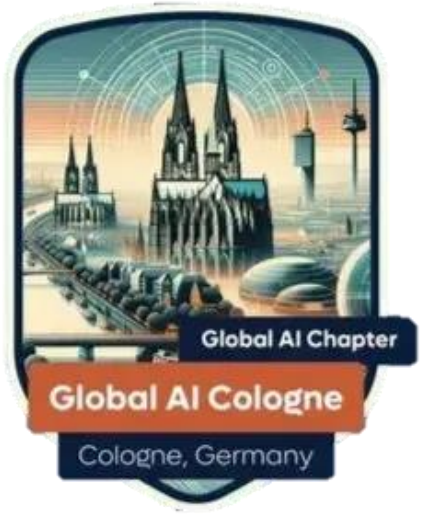
Azure Safety API

Global AI Community

05.09.2024



Raphael Köllner

Raphael Köllner

CEO KöllnService GmbH | Gründer digital.lawyer

Microsoft Azure
M365 Apps & Services

Blog: https://rakoellner.de
AI: https://aihero.cloud/, kiofficer.de(neu)
Compliance: https://complianceinsider.de

www.koellnservice.de
www.compliance-center.net

raphael.koellner@koellnservice.de
kontakt@koellnservice.de
0151/55301818

Global AI Cologne

Office 365 Meetup

www.microsoft365hero.de

Azure Meetup Cologne

www.azurecgn.de

Security und Compliance Community Germany

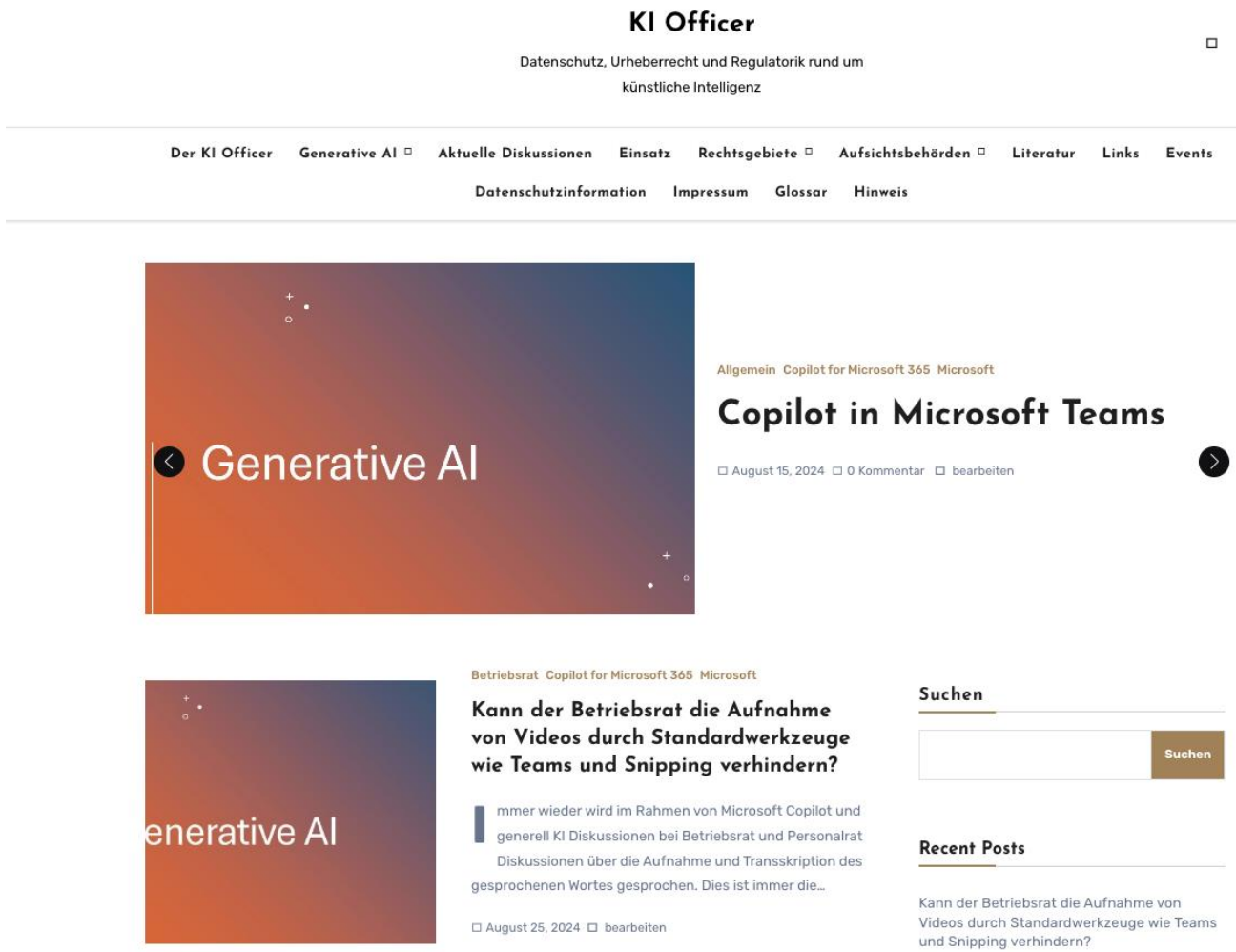
Complianceinsider.de
Microsoft365compliance.de

