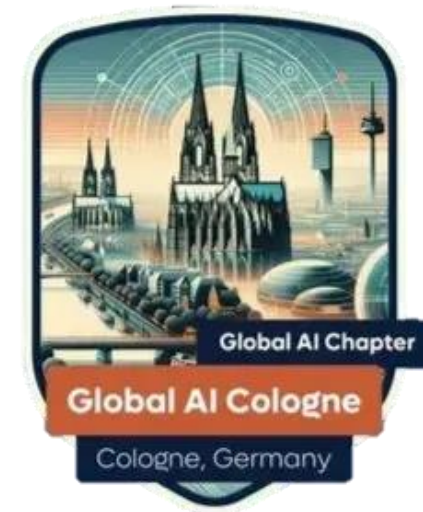


Azure Safety API

Global AI Community

05.09.2024



Raphael Köllner

Raphael Köllner

CEO KöllnService GmbH | Gründer digital.lawyer



Microsoft Azure
M365 Apps & Services



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



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

Blog: <https://rakoellner.de>

AI: <https://aihero.cloud/>, kiofficer.de(neu)

Compliance: <https://complianceinsider.de>



Global AI Cologne



Office 365 Meetup

www.microsoft365hero.de



Azure Meetup Cologne

www.azurecgn.de



Security und Compliance
Community Germany

[Complianceinsider.de](https://complianceinsider.de)
[Microsoft365compliance.de](https://microsoft365compliance.de)



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



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





Sarah Bird

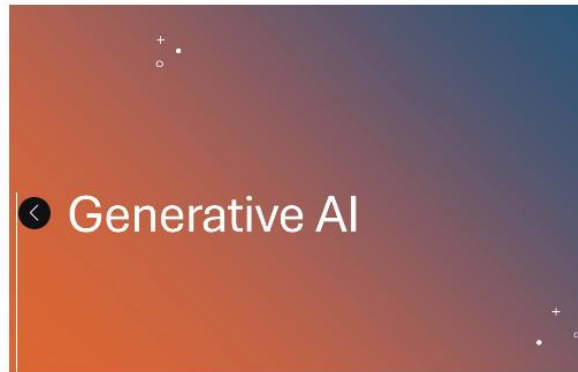
Neuer KI Blog

KI Officer

Datenschutz, Urheberrecht und Regulatorik rund um künstliche Intelligenz



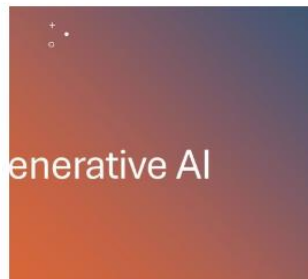
Der KI Officer Generative AI Aktuelle Diskussionen Einsatz Rechtsgebiete Aufsichtsbehörden Literatur Links Events
Datenschutzinformation Impressum Glossar Hinweis



Allgemein Copilot for Microsoft 365 Microsoft

Copilot in Microsoft Teams

August 15, 2024 0 Kommentar bearbeiten



Betriebsrat Copilot for Microsoft 365 Microsoft

Kann der Betriebsrat die Aufnahme von Videos durch Standardwerkzeuge wie Teams und Snipping verhindern?

Immer wieder wird im Rahmen von Microsoft Copilot und generell KI Diskussionen bei Betriebsrat und Personalrat Diskussionen über die Aufnahme und Transkription des gesprochenen Wortes gesprochen. Dies ist immer die...

August 25, 2024 bearbeiten

Suchen

Recent Posts

Kann der Betriebsrat die Aufnahme von Videos durch Standardwerkzeuge wie Teams und Snipping verhindern?

