

AI Security Blueprint: How Microsoft Security Tools Work Together to Mitigate AI Risk



Ben Wilcox

Chief Technology Officer &
Chief Information Security Officer
bwilcox@proarch.com
[LinkedIn.com/in/ben-wilcox](https://www.linkedin.com/in/ben-wilcox)



Jonathan Atlikhani

Senior Security Consultant



Founded
in 2006



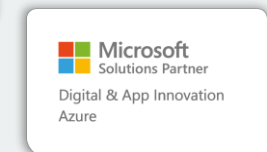
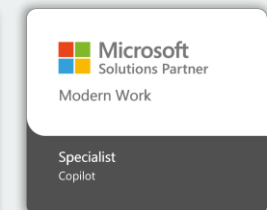
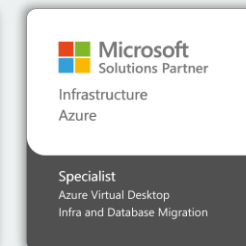
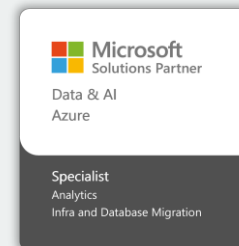
HQ in Atlanta | Offices in
Rochester, NY, UK, & India



Over 350 clients
& 450 employees

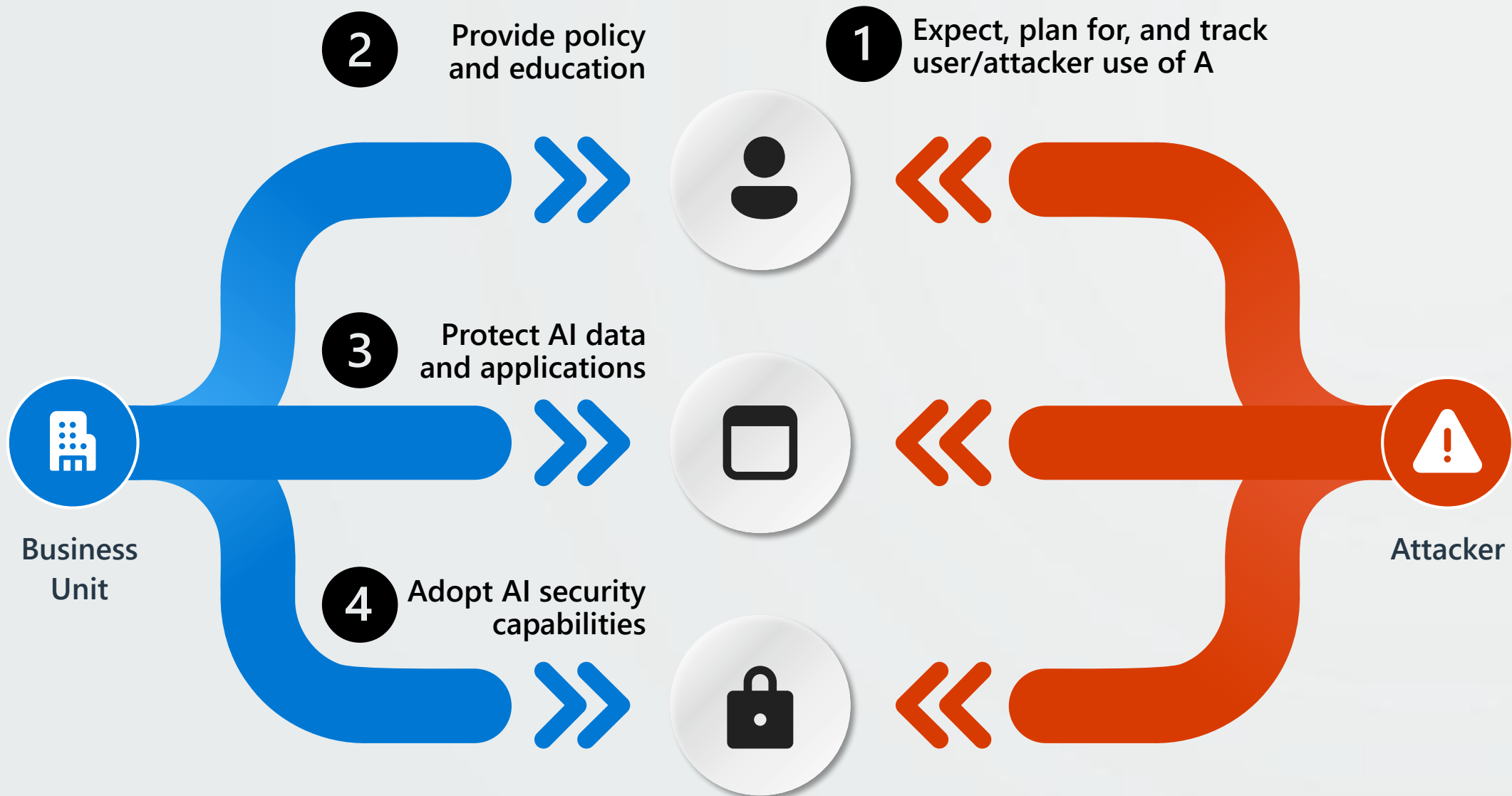


Top Microsoft
Solutions Partner



AI “just” another layer you need to protect.

AI Requires Security Leaders to Act on Multiple Fronts



“

We don't allow AI
in our organization.