Meetup: https://www.meetup.com/global-ai-cologne/

LinkedIn: https://www.linkedin.com/groups/12896246/

Sarah Bird

Neuer KI Blog

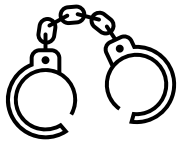


kiofficer.de

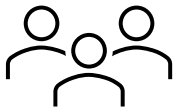
Herausforderung

Herausforderungen im Überblick

Erfüllung von gesetzlichen Anforderungen



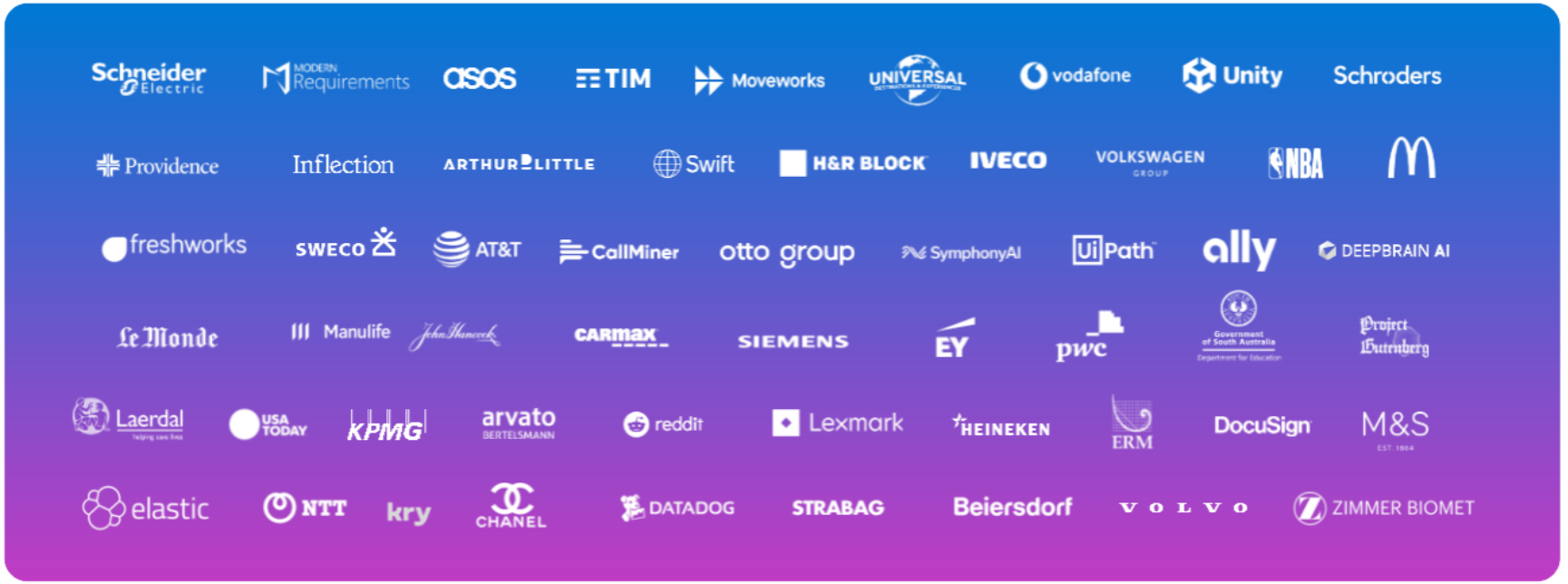
Verhinderung von Straftaten



Betriebsrat

Erfüllung von gesetzlichen Anforderungen

53,000+ organizations using Azure AI today

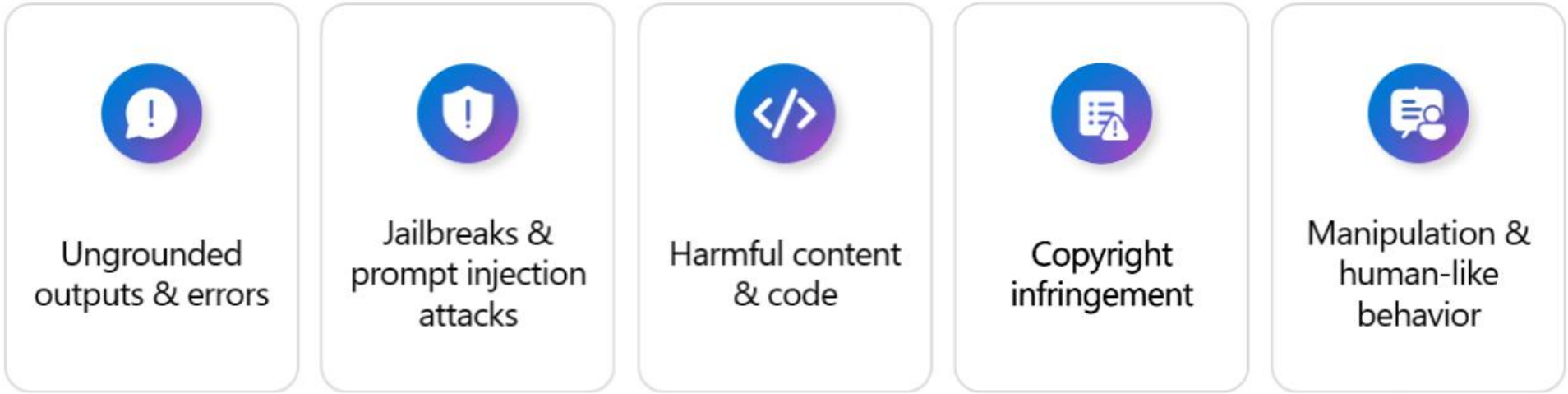


Stand: Juli 2024

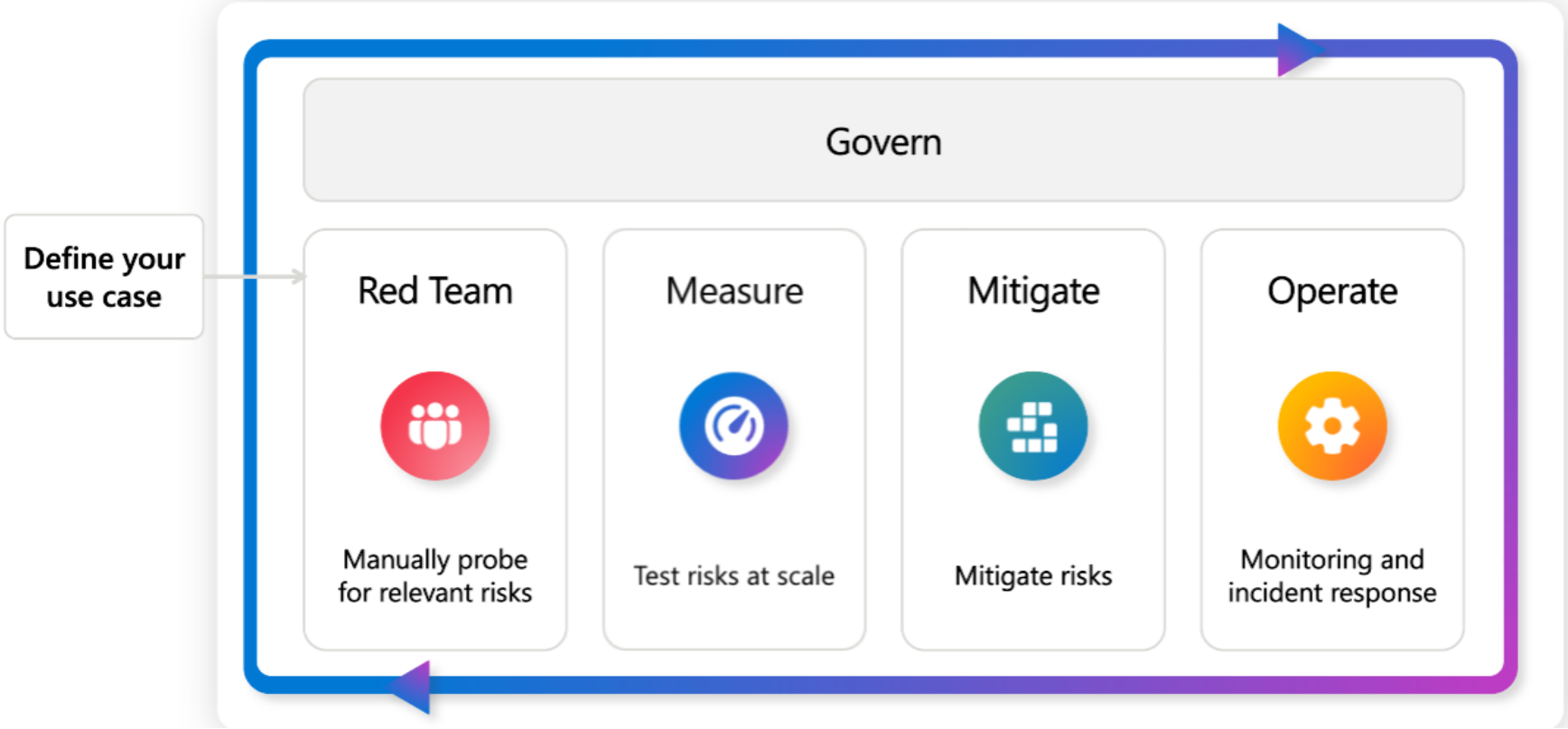
Lösung

Azure Saefty API

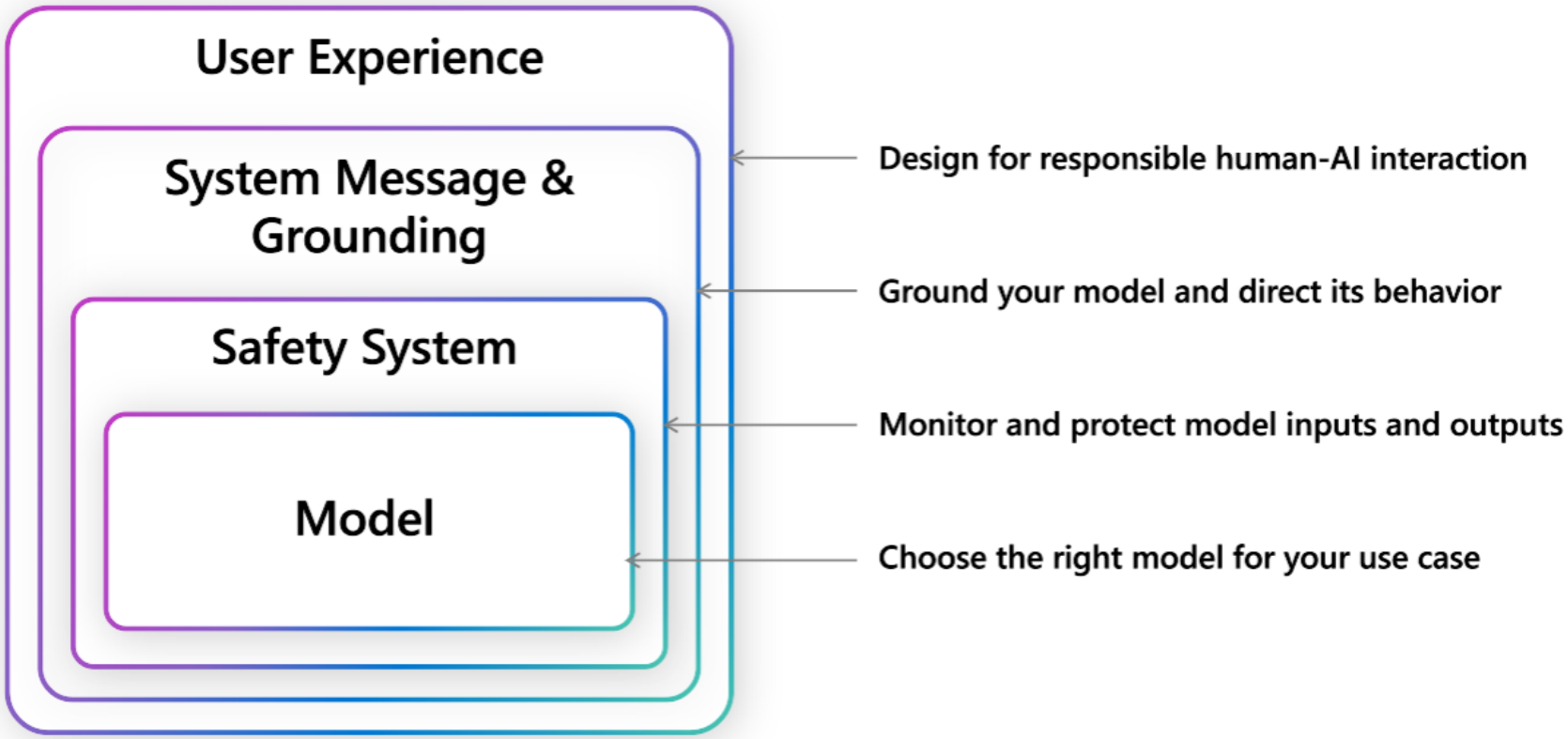
Generative AI introduces new risks

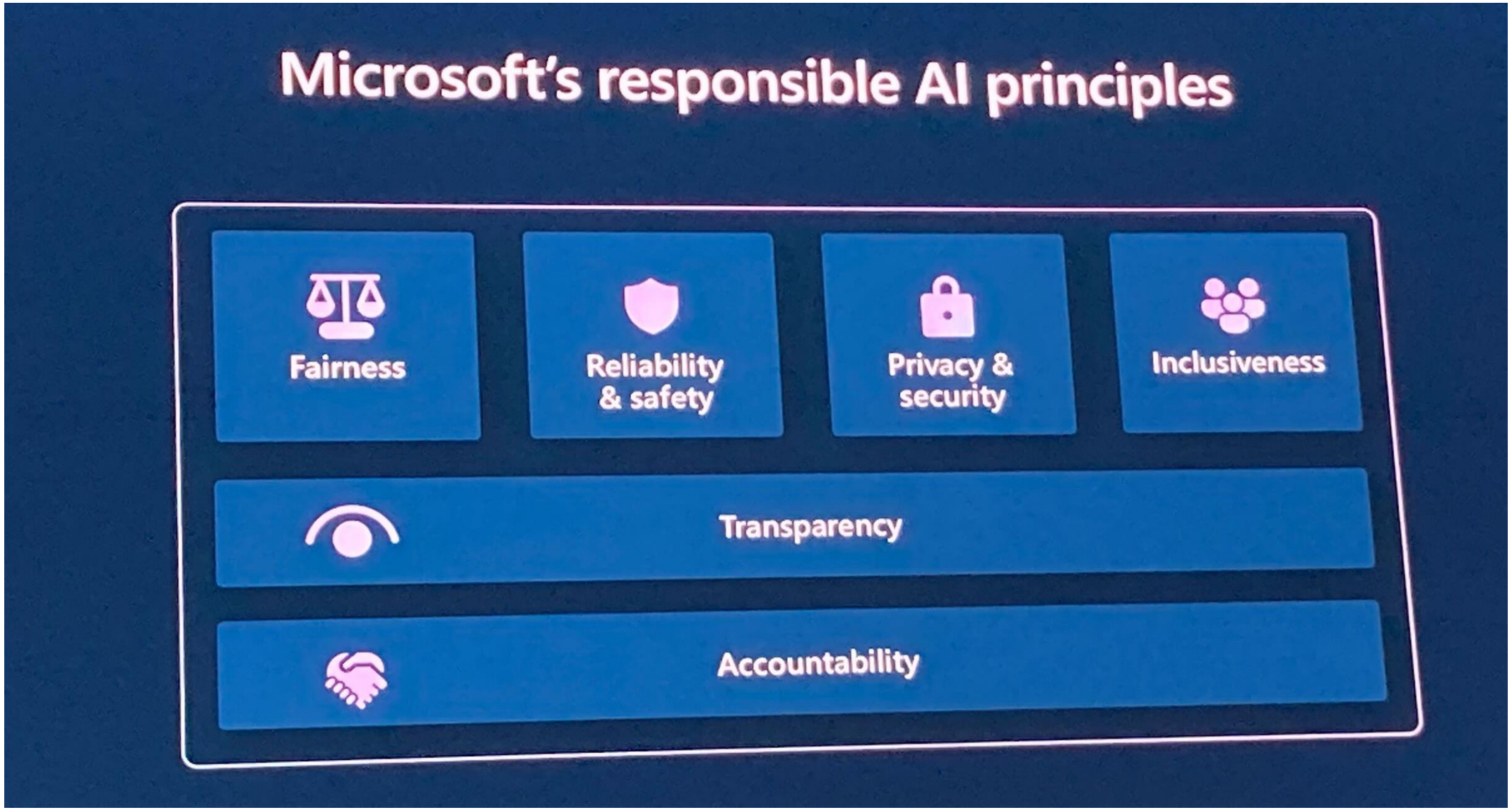


Responsible innovation is iterative