kiofficer.de

24.04.2025

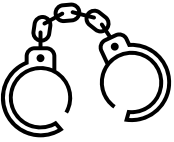
Azure Safety API

Herausforderung

Herausforderungen im Überblick



Erfüllung von gesetzlichen Anforderungen



Verhinderung von Straftaten

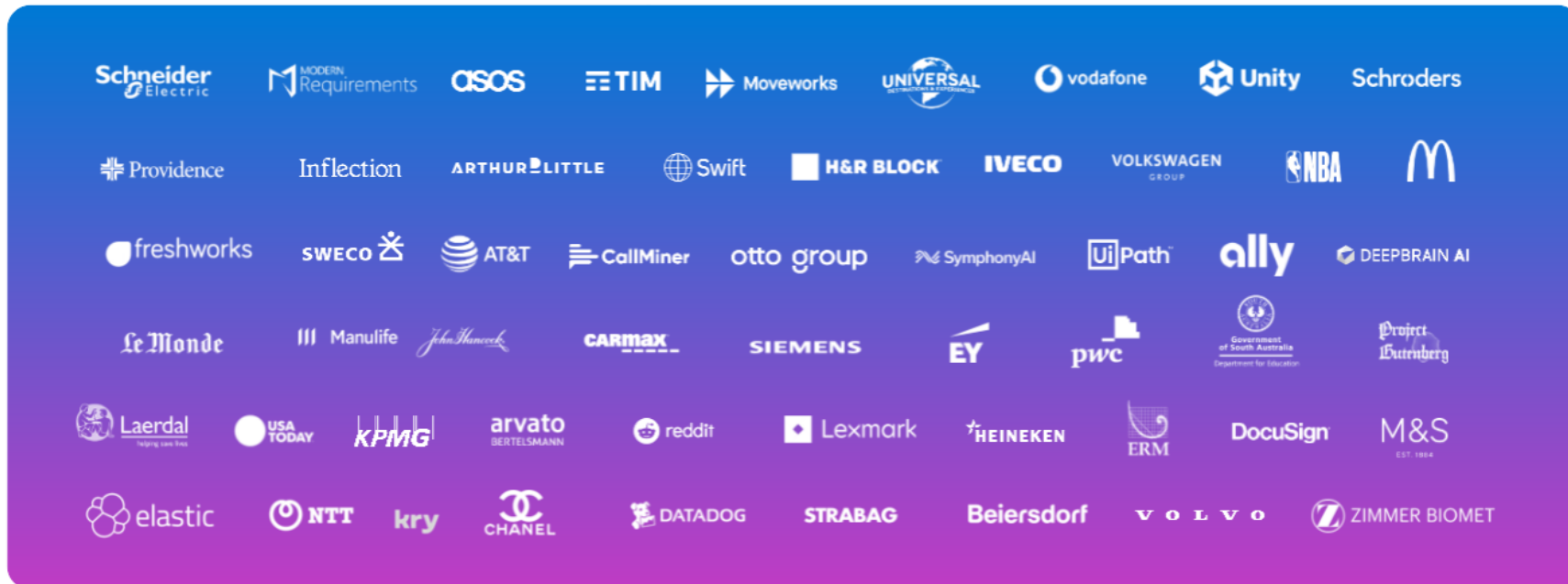


Betriebsrat



Erfüllung von gesetzlichen Anforderungen

53,000+ organizations using Azure AI today



Stand: Juli 2024

Lösung

Azure Saefly API

Generative AI introduces new risks



Ungrounded
outputs & errors



Jailbreaks &
prompt injection
attacks



Harmful content
& code

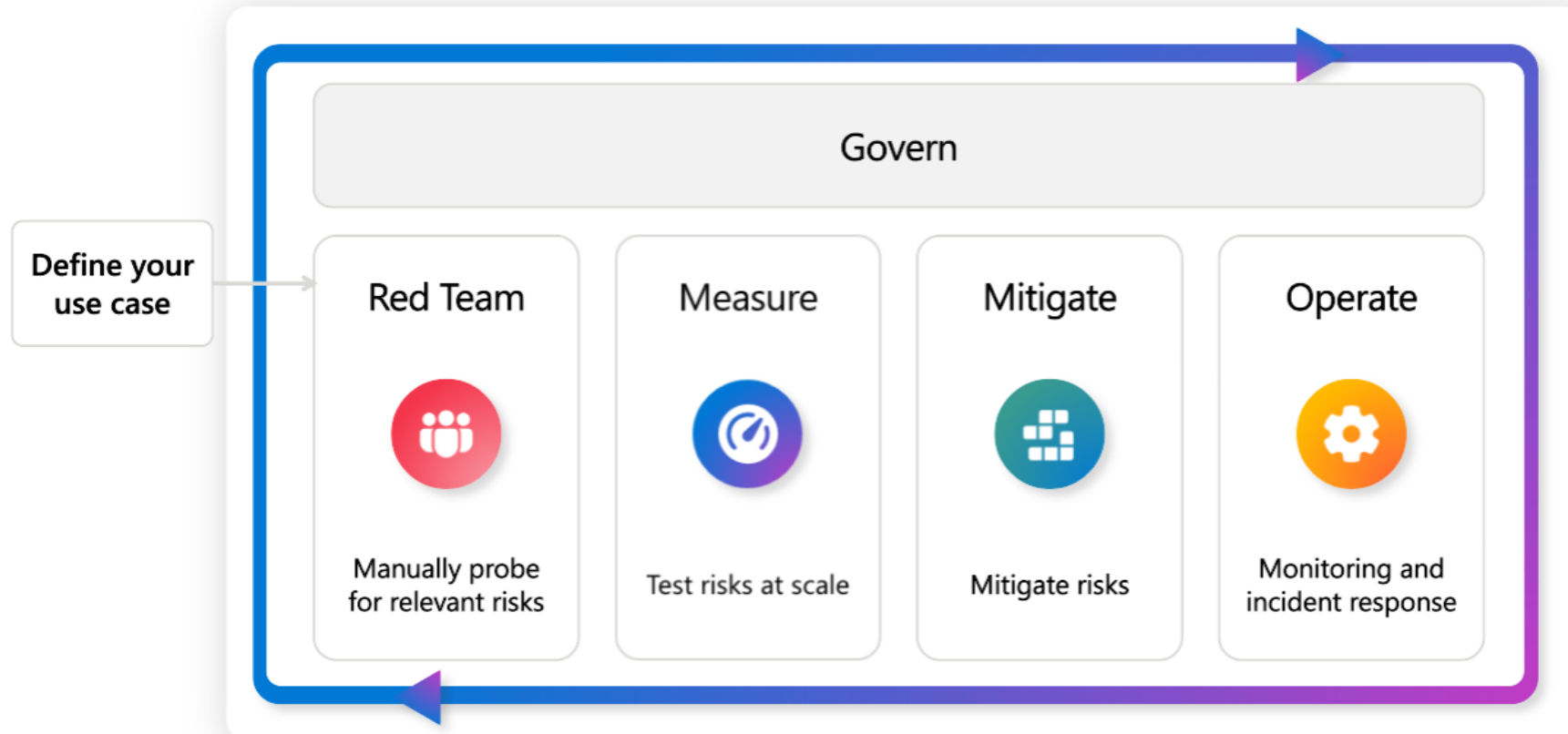


Copyright
infringement

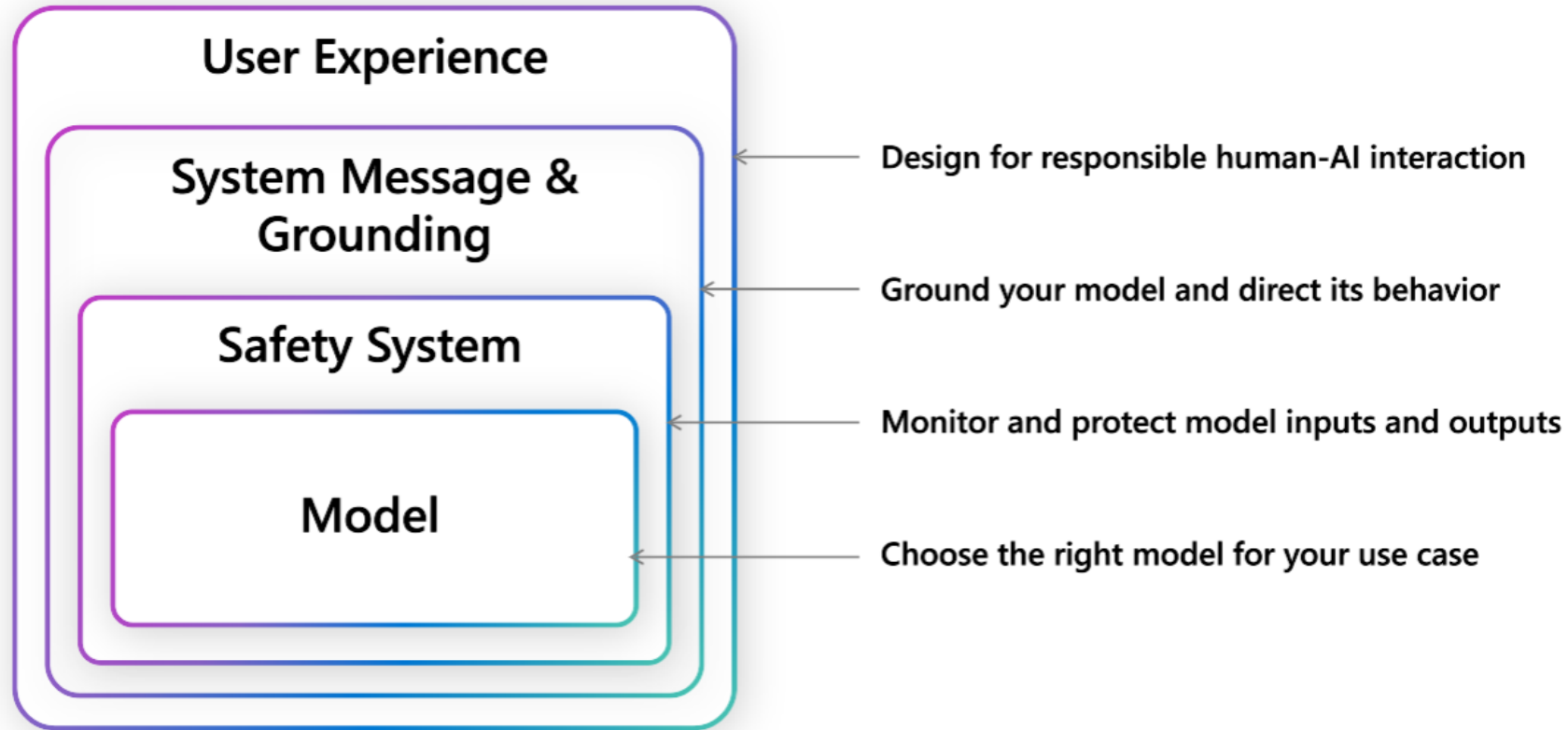


Manipulation &
