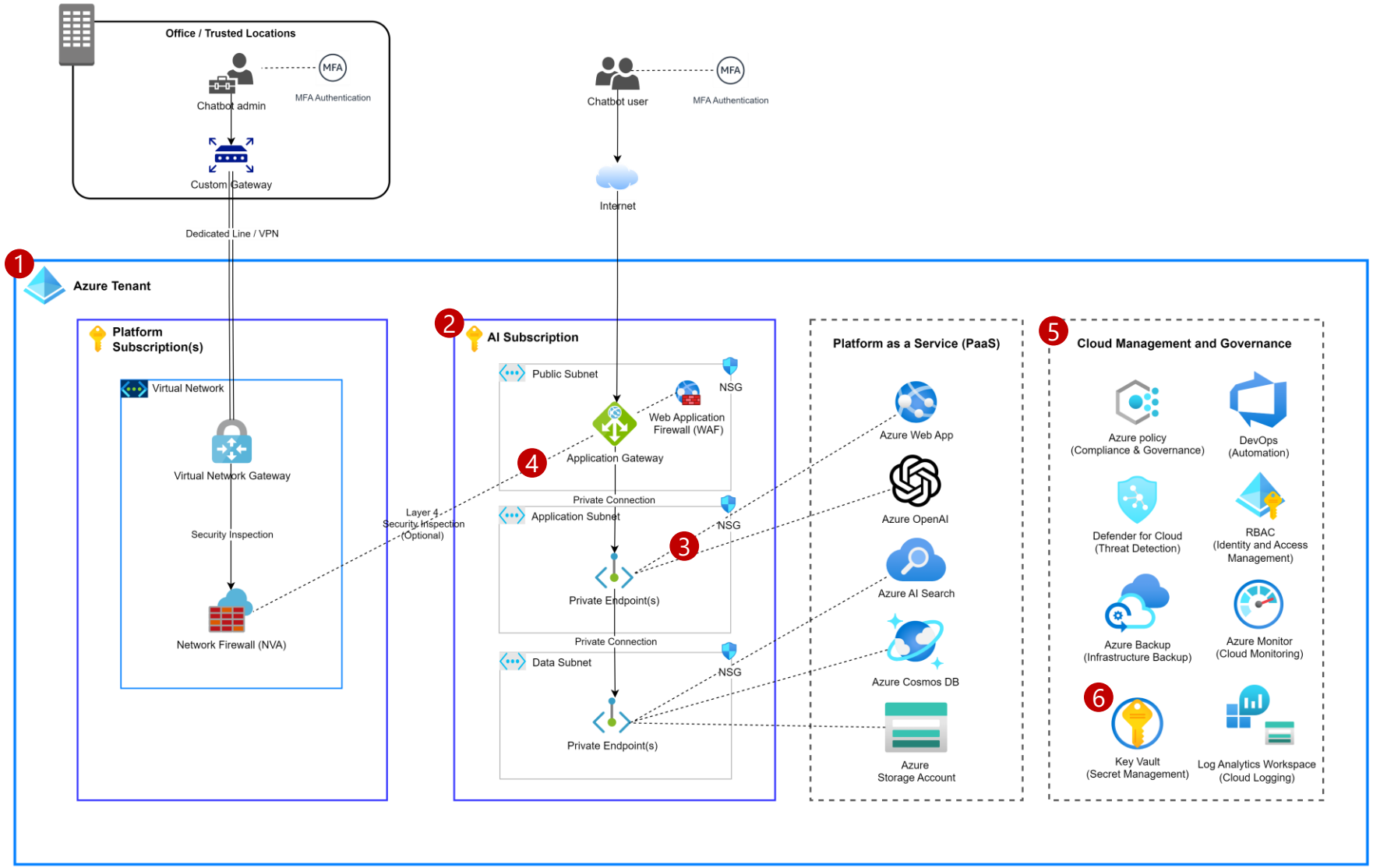
More consumer demand for transparency and explainability