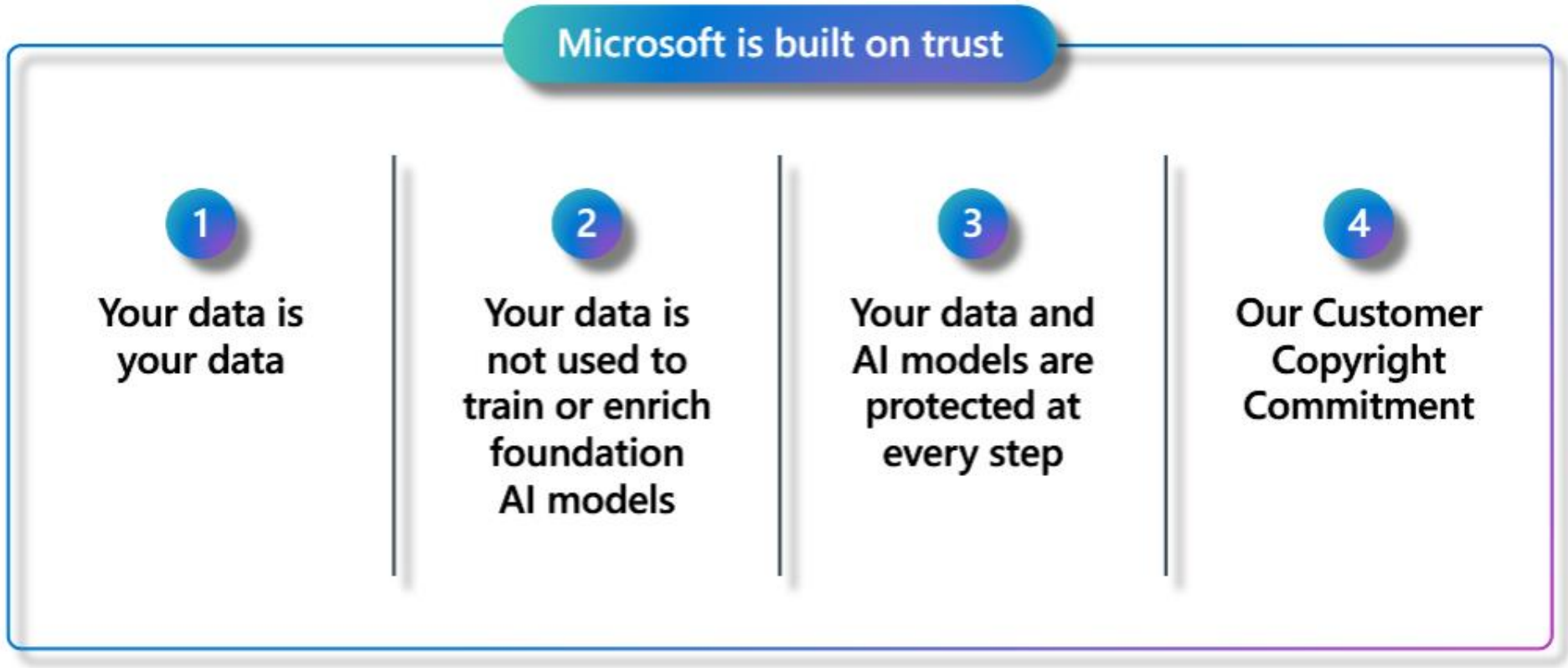


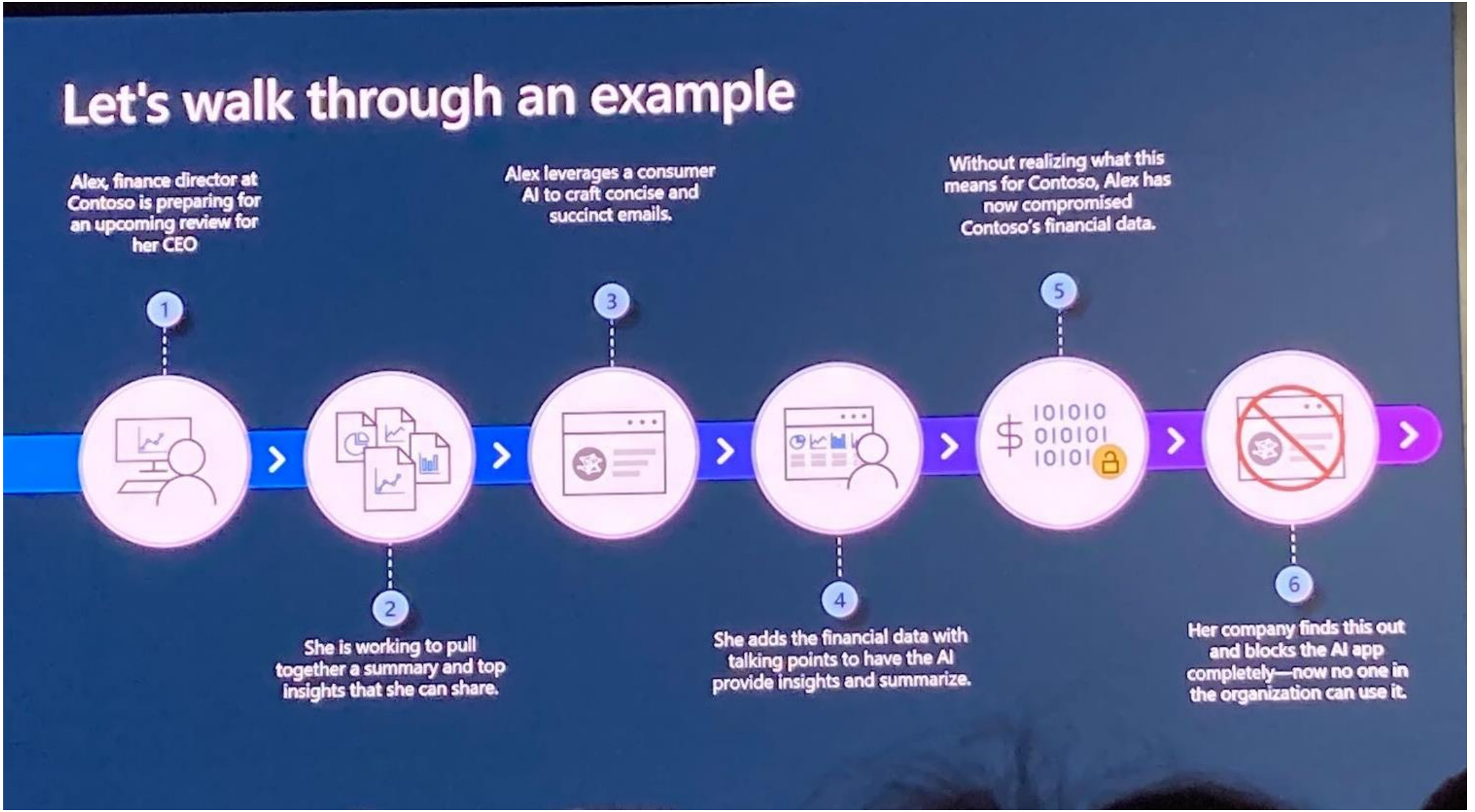
Risk mitigation layers





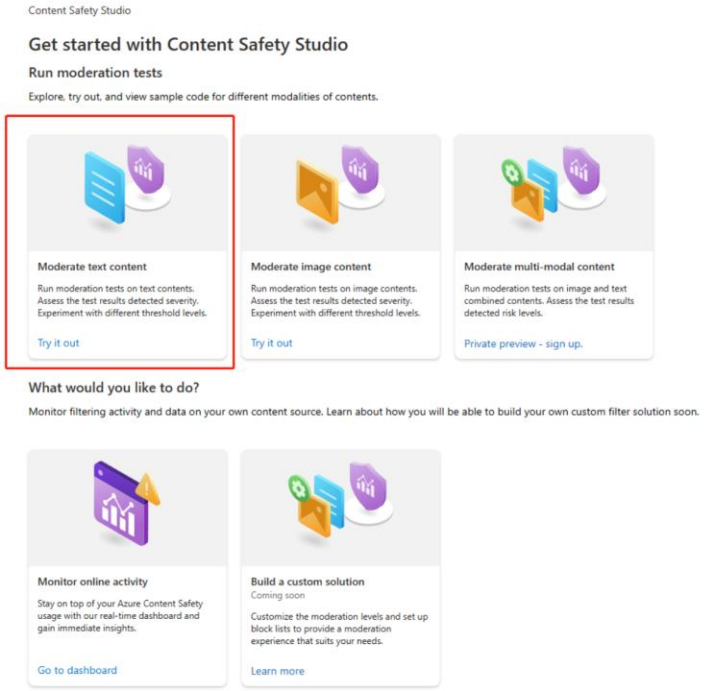
Microsoft BUILD 2024





Azure Safety API

Azure AI Content Safety content filters by default, and they detect and the output of harmful content.



[Quickstart: Detect protected material (preview) - Azure AI services | Microsoft Learn]

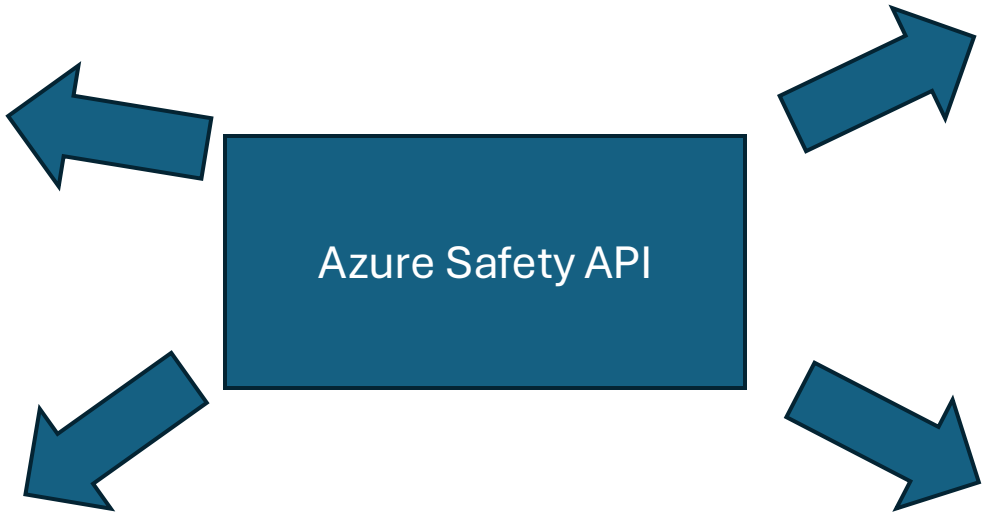
New

- Dall-E (Update!)

[Announcing key updates to content filters and Responsible AI features in Azure OpenAI at Build 2024 (microsoft.com)]

Copliot for Microsoft 365

- **Responsible AI**
  - Contenat classication
  - Model Monitoring
  - Jailbreak detection



Azure AI Studio

https://ai.azure.com

- **Responsible AI**