AI adoption is here.

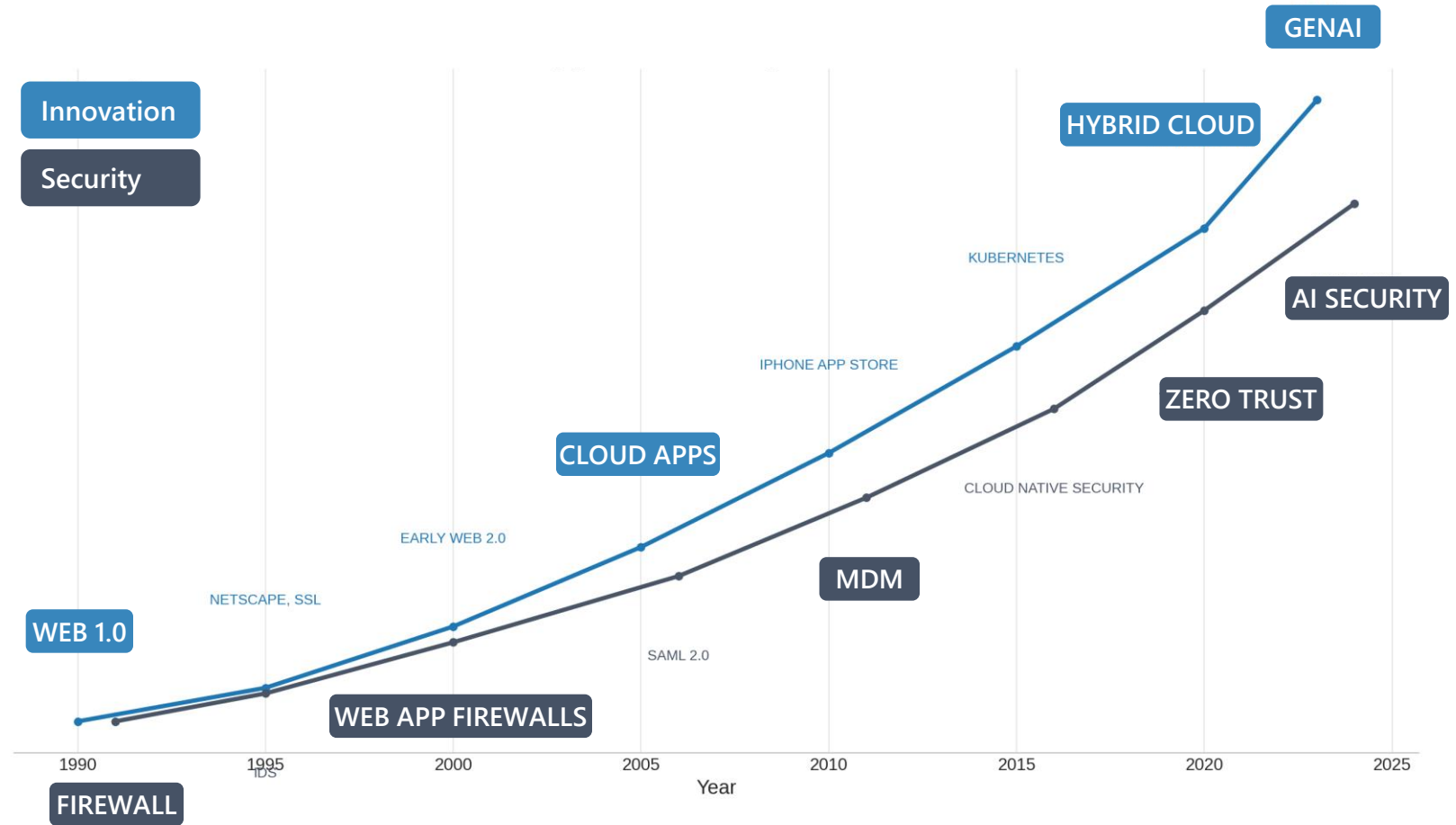
78% of organizations use AI in at least one business function

75% of knowledge workers use GenAI at work

2x increase
AI usage has nearly doubled in 6 months

79% of executives say AI agents are already being adopted

Security is almost keeping pace.



Healthcare SaaS start-up specializes in advanced analytics and AI-driven insights for oncology and clinical data.

BUSINESS SITUATION

Needed a robust, compliant, and scalable infrastructure to:

- Support sensitive PHI and PII data
- Secure remote operations
- Enable rapid innovation

AI SECURITY SOLUTION

- Architected and deployed Managed Detection and Response to their unique requirements
- 24x7 SOC monitoring and response
- End-to-end security, compliance, vulnerability management
- Leveraging Azure, Sentinel, and Defender ecosystem