human-like
behavior

Responsible innovation is iterative



Risk mitigation layers



Microsoft's responsible AI principles



Microsoft BUILD 2024

Microsoft is built on trust

1

**Your data is
your data**

2

**Your data is
not used to
train or enrich
foundation
AI models**

3

**Your data and
AI models are
protected at
every step**

4

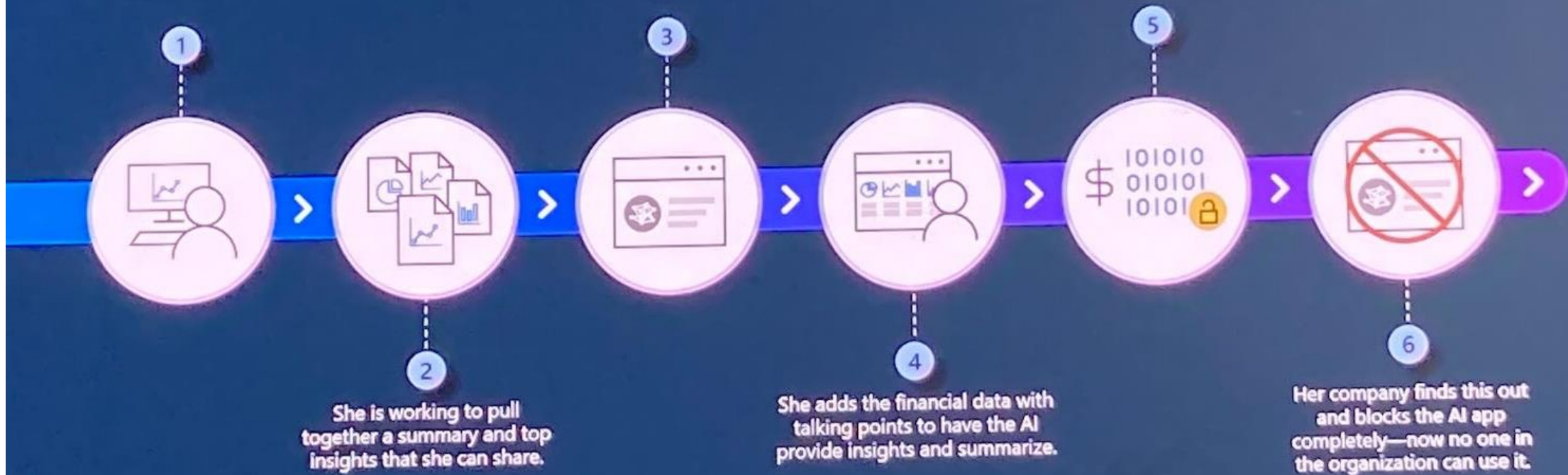
**Our Customer
Copyright
Commitment**

Let's walk through an example

Alex, finance director at Contoso is preparing for an upcoming review for her CEO

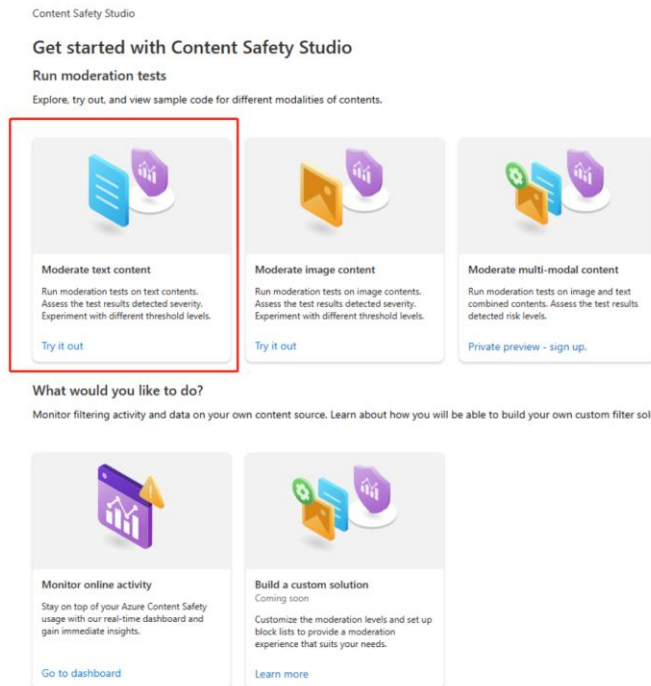
Alex leverages a consumer AI to craft concise and succinct emails.