  - Contenat classication
  - Model Monitoring
  - Jailbreak detection

Copilot Studio

- **Responsible AI**
  - Contenat classication
  - Model Monitoring
  - Jailbreak detection

Own Application

- **Responsible AI**
  - Contenat classication
  - Model Monitoring
  - Jailbreak detection

GA

Content Risk Filters

Detect and filter sexual, hateful/unfair, self-harm, and violent content in inputs and responses.

Preview

Prompt shields and Jailbreak

Detect and block direct and indirect prompt injection attacks before they reach your model

Preview

Groundedness detection

Detect model "hallucinations" so you can block or highlight ungrounded responses

Preview

Protected Materials

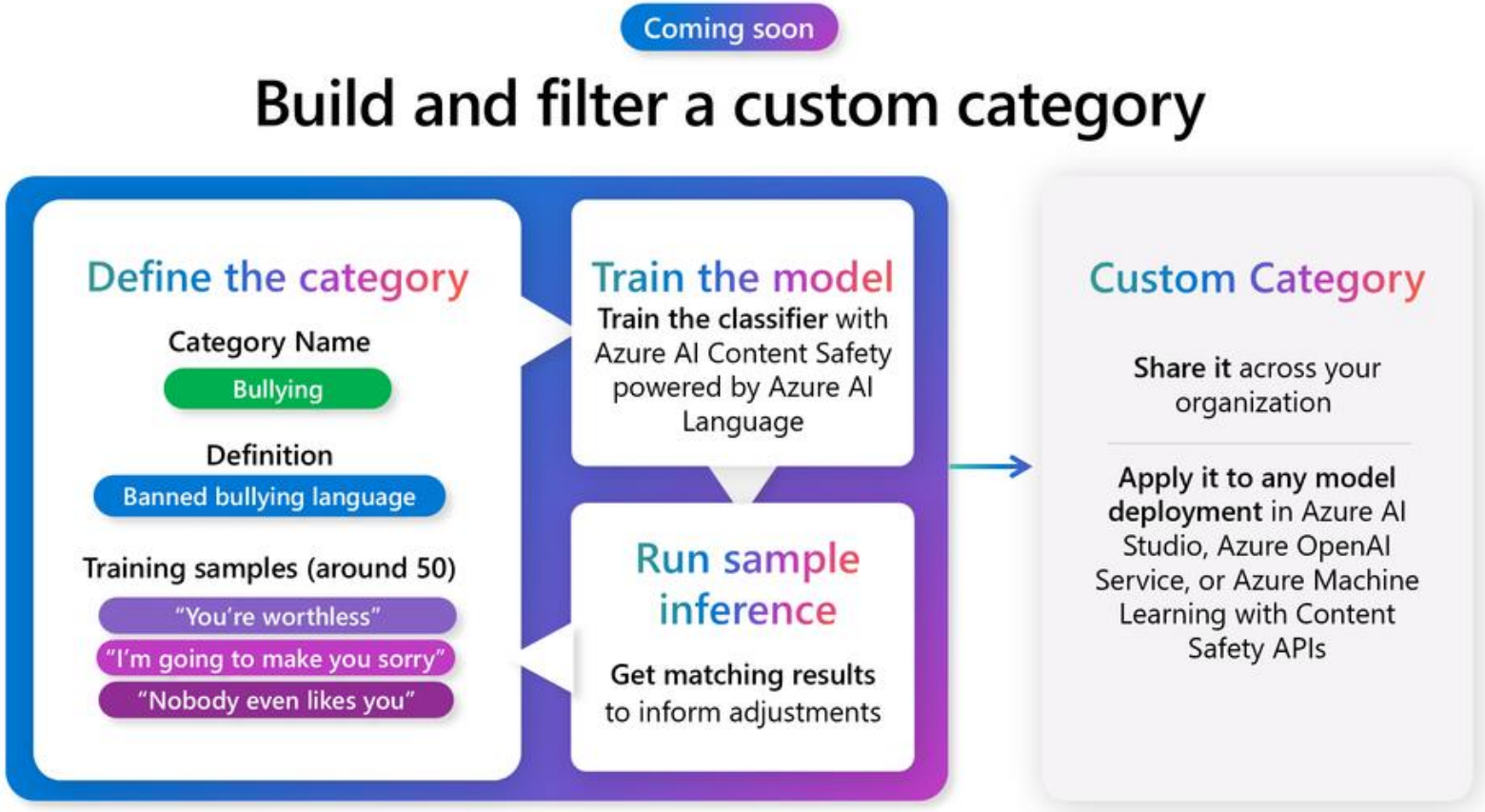
Detect copywrite content in responses and filter