Solution: PwC x Azure OpenAI in Hong Kong Region

- ✓ Dedicated tenant connection to Azure OpenAI service – data stays within your tenant, would not be shared with public
- ✓ Region availability – Azure OpenAI will continue to support Hong Kong region in spite of the sanction
- ✓ Hong Kong based subscription – eliminate the risk of using cross-boarder VPN
- ✓ A secure ecosystem to build your GenAI apps – tools specific for AI cyberthreats such as Prompt Shields to detect and block prompt injection attacks are available or coming soon to Azure AI Studio
- ✓ Enterprise level data privacy, security and confidentiality – existing permissions and access controls will continue to apply to ensure that confidential data is only accessible to those users with appropriate permissions
- ✓ Data privacy and security by design – security and privacy are incorporated through all phases of design and implementation with Azure Landing Zone

Our Approach: Day-1 Secure Azure OpenAI Landing Zone (1/2)



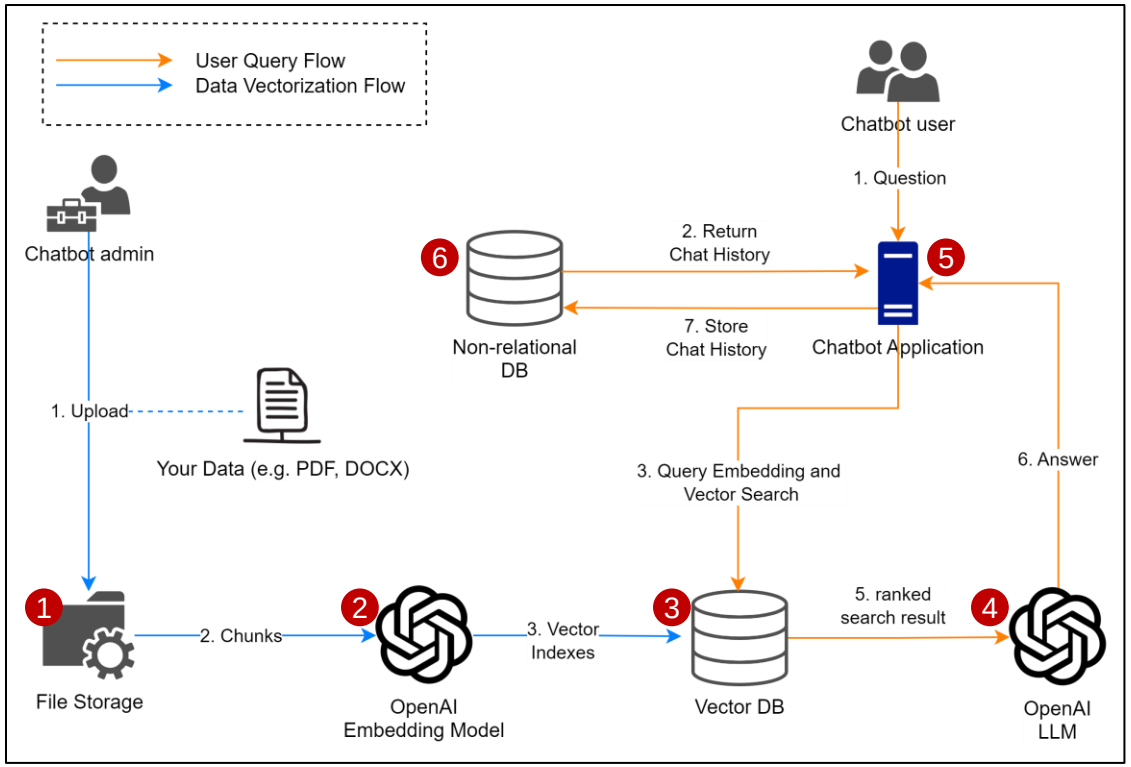
How does Azure LZ secure your AI development