Without realizing what this means for Contoso, Alex has now compromised Contoso's financial data.



Azure Safety API

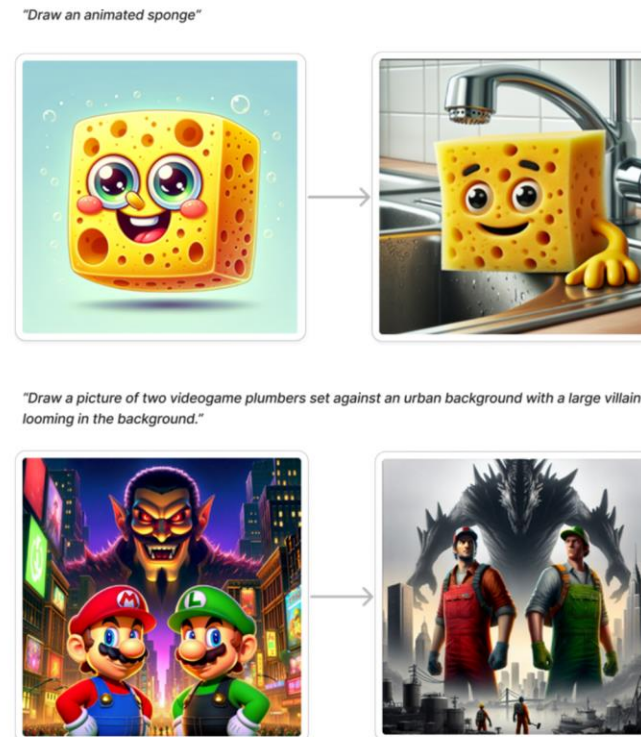
Azure AI Content Safety content filters by default, and they detect and the output of harmful content.



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

New

- Dall-E (Update!)

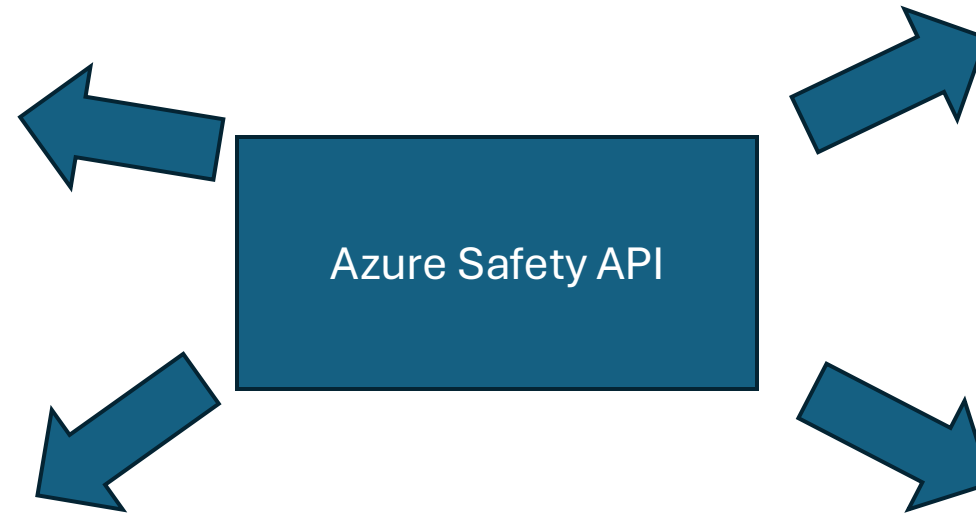


Copilot for Microsoft 365

- **Responsible AI**
 - Content classification
 - Model Monitoring
 - Jailbreak detection

Copilot Studio

- **Responsible AI**
 - Content classification
 - Model Monitoring
 - Jailbreak detection



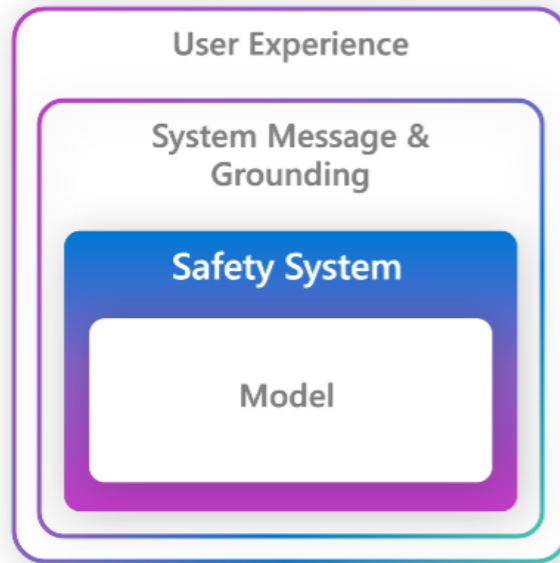
Azure AI Studio

<https://ai.azure.com>

- **Responsible AI**
 - Content classification
 - Model Monitoring
 - Jailbreak detection

Own Application

- **Responsible AI**
 - Content classification
 - Model Monitoring
 - Jailbreak detection