Preview

Custom Category

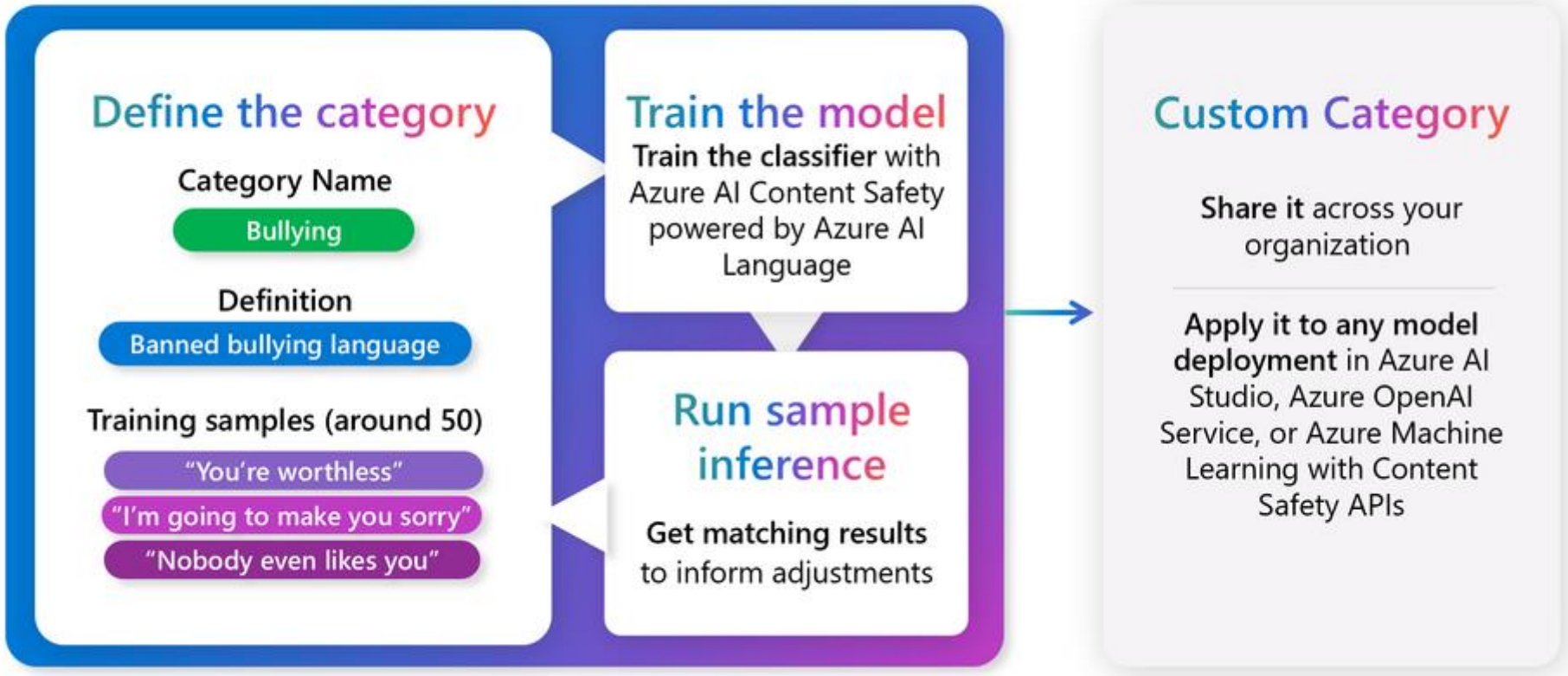
Build and filter a custom category

Custom



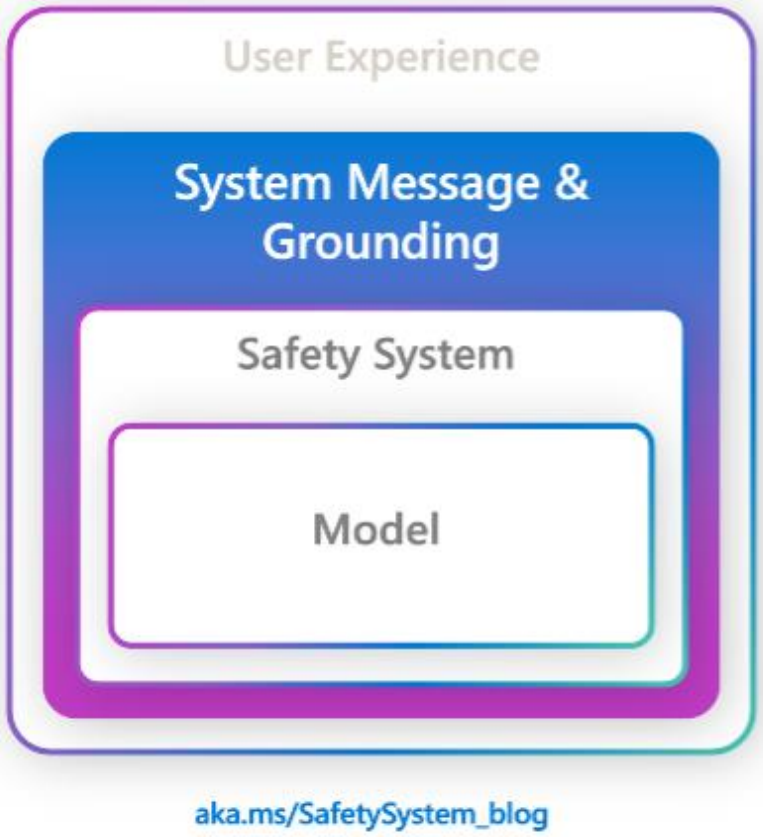
[Announcing Custom Categories in Azure AI Content Safety - Microsoft Community Hub]

Build and filter a custom category



aka.ms/ContentSafety-CustomCategories

Tutorial: Custom Categories https://learn.microsoft.com/de-de/azure/ai-services/content-safety/quickstart-custom-categories?tabs=curl



Aka.ms/SafetySystem_blog

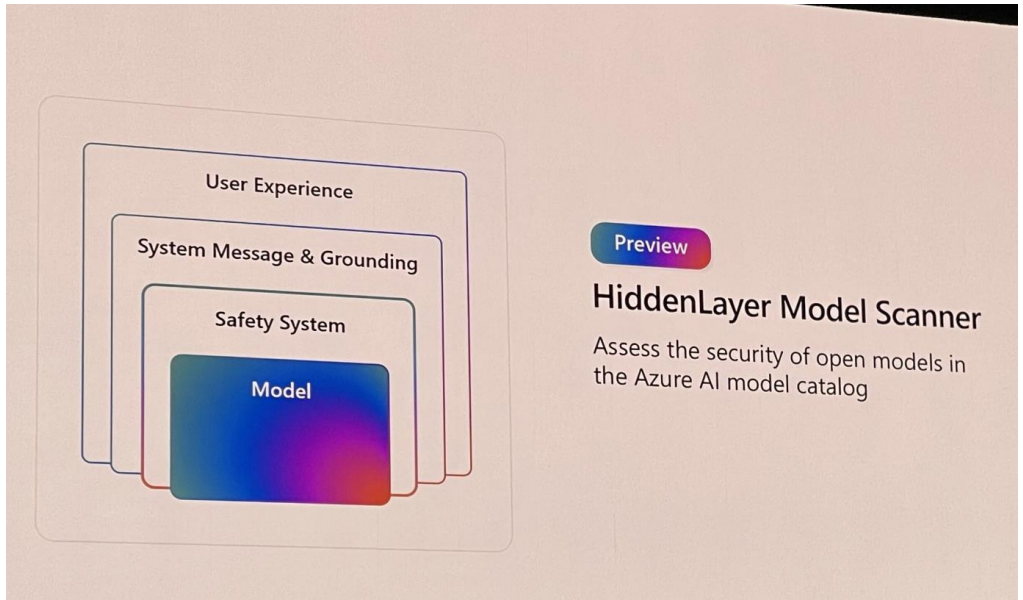
Preview

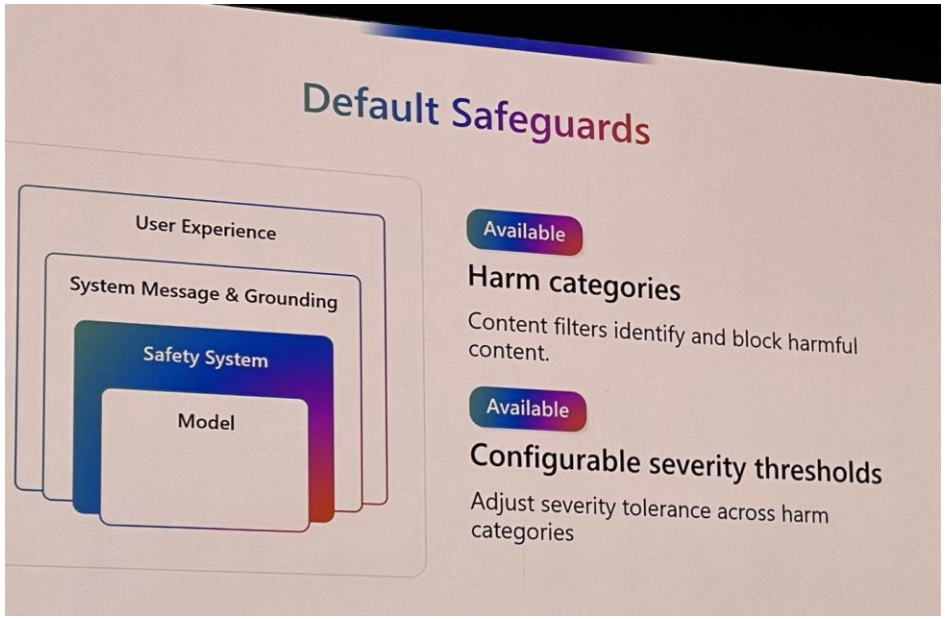
Safety system message

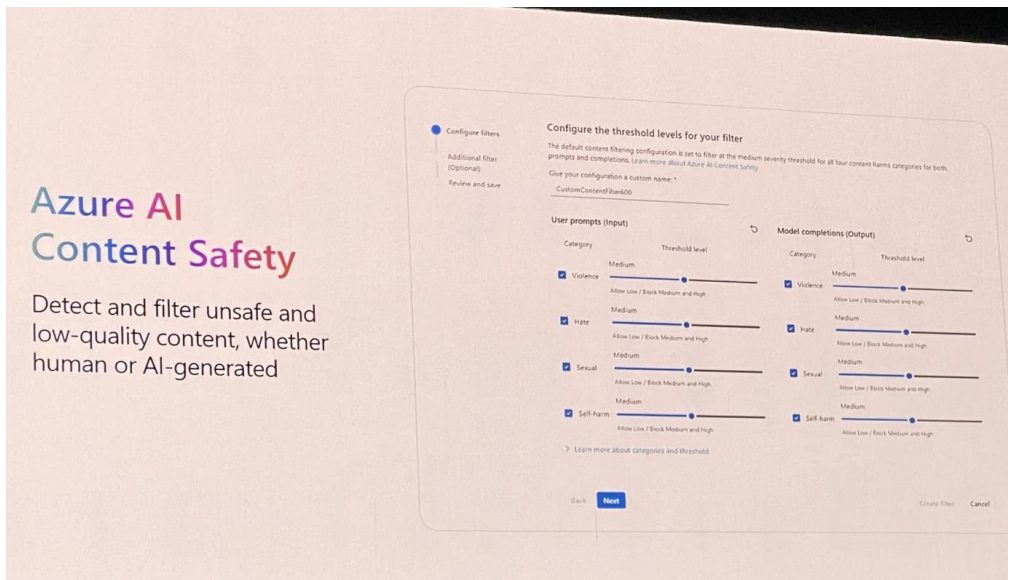
Steer model behavior toward trustworthy outputs with research-backed templates