459 Alerts Responded to in Last Year
(13 True Positive)

452 / 522
HIPAA & HITRUST
Security Controls Met

9 Azure
Subscriptions
Protected

3 Pillars of AI Security

AI SECURITY BLUEPRINT

3 Pillars for Secure Enterprise Adoption

1

VISIBILITY & GOVERNANCE

Discover, classify, and control AI usage

- AI Usage Layer
- User interactions, prompts

2

DATA PRIVACY & SECURITY CONTROLS

Protect sensitive data in AI workflows

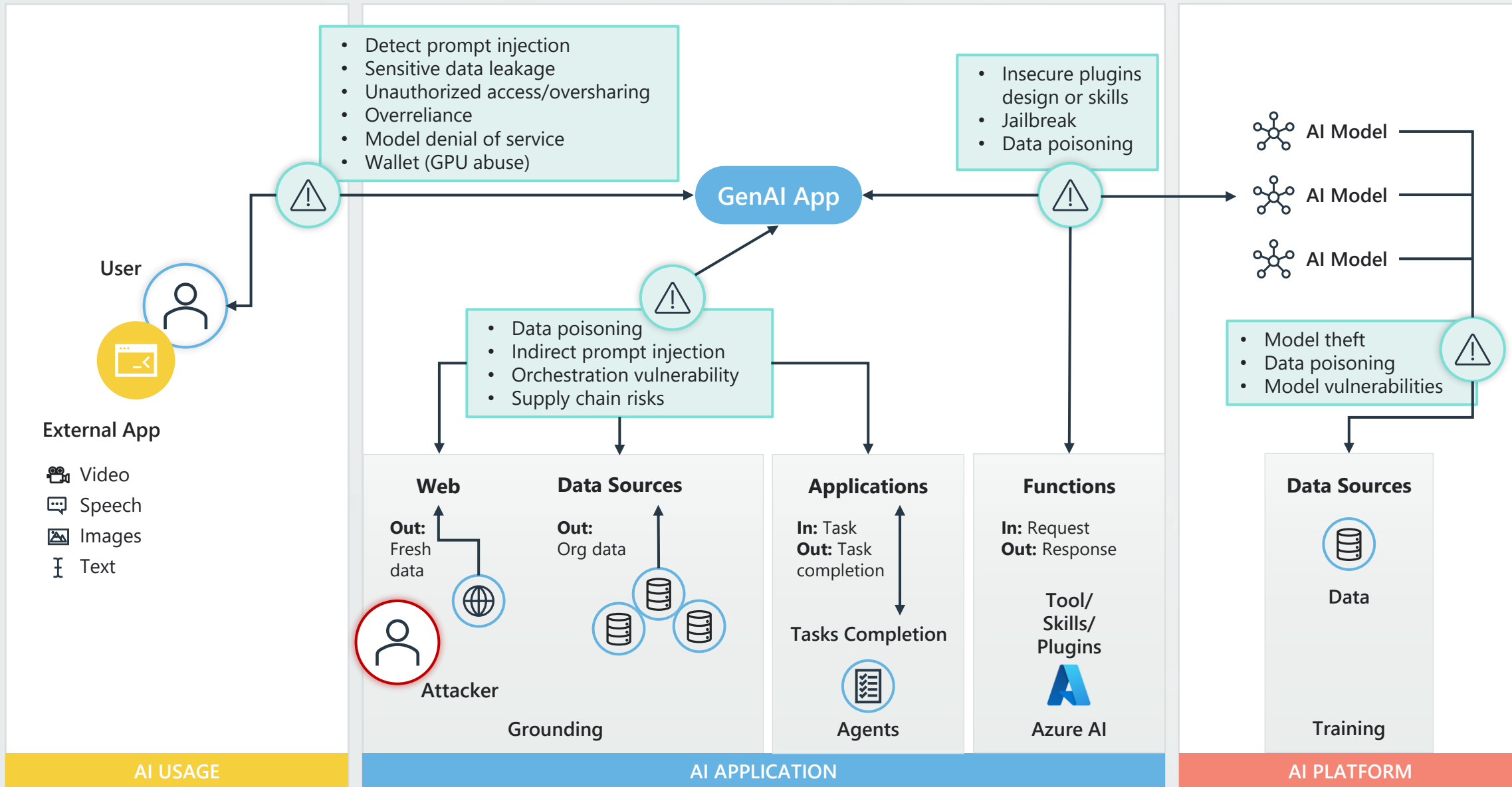
- AI Application Layer
- Apps, agents, plugins

3

ADAPTIVE SECURITY & INNOVATION

Evolve defenses with AI threat landscape

- Models, infrastructure



AI Usage Layer

User Interactions & Prompt Security

KEY THREATS

- Direct Prompt Injection
- Sensitive Data Leakage or Exposure
- Unauthorized Access/Oversharing
- Overreliance
- Model Denial of Service
- Inadequate Data Governance

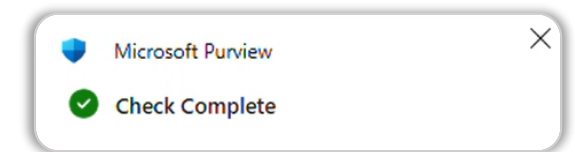
MICROSOFT SECURITY CONTROLS

Microsoft Defender for Cloud Apps

- Shadow AI discovery and visibility
- Session Policies for real-time monitoring
- OAuth app governance

Microsoft Entra ID

- Risk-based conditional access
- Anomalous usage detection
- Identity threat protection



Microsoft Purview

- **DLP:** prompt content inspection, sensitive data type detection, policy enforcement at data egress
- **Data Security:** discovery, classification, labeling, lifecycle management, compliance monitoring
- **DSPM for AI:** Identification of AI usage and risky behavior

AI Governance

Policy Requirements

Data classification for AI tool usage (public, internal, confidential, restricted)

Security Requirements for AI

Approved AI vendor list with risk management/security assessments

User training on AI threat recognition and safe usage practices

Monitoring and Compliance Practices

Governance Benefits

Security Budget Buy-in: Sharing risks with the AI steering committee will help you secure budget

Risk reduction: 63% of breached organizations lack AI governance policies

Compliance readiness: policies for AI tool usage, data classification, and vendor management

Cost avoidance: Global average breach cost is \$4.44M, with shadow AI adding costs

Return on Investment

AI governance prevents \$670K in additional costs associated with shadow AI

Reduces overall breach risk in an environment where 97% of AI-breached organizations lacked proper access controls.