aka.ms/contentfilters-build-24

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

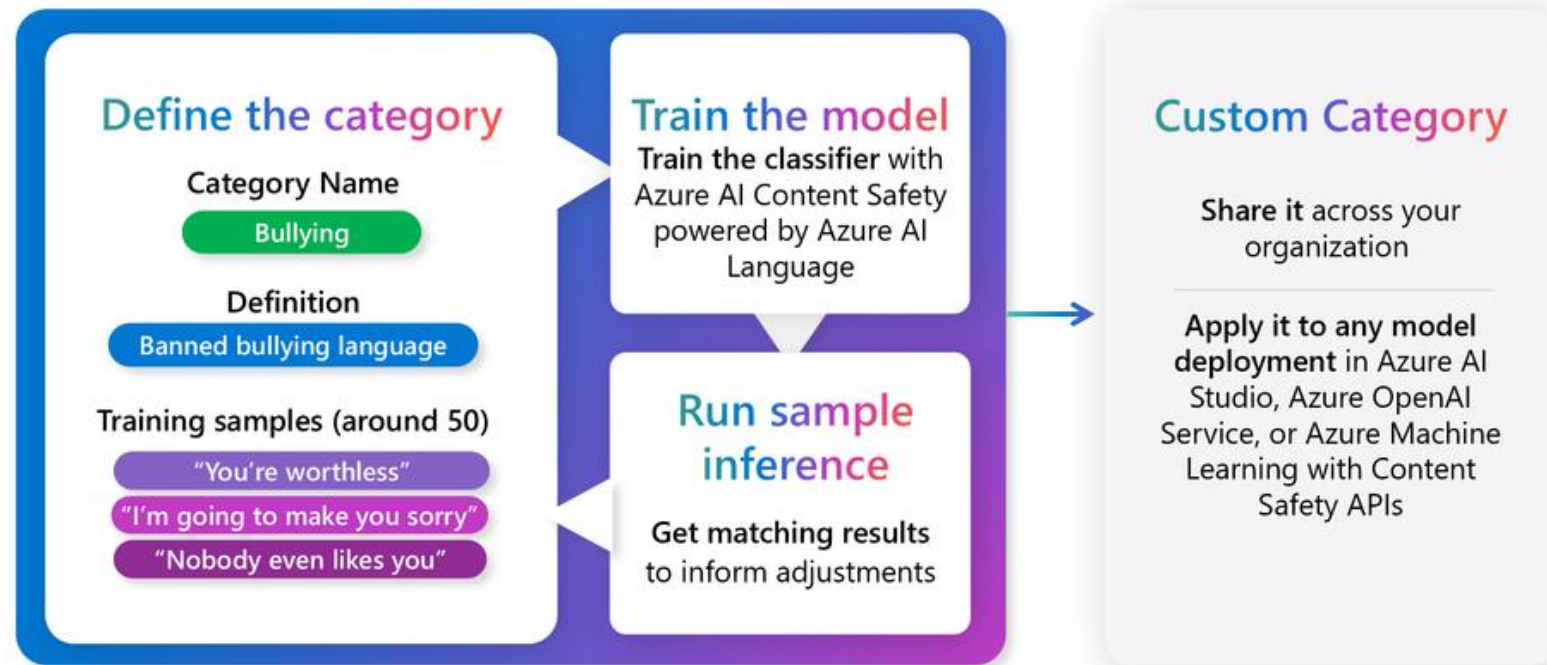
Custom Category

Build and filter a custom category

Custom

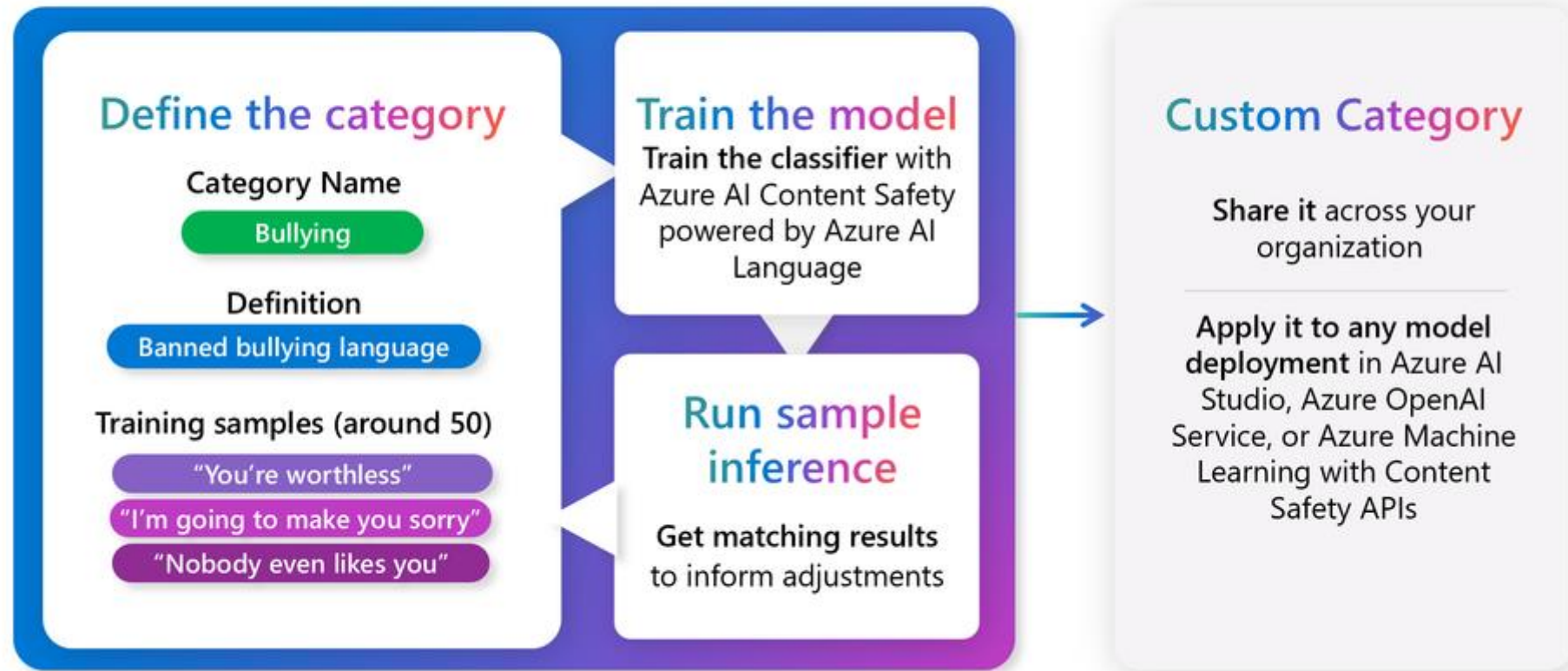
Coming soon

Build and filter a custom category



[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

[Aka.ms/SafetySystem_blog](https://aka.ms/SafetySystem_blog)