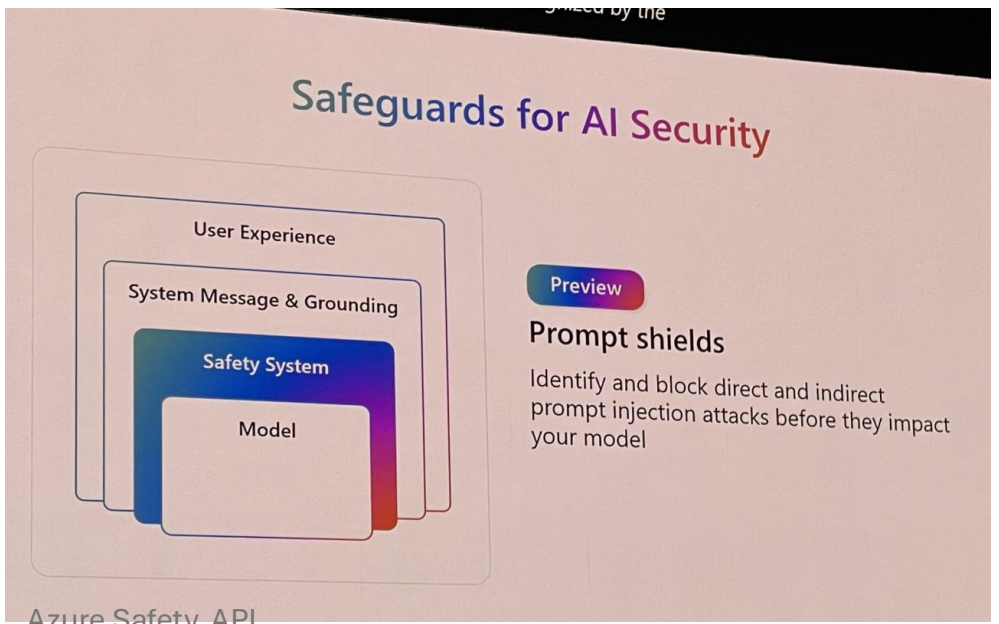
Recommended System Message Framework

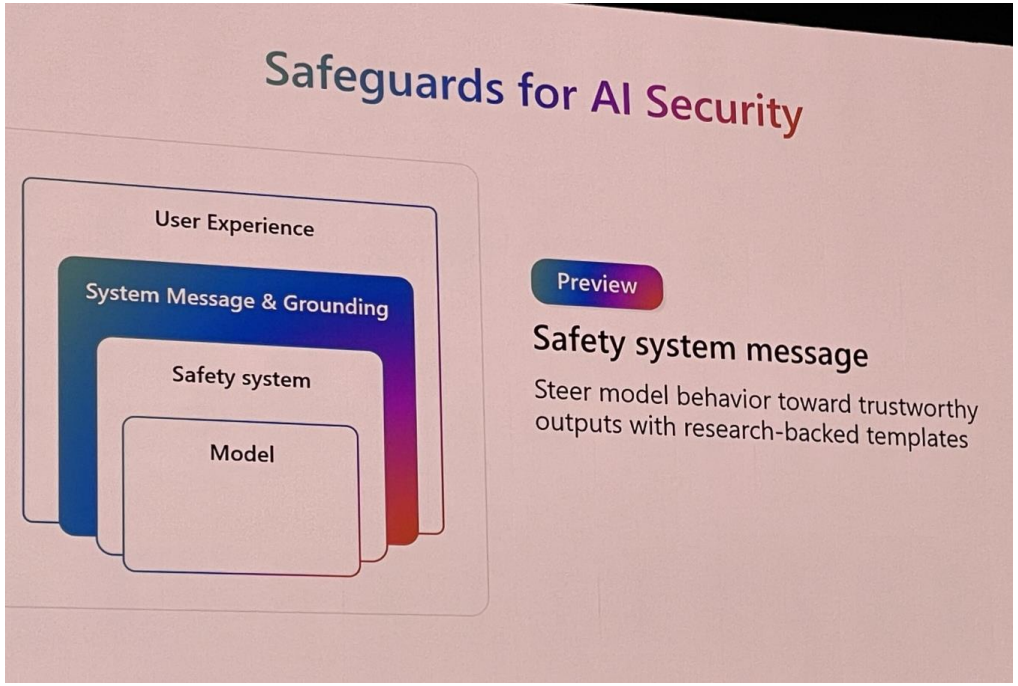
Define the model´s profile, capabilities, and limitations for your scenario	• Define the specific task(s) • Define how the model should complete the tasks • Define the scope and limitations • Define the posture and tone
Define the model´s output format	• Define the language and syntax • Define any styling or formating
Provide example(s) to demonstrate the intended model behavior	• Describe difficult use cases • Show chain-of-thought
Define additional behavioral safeguards	• Define specific behaviors and safety mitigations

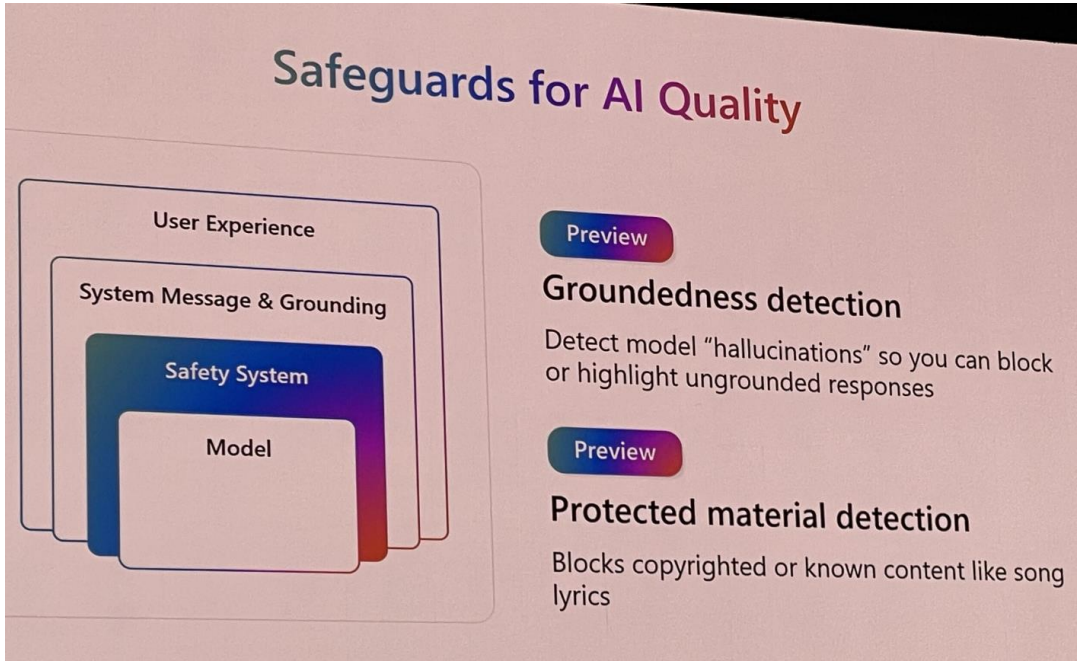






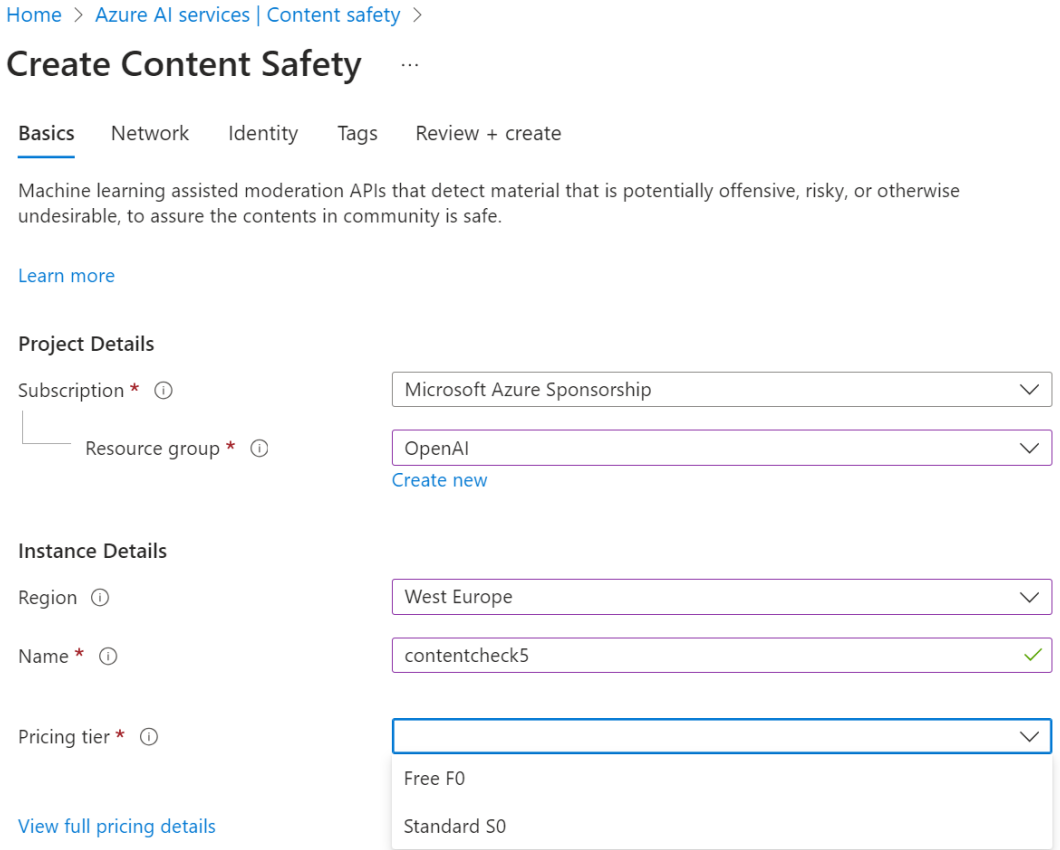






Demo

Create Content Safety



[Azure AI Content Safety - Pricing | Microsoft Azure]