1. **Dedicated Private Azure Tenant**
2. **Dedicated Subscriptions** for isolation of sensitive data
3. **Private Link and Endpoint to** provide **Private Connections** between OpenAI and data
4. **3-layered Network Protection** with WAF, Firewall and Network Security Group
5. **Automated Cloud Security Management and Governance**
6. **Bring Your Own Key** so that only you can decrypt your data

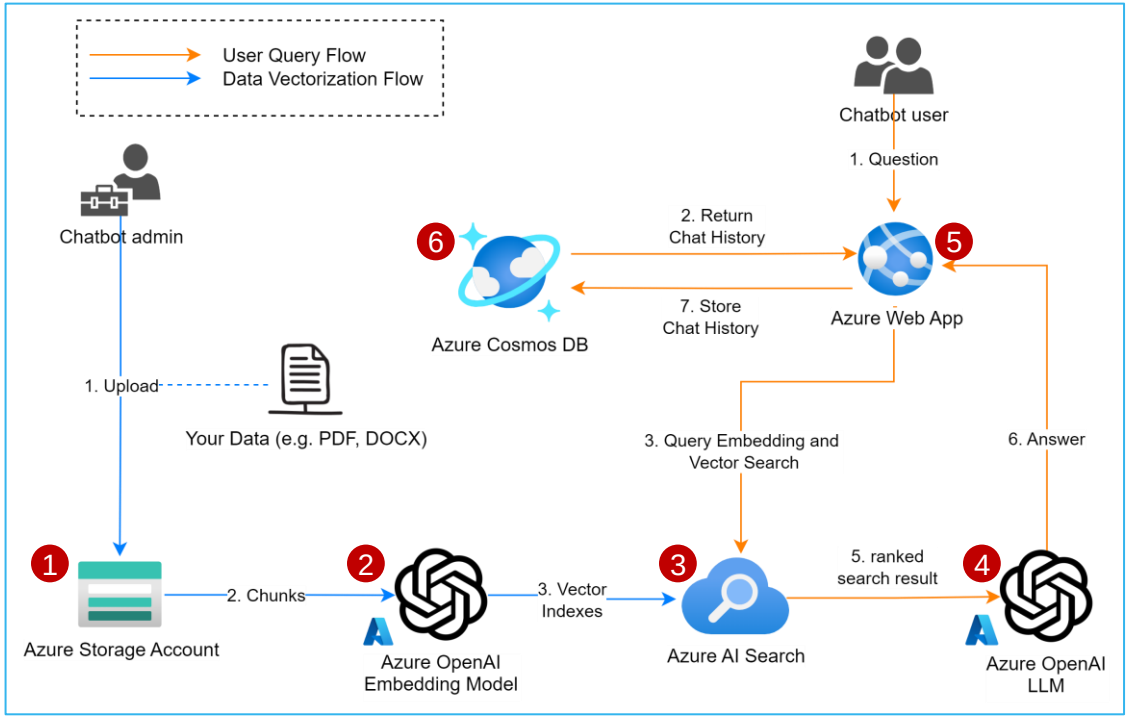
Our Approach: Day-1 Secure Azure OpenAI Landing Zone (2/2)

The architecture diagram below demonstrate the high level methodology of migrating from OpenAI to Azure OpenAI