Defender for Cloud Apps

MAKE SURE THESE FEATURES ARE ENABLED!

- App Connectors (M365 & Azure at minimum)
- Defender for Endpoint integration
- App Governance

Deepseek
Sanction
Unsanction
Tag app
...

Info

Unsanctioned app

5

DeepSeek is an artificial intelligence software company that focuses on developing open-source large language models.

General

Security

Compliance

Legal

Score

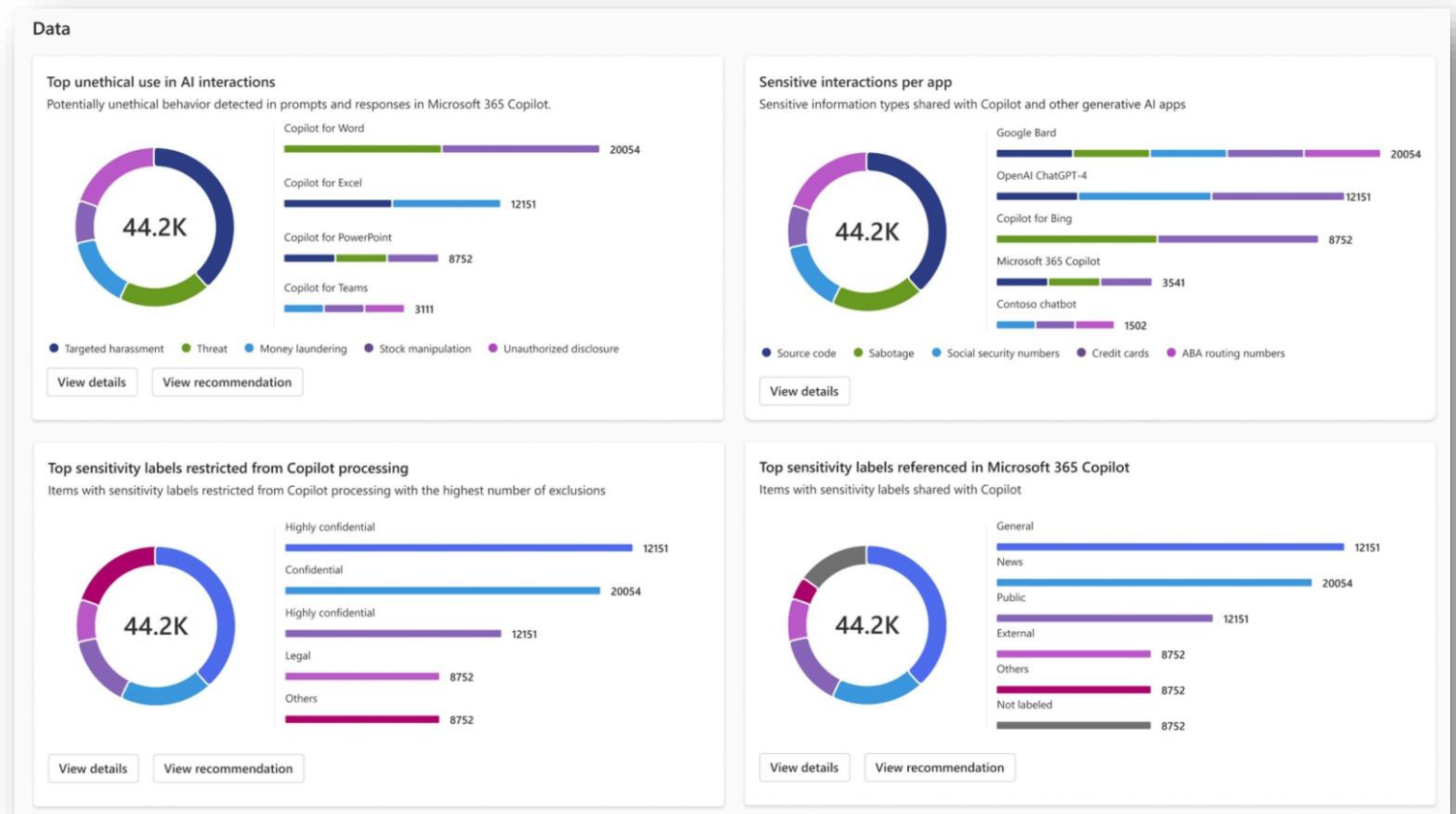
4

App	Risk score	Tags	Traffic	Upload	Trans...	Users	IP ad...	Devices	Last s...	Actions
Final Round Generative...	3	UNSANCTIONED	60 KB	—	1	1	1	1	Sep 18, 20...	
Blackbox AI Generative...	4	MONITORED	3 MB	—	1	1	1	1	Sep 19, 20...	
NoteGPT Generative...	4	MONITORED	1 MB	—	1	1	1	1	Sep 27, 20...	
Monica Generative...	5	MONITORED	32 KB	—	1	1	1	1	Sep 29, 20...	
Fotor Generative...	5	MONITORED	24 MB	135 KB	4	1	1	1	Sep 27, 20...	
Groq Generative...	5	MONITORED	1 MB	219 KB	2	1	1	1	Sep 12, 20...	
Wondershare Generative...	5	MONITORED	17 MB	—	1	1	1	1	Sep 20, 20...	
BeyondWords Generative...	5	MONITORED	1 MB	—	6	3	5	3	Oct 3, 2025	
UX Pilot Generative...	6	MONITORED	3 MB	259 KB	3	1	2	1	Sep 25, 20...	