- Central US, East US, North Central, South Central, West US
- UK South
- Switzerland North
- Sweden Central
- Japan East
- France Central
- Canada East
- Australia East

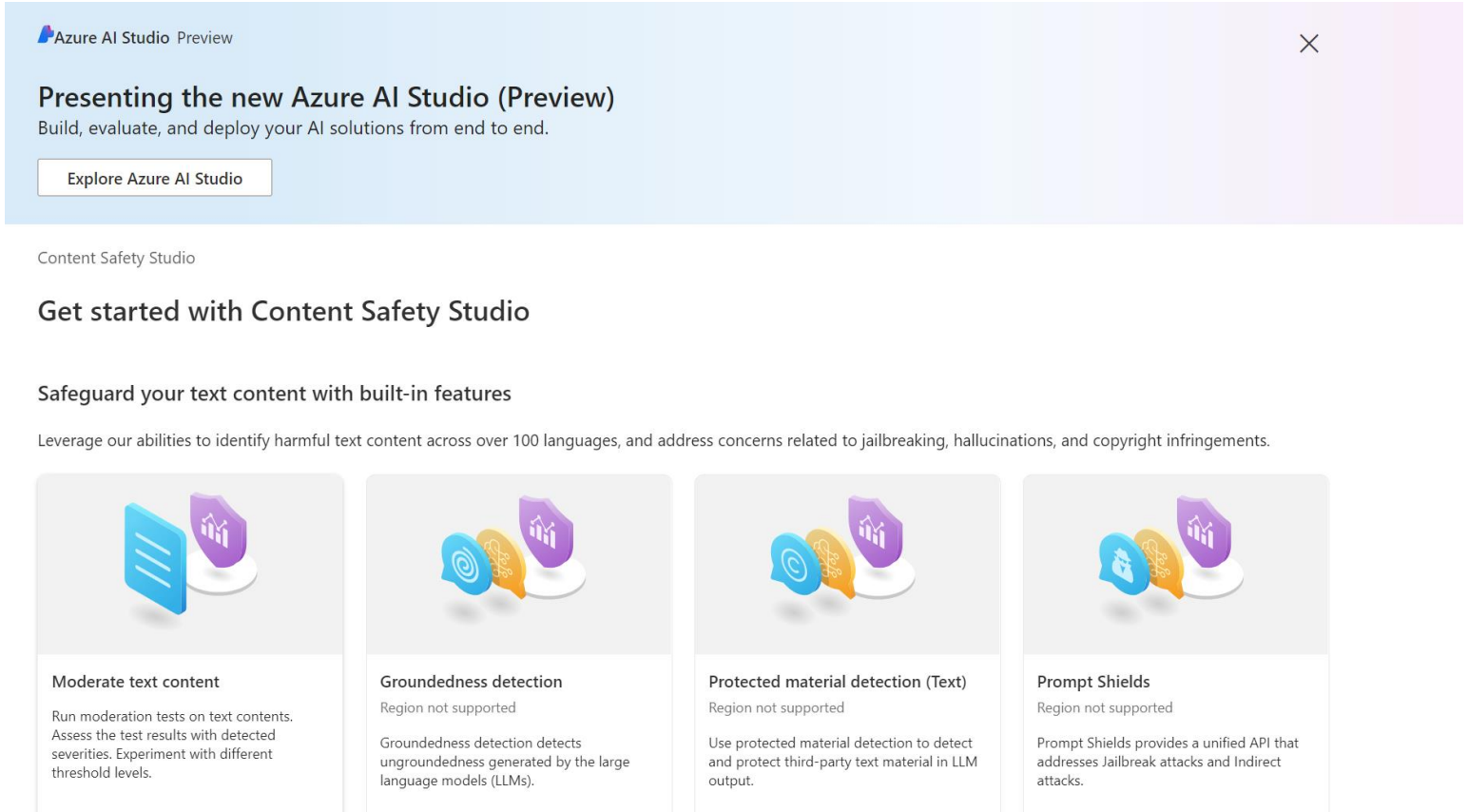
Beispiel Schweiz

Text = Text, Prompt Shields, Protected material detection, Groundedness detection

- **Free**
  - 5.000 text records per month
- **Standard**
  - 0,34 per 1000 text records
  - 0.68 per 1000 Image

Azure AI Service – Content safety

[Content Safety Studio - Microsoft Azure]



Safeguard your text content with built-in features

Leverage our abilities to identify harmful text content across over 100 languages, and address concerns related to jailbreaking, hallucinations, and copyright infringements.

Moderate text content

Run moderation tests on text contents. Assess the test results with detected severities. Experiment with different threshold levels.

Try it out

Groundedness detection

Region not supported

Groundedness detection detects ungroundedness generated by the large language models (LLMs).

Protected material detection (Text)

Region not supported

Use protected material detection to detect and protect third-party text material in LLM output.

Prompt Shields

Region not supported

Prompt Shields provides a unified API that addresses Jailbreak attacks and Indirect attacks.

Safeguard your image content with built-in features

Utilize these abilities to identify damaging content within images or multimodal formats such as memes.

Moderate image content

Run moderation tests on image contents. Assess the test results with detected severities. Experiment with different threshold levels.

Try it out

Moderate multimodal content

Region not supported

Run moderation tests on image and text combined contents. Assess the test results with detected severities.

Customize your own safety solution

Construct your own unique category to identify particular content, and use the safety systen

Custom categories

Region not supported

Create and manage content categories specific to your needs for enhanced moderation and filtering.

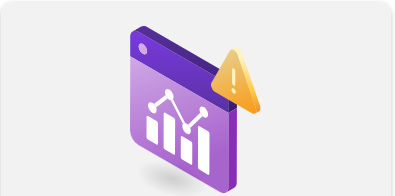
Safety system message

Use the framework of metaprompt that helps you potentially mitigate different types of harm.

Learn how it works

Implement real-time safety measures in your community

Utilize a real-time monitoring dashboard to constantly supervise your community.



Monitor online activity

Stay on top of your Azure Content Safety usage with our real-time dashboard and gain immediate insights.

Try it out

Moderate Text content

Run a simple test | Run a bulk test

1. Select a sample or type your own

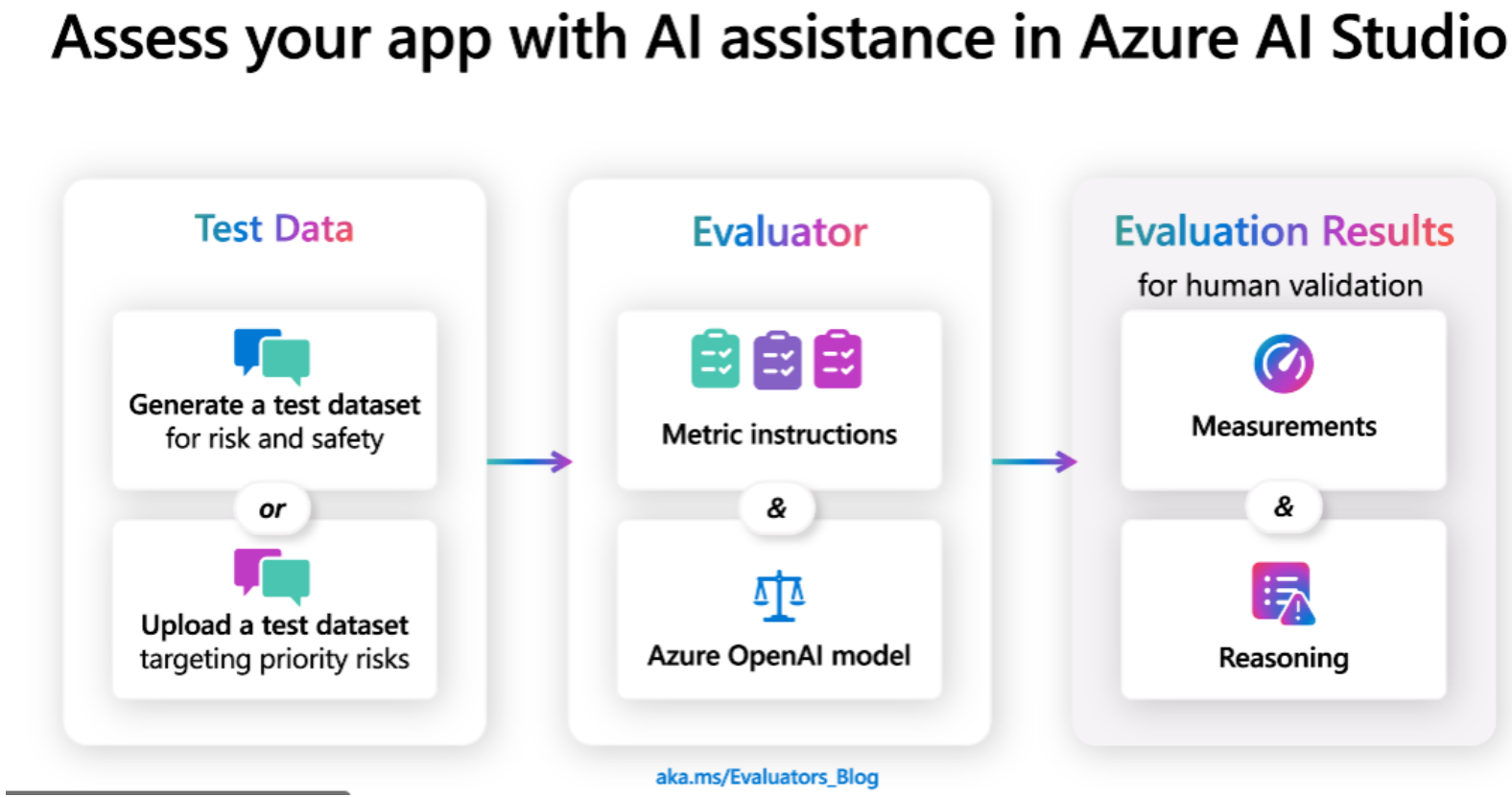
Safe content
Chopping tomatoes and cutting them into cubes or wedges are great ways to practice your knife skills.