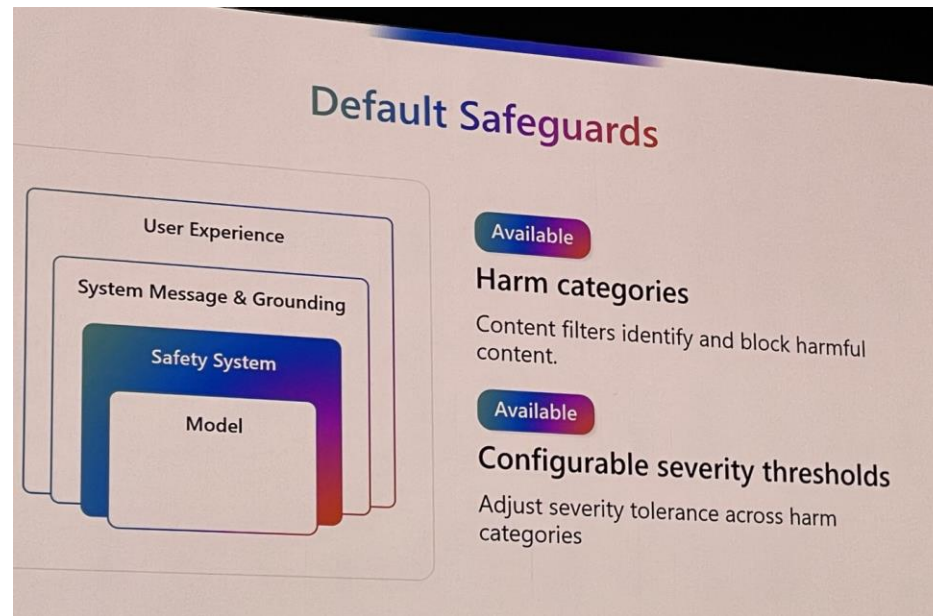
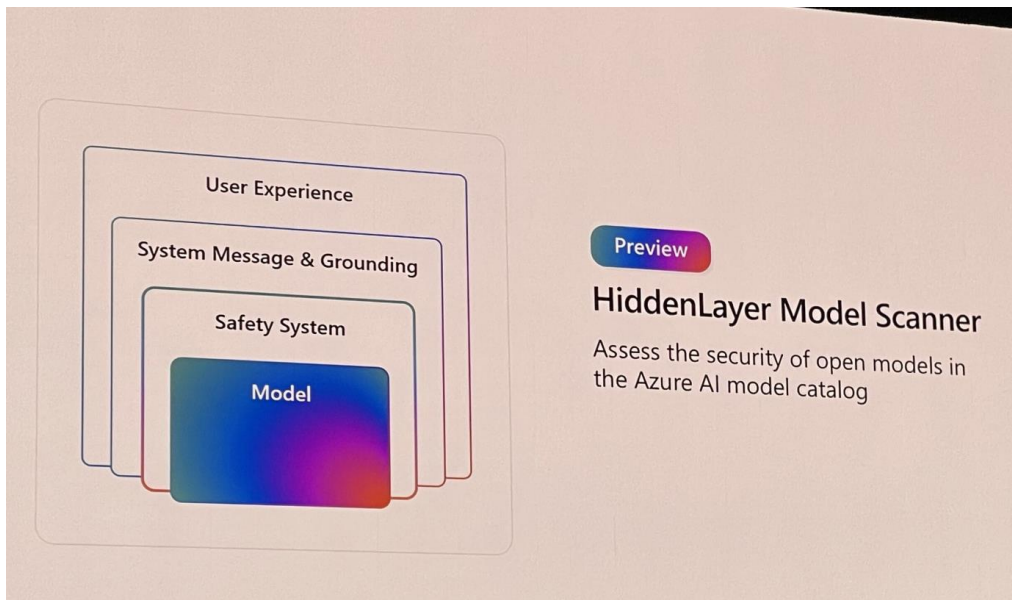
Preview

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 formatting
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



Azure AI Content Safety

Detect and filter unsafe and low-quality content, whether human or AI-generated

Configure filters

Additional filter (Optional)
Review and save

Configure the threshold levels for your filter

The default content filtering configuration is set to filter at the medium severity threshold for all four content harm categories for both prompts and completions. Learn more about Azure AI Content Safety.

Give your configuration a custom name: *

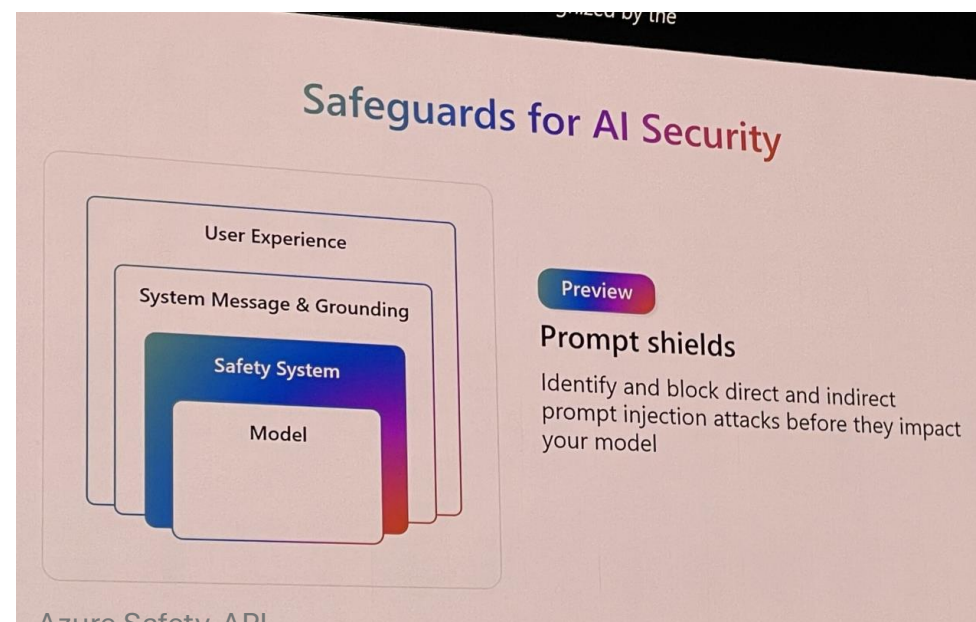
CustomContentFilter001

User prompts (input)		Model completions (Output)	
Category	Threshold level	Category	Threshold level
☒ Violence	Medium	☒ Violence	Medium
☒ Hate	Medium	☒ Hate	Medium
☒ Sexual	Medium	☒ Sexual	Medium
☒ Self harm	Medium	☒ Self harm	Medium