DSPM for AI

COMPLETE SETUP TASKS!

- Activate Microsoft Purview Audit
- Install Purview browser extension
- Onboard devices to Purview (or MDE)
- Enable monitoring policies





AI Application Layer

Apps, Agents & Plugins

KEY THREATS

- Insecure Plugins Design or Skills
- Jailbreak
- Data Poisoning
- Indirect Prompt Injection
- Orchestration Vulnerability

MICROSOFT SECURITY CONTROLS

Microsoft Defender for Endpoint

- Application control and sandboxing
- Behavioral monitoring
- Vulnerability management

App Governance

- AI app discovery and assessment
- Permission analysis
- Usage analytics and alerts

Entra ID Agent ID

- Secure agent authentication
- Managed identities for AI workloads
- Fine-grained access control

Defender for Cloud: AI Threat Protection

- AI-specific threat detection
- Runtime protection for AI workloads
- Security posture management

1

VISIBILITY &
GOVERNANCE

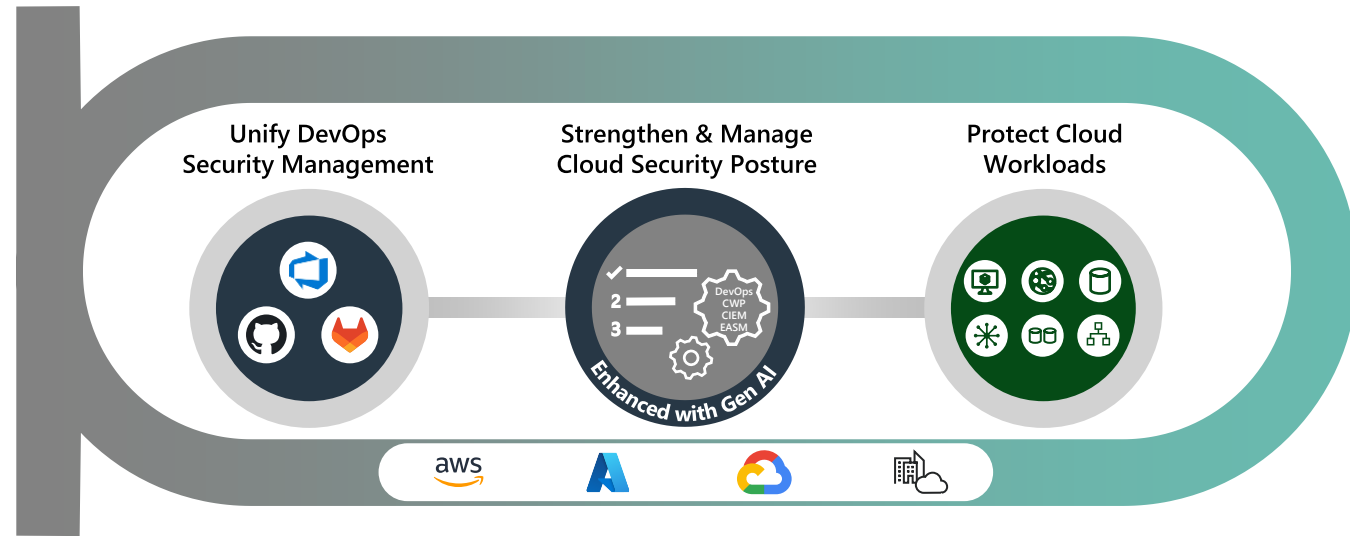
2

DATA PRIVACY &
SECURITY CONTROLS

3

ADAPTIVE SECURITY
& INNOVATION

Microsoft Defender for Cloud



INTEGRATED INSIGHTS

Application
Security

External Attack Surface
& Attack Paths

Cloud Entitlement &
Access Permissions

Sensitive Data
Protection

Network
Segmentation

SECOPS INTEGRATIONS

SIEM

Native-XDR

Workflow Automation

Ticketing & Collaboration



AI Platform Layer

Models & Infrastructure

KEY THREATS

- Model Theft
- Data Poisoning
- Model Vulnerabilities
- Infrastructure Misconfiguration
- Supply Chain Risks

MICROSOFT SECURITY CONTROLS

Microsoft Defender for Cloud

- Cloud Security Posture Management
- Vulnerability assessment
- Compliance monitoring

Azure Confidential Computing

- Hardware-based data encryption
- Secure enclaves for model interference
- Protected model weights

Azure AI Foundry

- Content filters (hate, violence, etc.)
- Custom blocklists and allowed terms
- Security recommendations and best practices
- Prompt shields and grounding detection

1

VISIBILITY &
GOVERNANCE

2

DATA PRIVACY &
SECURITY CONTROLS

3

ADAPTIVE SECURITY
& INNOVATION

Defender for Cloud: Proactive Security

Microsoft Azure

Home >

contoso-openAI-B

Azure OpenAI

Search

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource Management

Security

Microsoft Defender for Cloud

Monitoring

Automation

Help

Internet exposed AI application has access to sensitive data (Preview)

Medium

0 Active Recommendations

12:00:00 Freshness interval

Attack path

Remediation

mdc-demo-w2019

Entry point

xdrstaticstorage

Target

Internet

mdc-demo-w2019 Virtual machine

can authenticate as

221f60a3-2744-4381... Managed identity

has permissions to

xdrstaticstorage Storage account

Azure AI Services resources should restrict network access

Cognitive Services should use private link

[Enable if required] Cognitive Services accounts should enable data encryption with a customer-managed key (CMK)

Azure AI Services resources should have key access disabled (disable local authentication)

Audit diagnostic setting for selected resource types

Showing 1 - 5 of 6 results.