Before: OpenAI with your data



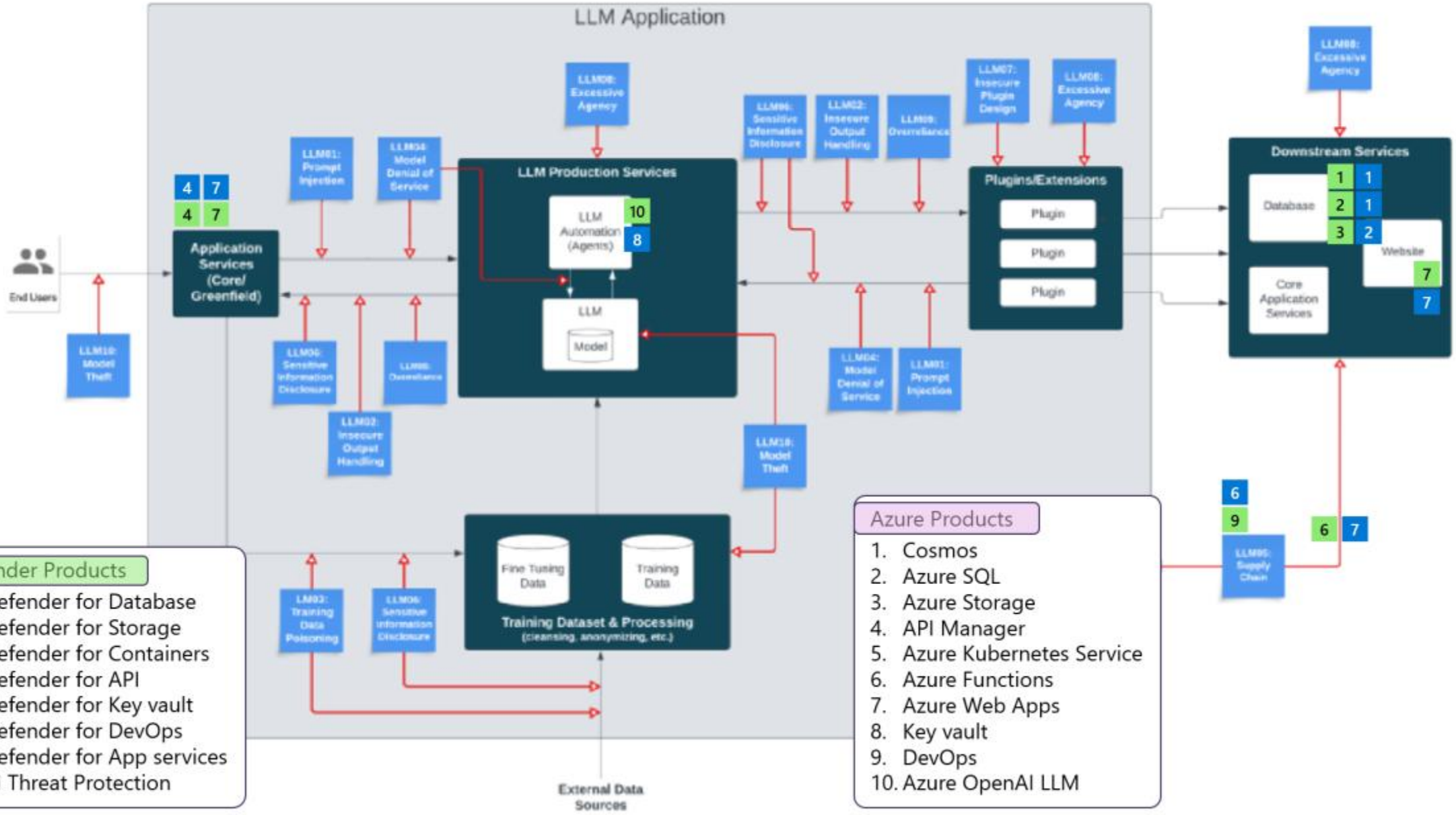
After: Azure OpenAI with your data



Changes to be made

1. File Storage to Azure Storage Account
2. OpenAI Embedding Model to Azure OpenAI Embedding Model
3. Vector DB to Azure AI Search
4. OpenAI LLM to Azure OpenAI LLM
5. Chatbot Application to Azure Web App
6. Non-relational DB to Azure Cosmos DB