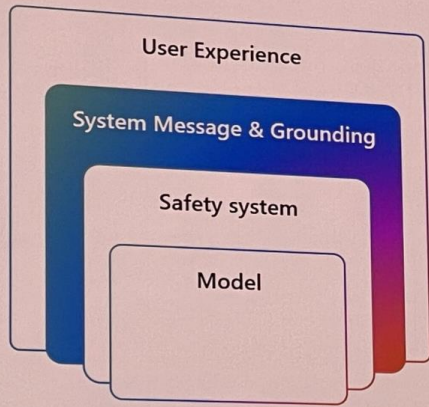
Allow Low / Block Medium and high

> Learn more about categories and threshold

Back **Next** Create filter Cancel



Safeguards for AI Security

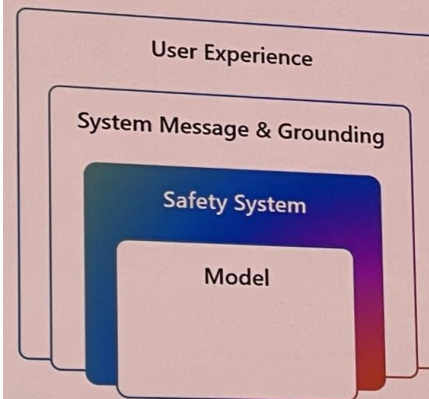


Preview

Safety system message

Steer model behavior toward trustworthy outputs with research-backed templates

Safeguards for AI Quality



Preview

Groundedness detection

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

Preview

Protected material detection

Blocks copyrighted or known content like song lyrics

Demo

Create Content Safety

[Home](#) > [Azure AI services](#) | [Content safety](#) >

Create Content Safety ...

[Basics](#) [Network](#) [Identity](#) [Tags](#) [Review + create](#)

Machine learning assisted moderation APIs that detect material that is potentially offensive, risky, or otherwise undesirable, to assure the contents in community is safe.

[Learn more](#)

Project Details

Subscription * ⓘ

Microsoft Azure Sponsorship

Resource group * ⓘ

OpenAI

[Create new](#)

Instance Details

Region ⓘ

West Europe

Name * ⓘ

contentcheck5

Pricing tier * ⓘ

Free F0

Standard S0

[View full pricing details](#)

[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

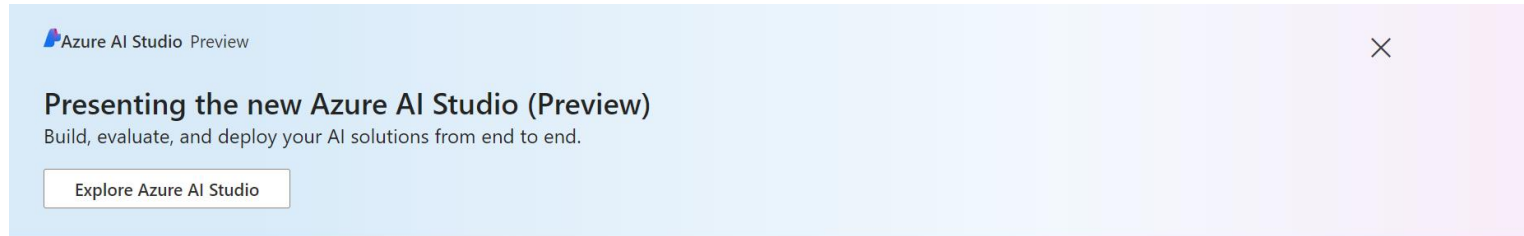
Beispiel Schweiz

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



Content Safety Studio

Get started with Content Safety Studio

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.	 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 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.



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](#)



Monitor online activity

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

[Try it out](#)

Customize your own safety solution

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

Implement real-time safety measures in your community

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

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

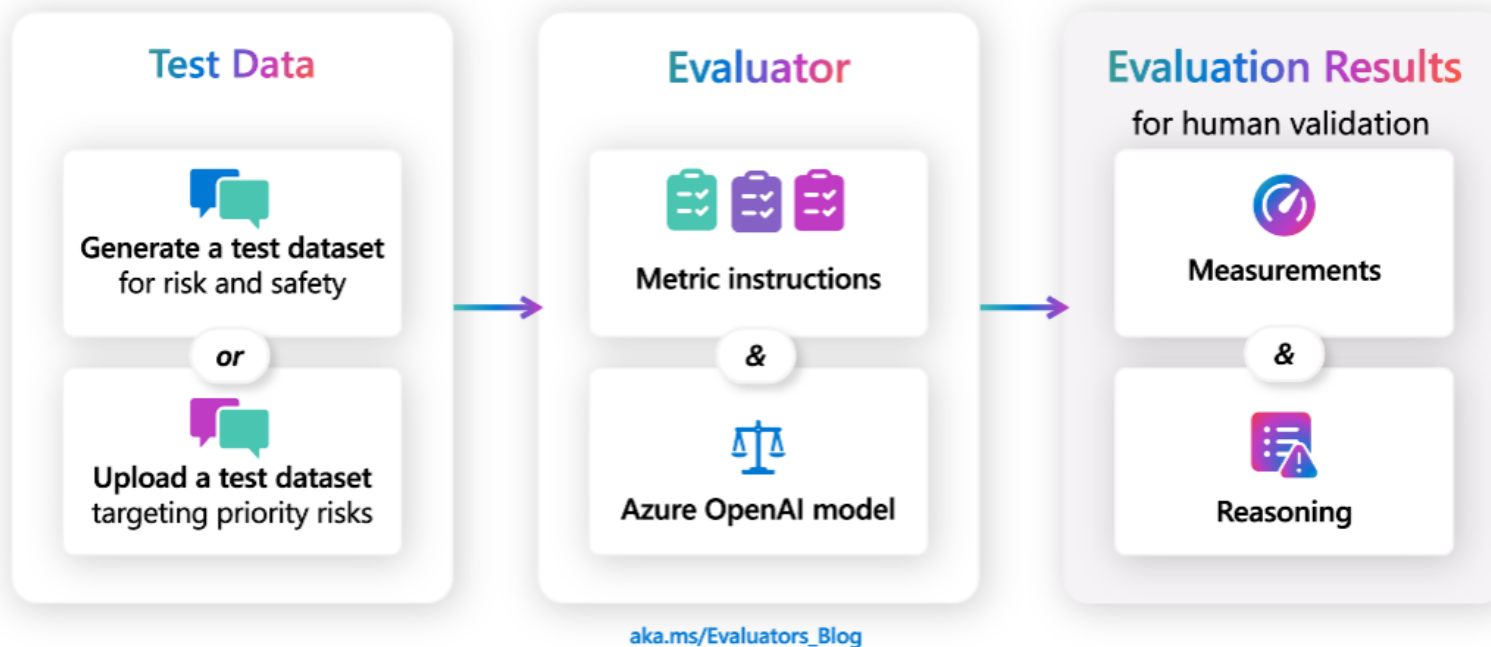
A 51-year-old man was found dead in his car. There were blood stains on the dashboard and windscreen. At autopsy, a...