View additional recommendations in Defender for Cloud >

Security incidents and alerts

Defender for Cloud uses advanced analytics and global threat intelligence to alert you to malicious activity. Alerts displayed below are from the past 21 days.

Alert title	Count	Detected	State	Activity time	Severity
A jailbreak attempt on your Azure Open AI model deployment was blocked b...	5	Microsoft	Active	05/16/24	Medium
Sensitive data exposure by your Azure Open AI model deployment was detec...	1	Microsoft	Active	05/16/24	Medium

Microsoft Defender for Cloud | Recommendations

Showing subscription 'CyberSecSOC'

Search

Refresh

Download CSV report

Open query

Governance report

Guides & Feedback

Switch to classic view

Analyze with

General

Overview

Getting started

Recommendations

Attack path analysis

Security alerts

Inventory

Cloud Security Explorer

Workbooks

Community

Diagnose and solve problems

Cloud Security

Secure score and custom recommendations can still be found in the classic view >

Scope: Azure subscriptions 1 AWS accounts 6 GCP projects 3 GitHub connectors 3 AzureDevOps connectors 1 GitLab connectors 1

Defender for CSPM

Recommendations by risk

Prioritized by resource level risk factors and context. Learn more

Risk based recommendations

172 Critical High (193) Medium (921) Low (4.5K)

Foundational CSPM

Recommendations

0 No risk calculated

Search by title / resource

Status == 3 selected

Risk factors == All

Risk level == All

Add filter

Group by title

Title	Affected resource	Risk level	Risk factors	Attack paths	Owner	Status	Insights
Machines should have vulnerability findings resolved	contoso-dsvm	Critical	Contains Verified Sec...	+5 9		Unassigned	
Machines should have vulnerability findings resolved	mdc-demo-w2019	Critical	Exposure to the Inter...	+3 7	Lior Aniv	Overdue	

Defender for Cloud: Alert & Remediation

Security alert

2516865245034355829_cb4b2c24-51a7-4d12-b85c-3654b4b5f099

A Jailbreak attempt on your Azure Open AI model deployment was blocked by Prompt Shields (Preview)

Medium

Active

07/22/25, 06:44 AM

Alert description

There were 3 blocked attempts of a Jailbreak attack on model deployment Contoso-35-SI on your Azure Open AI resource contoso-openAI-B.

A Jailbreak attack is also known as User Prompt Injection Attack (UPIA). It occurs when a malicious user manipulates the system prompt, and its purpose is to bypass a generative AI's large language model's safeguards in order to exploit sensitive data stores or to interact with privileged functions. Learn more at <https://aka.ms/RAI/jailbreak>.

The attempts on your model deployment were using direct prompt injection techniques and were blocked by Azure Responsible AI Content Filtering. The prompts were not completed. However, to block further malicious attempts by the suspected user and to handle possible undetected prompt injections, we recommend taking immediate action:

1. Investigate the user who created the attempts by looking at the source application's history and consider removing their access.

2. Consider there may have been undetected successful prompt injections – investigate to validate no sensitive data was revealed by the model, and that no data poisoning took place.

To get detailed information on the prompt injection attempts, refer to the 'Supporting evidence events' section in the Azure Portal.

Affected resource

contoso-openAI-B

Azure AI Service

CyberSecSOC

Alert details

Take action

Inspect resource context

Start with examining the resource logs around the time of the alert.

Open logs

Mitigate the threat

If the user who accessed the application and entered the prompt seems to be malicious, consider removing their access.

Validate your model deployment has minimal access to sensitive data and to privileged actions, and that you have strong safeguards in place so it doesn't allow unprivileged users access these data sources and actions.

You have 46 more alerts on the affected resource. [View all >>](#)

Prevent future attacks

Your top 3 active security recommendations on contoso-openAI-B:

Medium

Azure AI Services resources should restrict network access

Medium

Cognitive Services should use private link

Medium

Azure AI Services resources should have key access disabled (disable local authentication)

Solving security recommendations can prevent future attacks by reducing attack surface.

[View all 6 recommendations >>](#)

Trigger automated response

Suppress similar alerts

Configure email notification settings

Automatic remediation script content (Preview)

```
1 #The following script will disable local authentication to your Azure AI instance To disable local authentication to your Azure AI instance:
2 az resource update --ids /subscriptions/d1d8779d-38d7-4f06-91db-9cbc8de0176f/resourcegroups/contoso_clouddatasecurity_rg/providers/microsoft.cognitiveservices/accounts/contoso-openai-b
```

AI Foundry Security

Guardrails + Controls

Content filters work alongside core models. Create filters and assign to deployments to manage content by c
[Learn more about content filters and blocklists](#)

Overview Try it out Content filters **Blocklists** PREVIEW Security recommendations PREVIEW

+ Create blocklist Edit Delete Refresh

Find in page

☐	Name	Created at
☐	Blocklist-ProArch-Standard	Aug 8, 2025 11:37 AM

Guardrails + Controls

Overview Try it out Content filters Blocklists PREVIEW **Security recommendations** PREVIEW

You have 0 security alerts. Monitor and review your alerts in Risks + alerts.