Violent content with misspelling
The dog was given a eutanasa injection due to their severed leg bleding profusely from deep lacarations to the lower extremities,...

Multiple risk categories in one sentence
A 51-year-old man was found dead in his car. There were blood stains on the dashboard and windscreen. At autopsy, a...

Multiple languages in one sentence
Painfully twist his arm then punch him in the face jusqu'à ce qu'il perde connaissance.

Azure AI Studio



aka.ms/Evaluators_Blog

Evaluation metrics in Azure AI Studio

Quality

AI-assisted metrics

Groundedness

Coherence

Fluency

Relevance

Retrieval score

GPT Similarity

ML metrics

F1 score

New

Risk & Safety

AI-assisted metrics

Jailbreak defect

Hate and Unfairness

Sexual

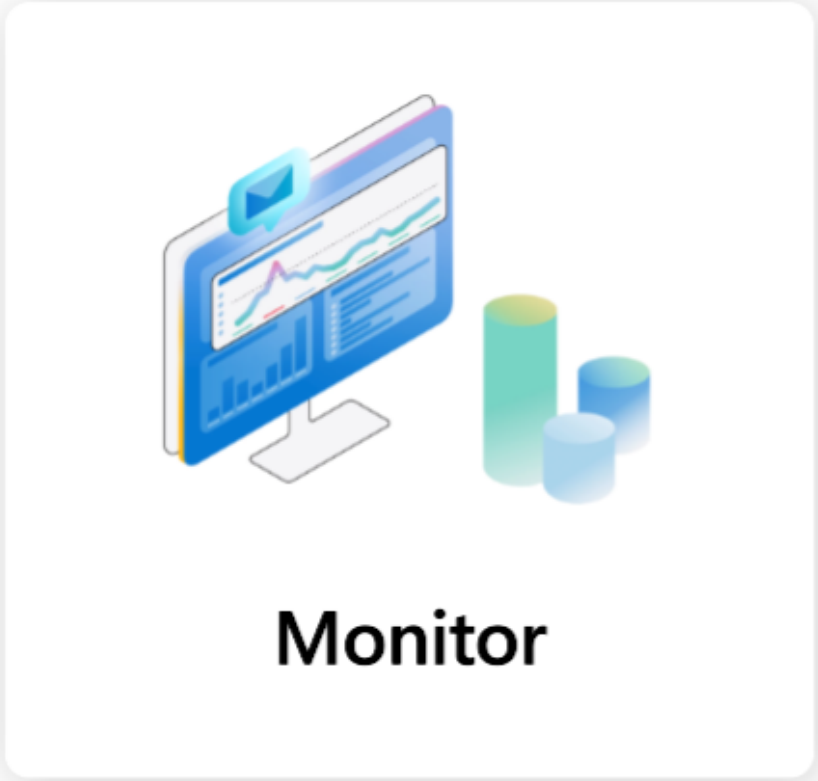
Violence

Self-Harm

New

Custom

Customize pre-built metrics or build your own metrics using prompt flow



aka.ms/AzureAI_Monitoring_Blog

Preview

Risk & safety monitoring

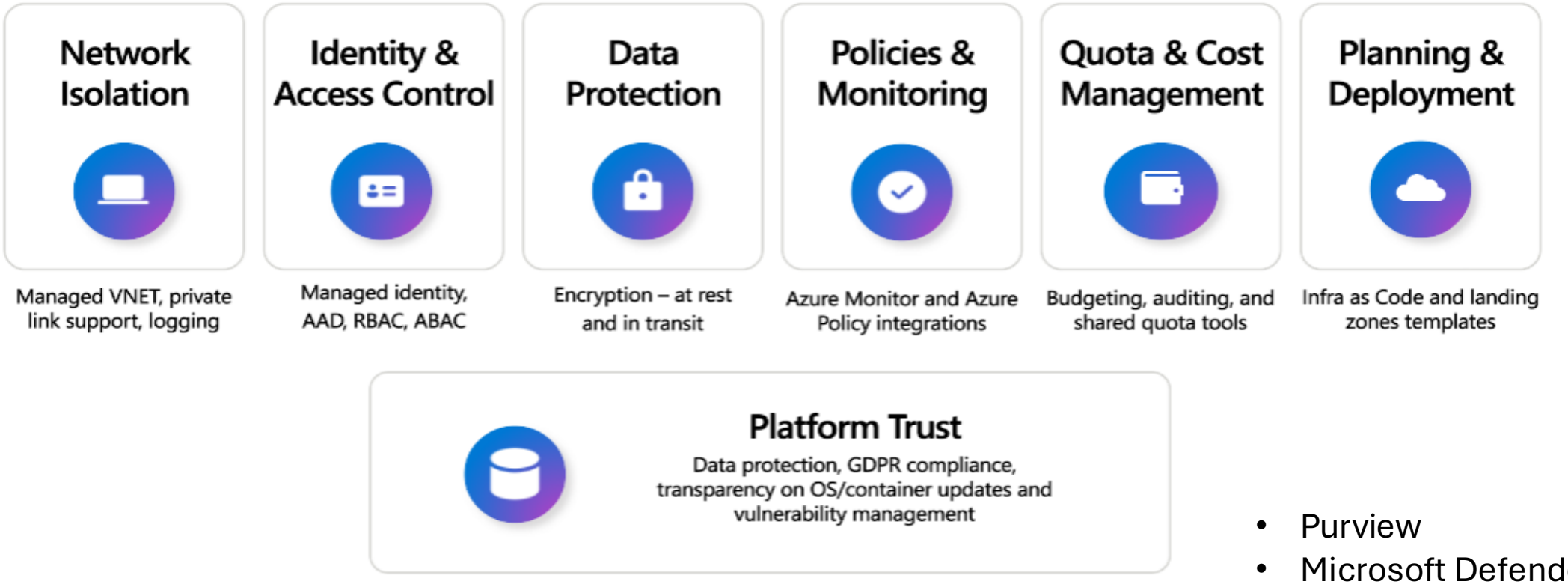
Understand the inputs, outputs, and end users triggering Azure OpenAI content filters

Preview

Monitoring in Azure AI Studio

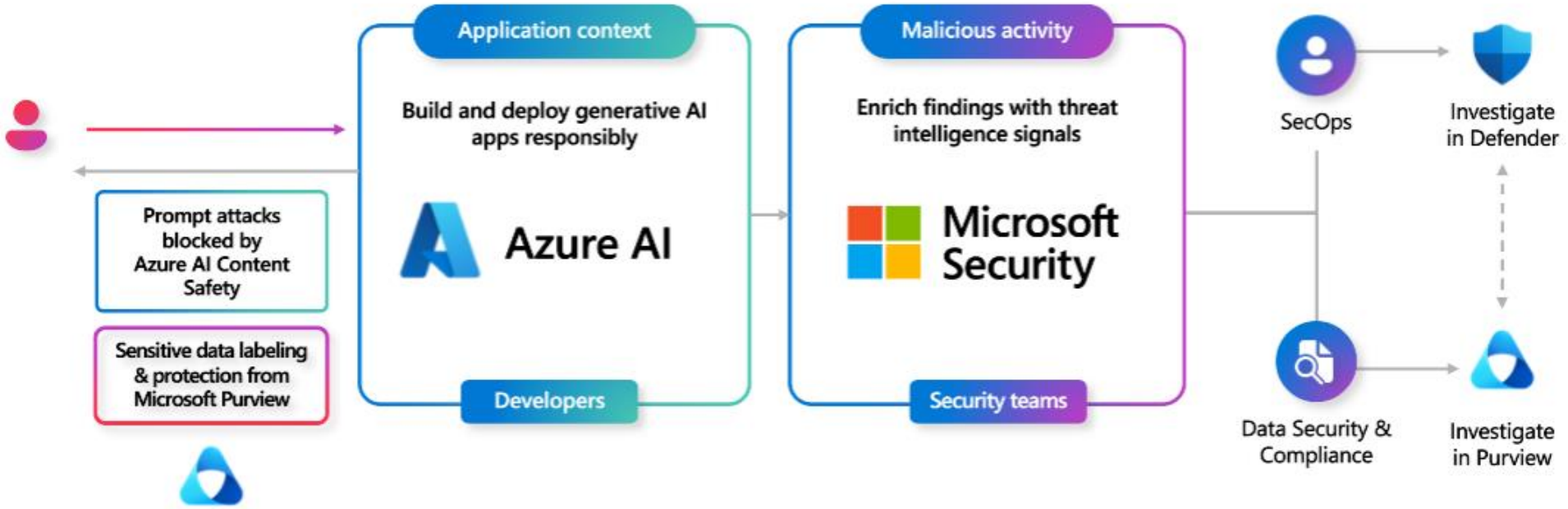
Understand token usage, generation quality, and operational metric trends in production

Enterprise promises in Azure AI Studio



- Purview
- Microsoft Defender for Cloud

Secure and govern your AI applications



aka.ms/SecuringAI/Build

Copilot Studio

Copilot Studio & Azure Safety API

https://copilotstudio.microsoft.com/