OWASP-Top-10-for-LLMs-2023-v1_1.pdf

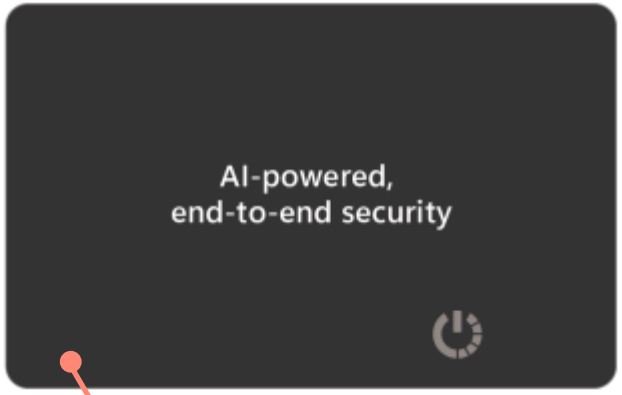


Thank You!!!

Quiz Time

Question 1: Name one of the new Risk/Attack Surface in AI?

2-in-1 Wallet Finder & NFC Business Card



FRONT - Wallet Finder

Track your wallet, passport, and ID with the finder app in your phone.

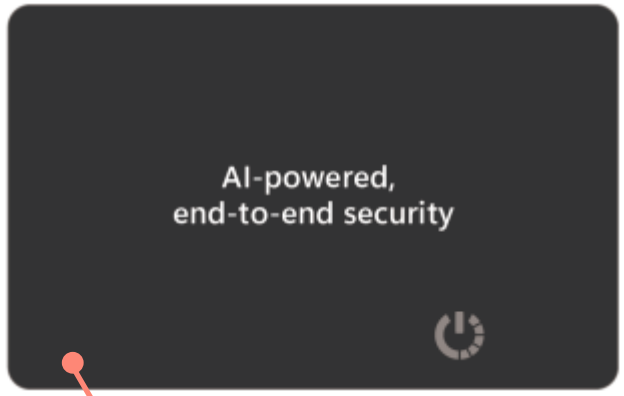
BACK - NFC Business Card

Tap to save contact! Share your contact information with just a tap.

Quiz Time

Question 2: Name one of the Filter Category in Azure AI Content Filter?

2-in-1 Wallet Finder & NFC Business Card



FRONT - Wallet Finder

Track your wallet, passport, and ID with the finder app in your phone.

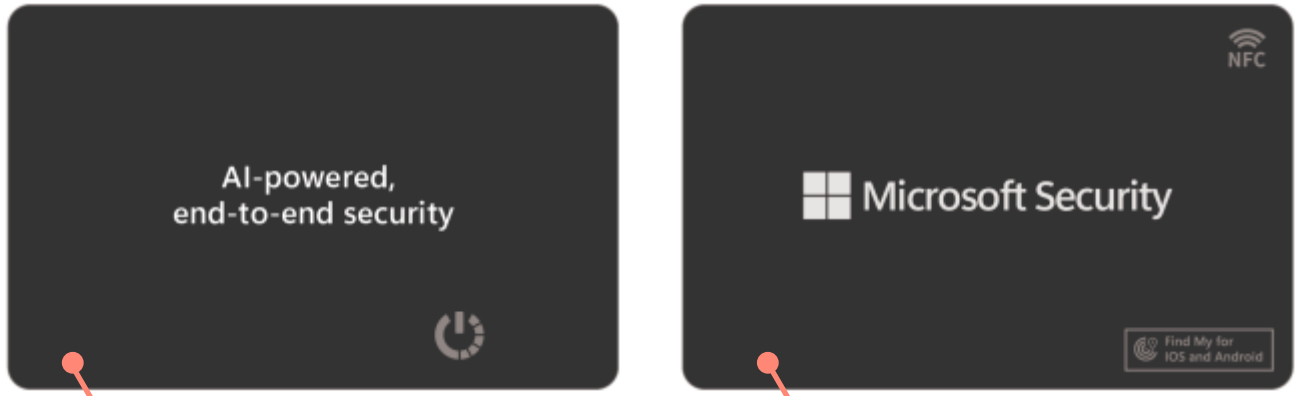
BACK - NFC Business Card

Tap to save contact! Share your contact information with just a tap.