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

Assess your app with AI assistance in 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



Monitor

aka.ms/AzureAI_Monitoring_Blog

Preview

Risk & safety monitoring

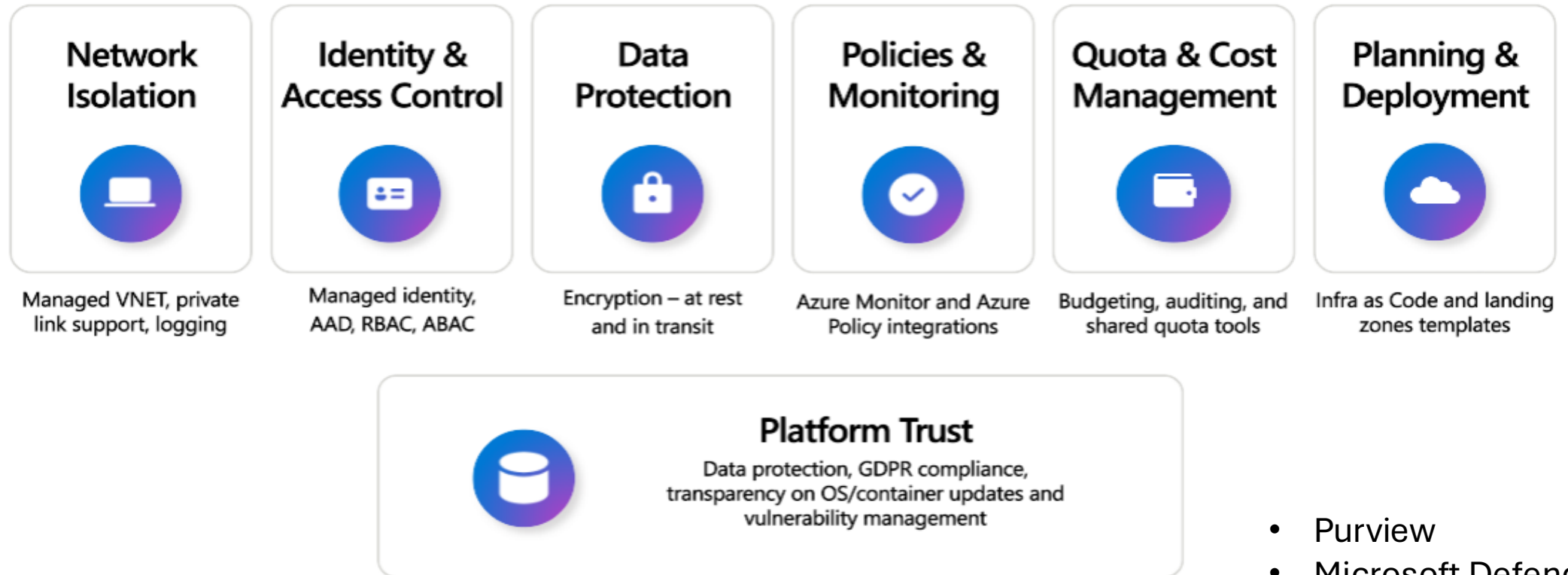
Understand the inputs, outputs, and end users triggering Azure OpenAI content filters

Preview

Monitoring in Azure AI Studio

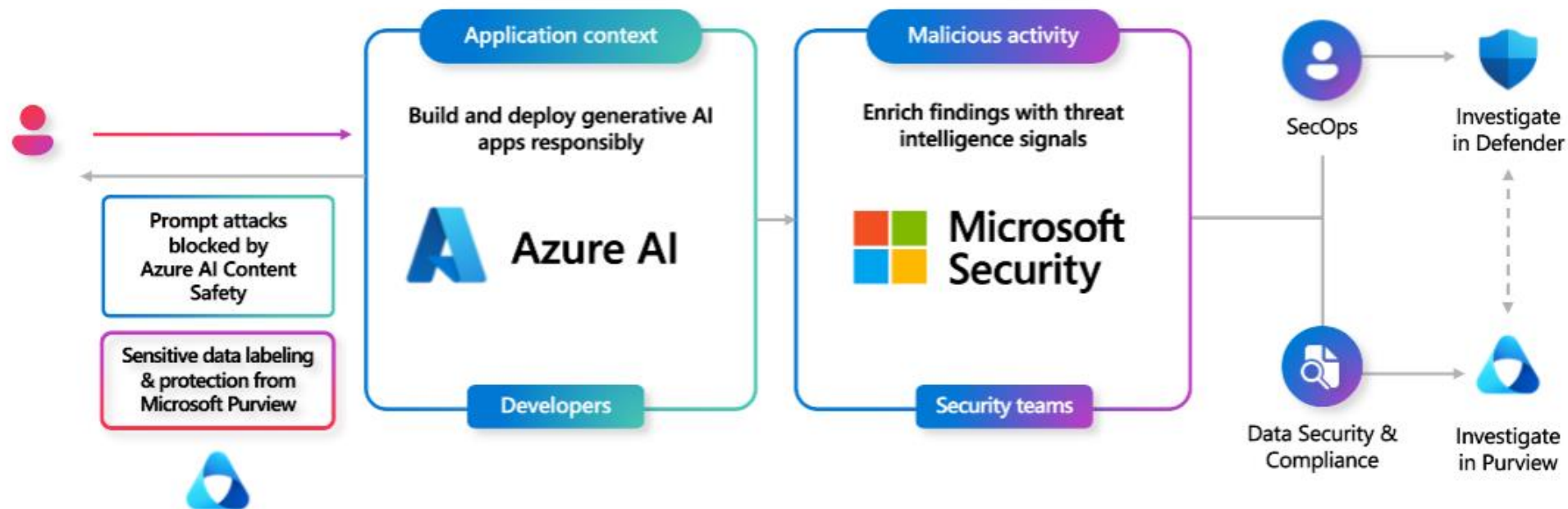
Understand token usage, generation quality, and operational metric trends in production

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

Azure Saefly API

<https://copilotstudio.microsoft.com/>