Recommendations uses Defender for Cloud to monitor your Azure resource configurations for potential security vulnerabilities and to suggest remediation actions. [Learn more about Microsoft Defender for Cloud](#)

Search

Recommendation	Affected resource	Resource type	Risk level
Azure AI Services resources should have key access disabled (disable local authentication)	pause2rgdvaoa02	Azure AI Services	Critical
Azure AI Services resources should restrict network access	pause2rgdvaoa02	Azure AI Services	High
Azure AI Services resources should use Azure Private Link	pause2rgdvaoa02	Azure AI Services	High
D diagnostic logs in Azure AI services resources should be enabled	pause2rgdvaoa02	Azure AI Services	Low
Audit diagnostic setting for selected resource types	pause2rgdvaoa02	Azure AI Services	Low

Edit filters to allow or block specific types of content

- Basic information
- Input filter**
- Output filter
- Connection
- Review

Set input filter

Select the level of content severity to block for each category. Note the lowest setting blocks only the worst severity, and vice versa.
[Learn more about categories and blocking threshold levels](#)

Category	Media	Action	Blocking threshold level
Violence	☐ Text ☐ Image	Annotate and block	Medium blocking Blocks both moderate and highly severe unwanted content
Hate	☐ Text ☐ Image	Annotate and block	Medium blocking Blocks both moderate and highly severe unwanted content
Sexual	☐ Text ☐ Image	Annotate and block	Medium blocking Blocks both moderate and highly severe unwanted content
Self-harm	☐ Text ☐ Image	Annotate and block	Medium blocking Blocks both moderate and highly severe unwanted content
Prompt shields for jailbreak attacks	☐ Text	Annotate and block	Jailbreak attacks will be blocked
Prompt shields for indirect attacks	☐ Text	Annotate and block	Indirect attacks will be blocked

Blocklist (Preview)