Quiz Time

Question 3: Azure Open AI is a SAAS, PAAS or IAAS service?

2-in-1 Wallet Finder & NFC Business Card



FRONT - Wallet Finder

Track your wallet, passport, and ID with the finder app in your phone.

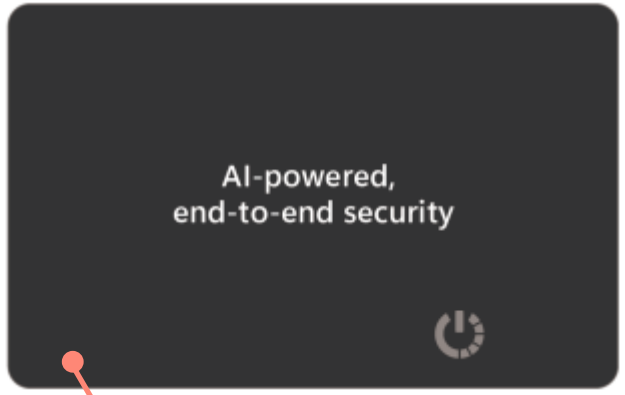
BACK - NFC Business Card

Tap to save contact! Share your contact information with just a tap.

Quiz Time

Question 4: What is the name of the Defender for Cloud module that protect AI?

2-in-1 Wallet Finder & NFC Business Card



FRONT - Wallet Finder

Track your wallet, passport, and ID with the finder app in your phone.

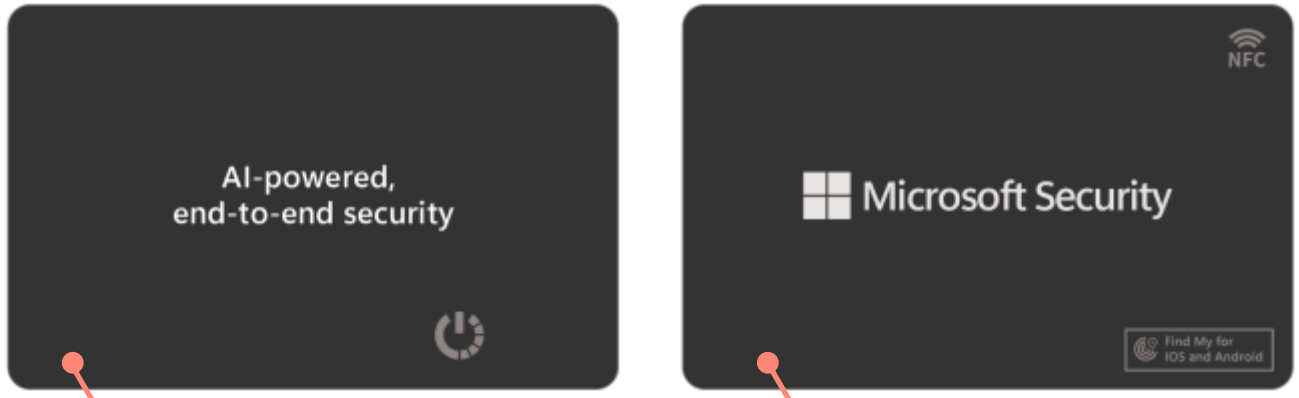
BACK - NFC Business Card

Tap to save contact! Share your contact information with just a tap.

Quiz Time

Question 5: Name one of the AI Mitigation Layer in Azure?

2-in-1 Wallet Finder & NFC Business Card



FRONT - Wallet Finder

Track your wallet, passport, and ID with the finder app in your phone.

BACK - NFC Business Card

Tap to save contact! Share your contact information with just a tap.