uilt-in or customized blocklist *

ity,Blocklist-ProArch-Standard

Where Your Organization Needs Security Coverage

Endpoints

Servers: Linux, Windows

Workstations: Linux, Windows, MacOS

Mobile Devices: iOS, Android

Identity

On-Premises Active Directory

Cloud Azure AD/Entra ID

AI Application Identities

AI Agent Identities

Collaboration

Exchange Online

Microsoft Teams

Microsoft SharePoint

Microsoft OneDrive

Cloud Apps

Microsoft 365 Apps

3rd party Cloud Apps

Cloud & AI

Microsoft Azure

Amazon Web Services

Google Cloud Platform

Microsoft Copilot

Azure OpenAI Service

Azure AI Foundry

Azure Machine Learning

Amazon Bedrock

Google Vertex AI

Azure Confidential Computing

Confidential VM Instances

Confidential AI Training and Inference

Azure Disk Encryption

Azure Key Vault

People & Processes

Security Awareness Training

Processes and Security

Playbooks

Tabletop Exercises

SIEM Telemetry

Endpoints

Firewalls

Switches

Routers

Web traffic

Databases

Cloud logs

Access/Activity logs

AI Prompts

300+ Connectors Available

XDR

Multi-layer

Security Solutions

Alerts

Telemetry

Data

1st party sources

3rd party sources

Data

Sensitive Data

M365/Exchange Online Data

On-Premise Data

Endpoint Data

Cloud App Data

Multi-cloud Data

Database Data

File services Data

IoT/OT

Control Systems

HMI

Field Devices

PLCs

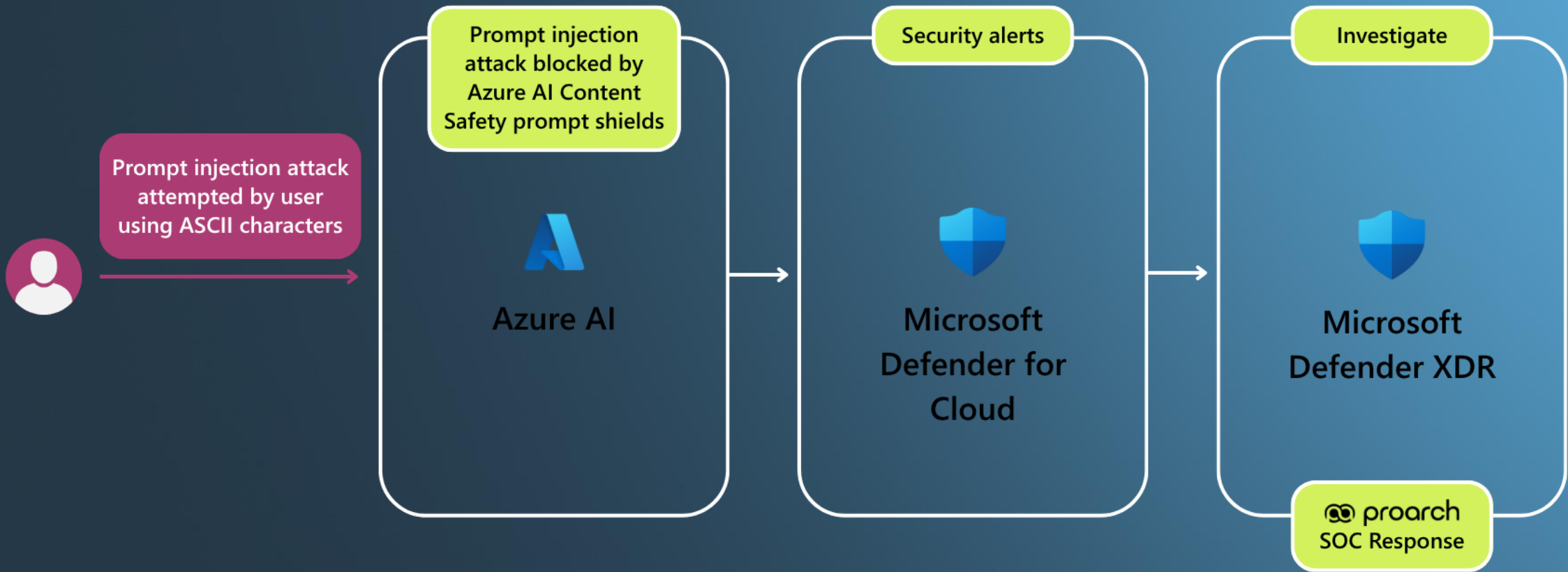
Sensor & Measurement

SCADA

Smart Grids

Industrial Robots

Demo: Microsoft tools block and respond to AI-related threat



Part 1

Staging the Gen AI App Attack

“Take home” Next Steps



AI Security Responsibility Matrix

LAYER	MICROSOFT RESPONSIBILITY	YOUR RESPONSIBILITY	TAKEAWAY
INFRASTRUCTURE	Physical DCs, compute, storage, networking, hypervisor security, patching managed OS	VM config, firewalls, NSGs, conditional access, compliance monitoring	Same as cloud today. Microsoft runs the data centers, you secure your configs.
AI PLATFORM	Secure APIs, model lifecycle, baseline safety filters, AAD/RBAC, managed identities	Model selection, fine-tuning, safety filters, governance on approved models	Microsoft secures the platform; you govern how AI services are configured and used.
DATA	Encryption (rest/transit), tenant isolation, regulatory compliance (GDPR, HIPAA, FedRAMP)	Data quality, bias mitigation, masking PII, DLP & Purview classification, compliance checks	Your largest risk. Data governance and compliance are your responsibility.
APPLICATIONS & AGENTS	Secure APIs, managed connectors, baseline Copilot guardrails	Configure agent access, restrict dangerous automations, monitor logs, enforce Zero Trust	High customer responsibility: Where AI acts — security depends on how you control agent access, what tools they have access to, and actions. Know your data flows.
MICROSOFT 365 COPILOT	Secure M365 integration, baseline Responsible AI filters, Entra ID integration, access control enforcement	License/user enablement, data governance (Purview/DLP), access policies, monitoring/audit, user training	Mostly Microsoft-secured, but data governance and user behavior determine your risk.
COPILOT STUDIO	Platform security, connector catalog, isolation, identity integration, lifecycle management	Design prompts/flows, govern connectors, secure data IO, monitor copilots, developer RBAC, kill switch	High customer responsibility. You own agent design, integrations, and guardrails.



What To Do Next

PROARCH RECOMMENDATIONS

UNDERSTAND AI USAGE

Visibility into how AI is being used in your environment.

- Purview, Defender for Cloud Apps

STRATEGIZE AI ADOPTION

Decide whether to build or buy GenAI apps.

- Copilot, Copilot Studio, Azure AI Foundry

FORM A CROSS-FUNCTIONAL TEAM

Define governance board and security team/process for AI decisions

CLARIFY ROLES

Define responsibilities across cloud providers, resource owners, and security teams.

STRENGTHEN INFRASTRUCTURE

- Upgrade systems
- Tighten access controls
- Enhance data governance
- Invest in compliance, legal, and training resources.

IMPLEMENT ZERO TRUST

Validate every access request, user, device, and identity.

- Purview for data classification, labeling, and protection

SECURE AI WORKLOADS

Maintain inventory of models, plugins, and sensitive data. Monitor custom AI apps for misconfigurations and risks.

- Purview to govern data flowing into AI systems and enforce DLP policies

MAKE A LONG-TERM PLAN

Have the ability to:
Identify, respond and mitigate threats, misconfigurations, leaked data, unauthorized changes, shadow AI/tools.

AI SECURITY BLUEPRINT

3 Pillars for Secure Enterprise Adoption

1

VISIBILITY & GOVERNANCE

Discover, classify, and control AI usage



- MDR: Shadow AI Discovery
- Governance Frameworks
- Virtual Chief AI Officer
- AI Security Readiness Assessment

2

DATA PRIVACY & SECURITY CONTROLS

Protect sensitive data in AI workflows



- MDR: Data Security
- DLP (Purview) Implementation
- Zero Trust Assessment
- Secure AI and Data Implementation

3

ADAPTIVE SECURITY & INNOVATION

Evolve defenses with AI threat landscape



- MDR: AI Threat Intelligence
- Use Case Security Design
- Continuous Posture Improvement

Questions?



THANK YOU FOR JOINING US | [PROARCH.COM](https://proarch.com)