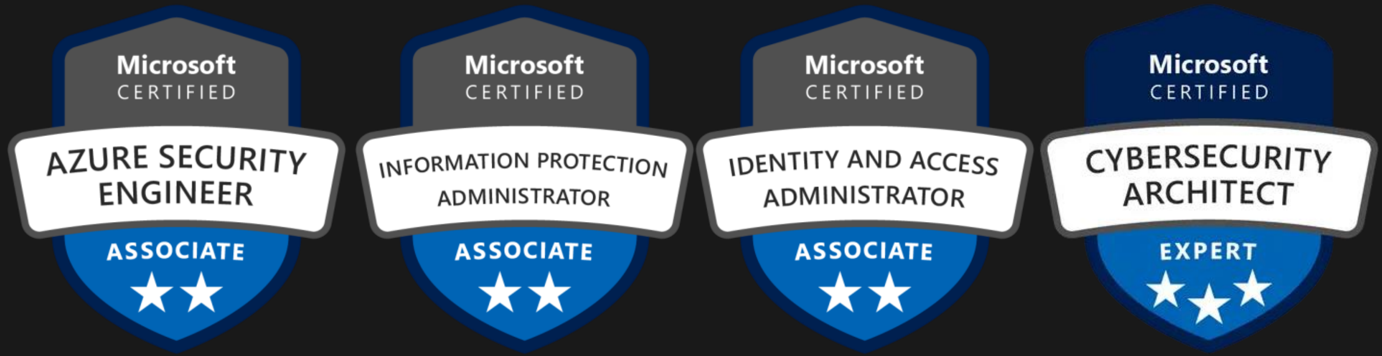
Protecting the Generative-AI content with Azure AI Content Safety

MSUG 23.10.2024

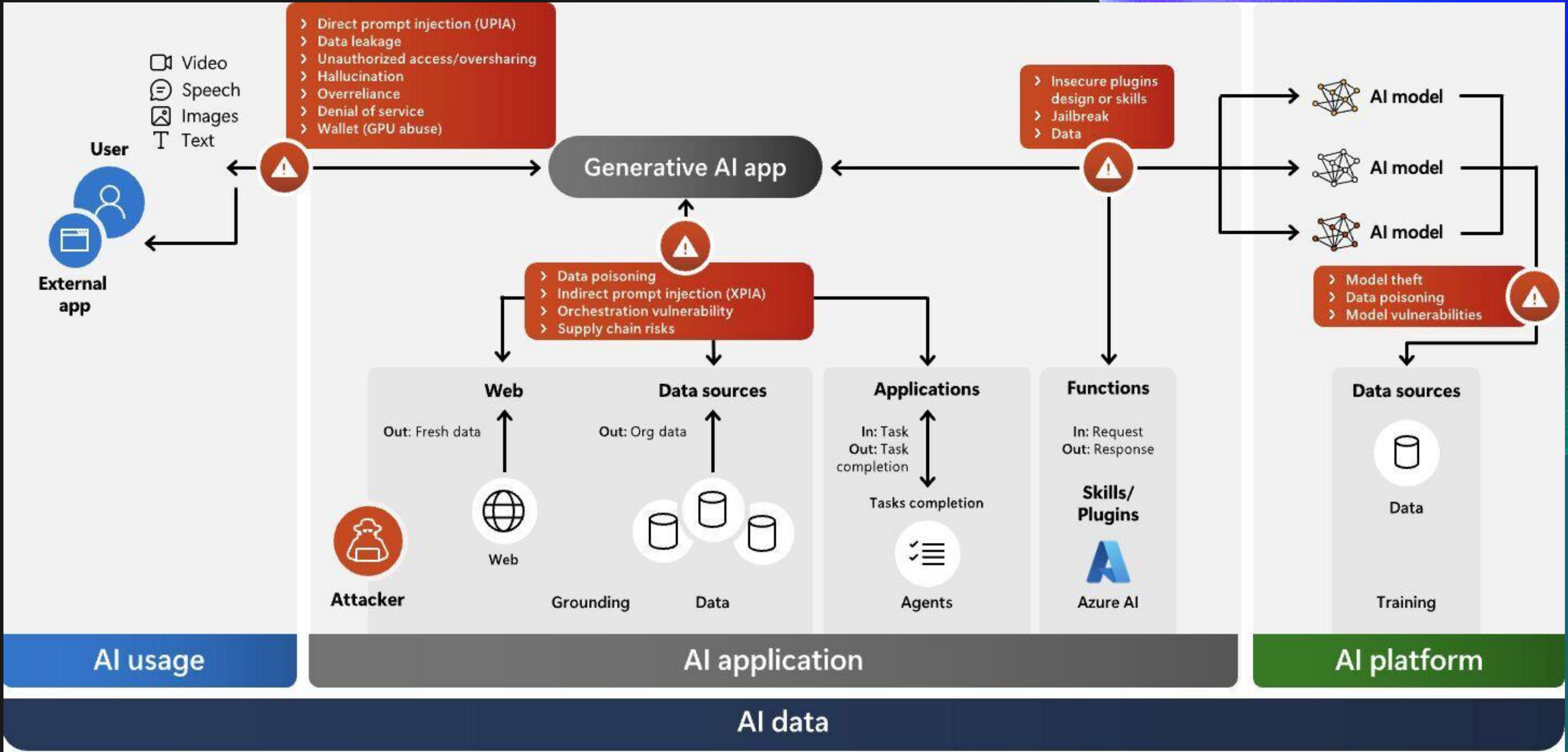
Petrus Vasenius

Introduction

- Petrus Vasenius
- Lead Cloud Security Advisor @ Zure
- Twitter: @PetrusVasenius
- Mastodon/Bluesky: @pvasenius.bsky.social



Security risks with Generative AI



- Direct prompt injection (UPIA)
- Data leakage
- Unauthorized access/oversharing
- Hallucination
- Overreliance
- Denial of service
- Wallet (GPU abuse)

Video
Speech
Images
Text

User

External app

Generative AI app

- Insecure plugins design or skills
- Jailbreak
- Data

AI model
AI model
AI model

- Model theft
- Data poisoning
- Model vulnerabilities

- Data poisoning
- Indirect prompt injection (XPIA)
- Orchestration vulnerability
- Supply chain risks

Web
Out: Fresh data
Web

Data sources
Out: Org data
Grounding
Data

Attacker

Applications
In: Task
Out: Task completion
Tasks completion
Agents

Functions
In: Request
Out: Response
Skills/ Plugins
Azure AI

Data sources
Data
Training

AI usage

AI application

AI platform

AI data

Picture: Microsoft

Azure AI Content Safety

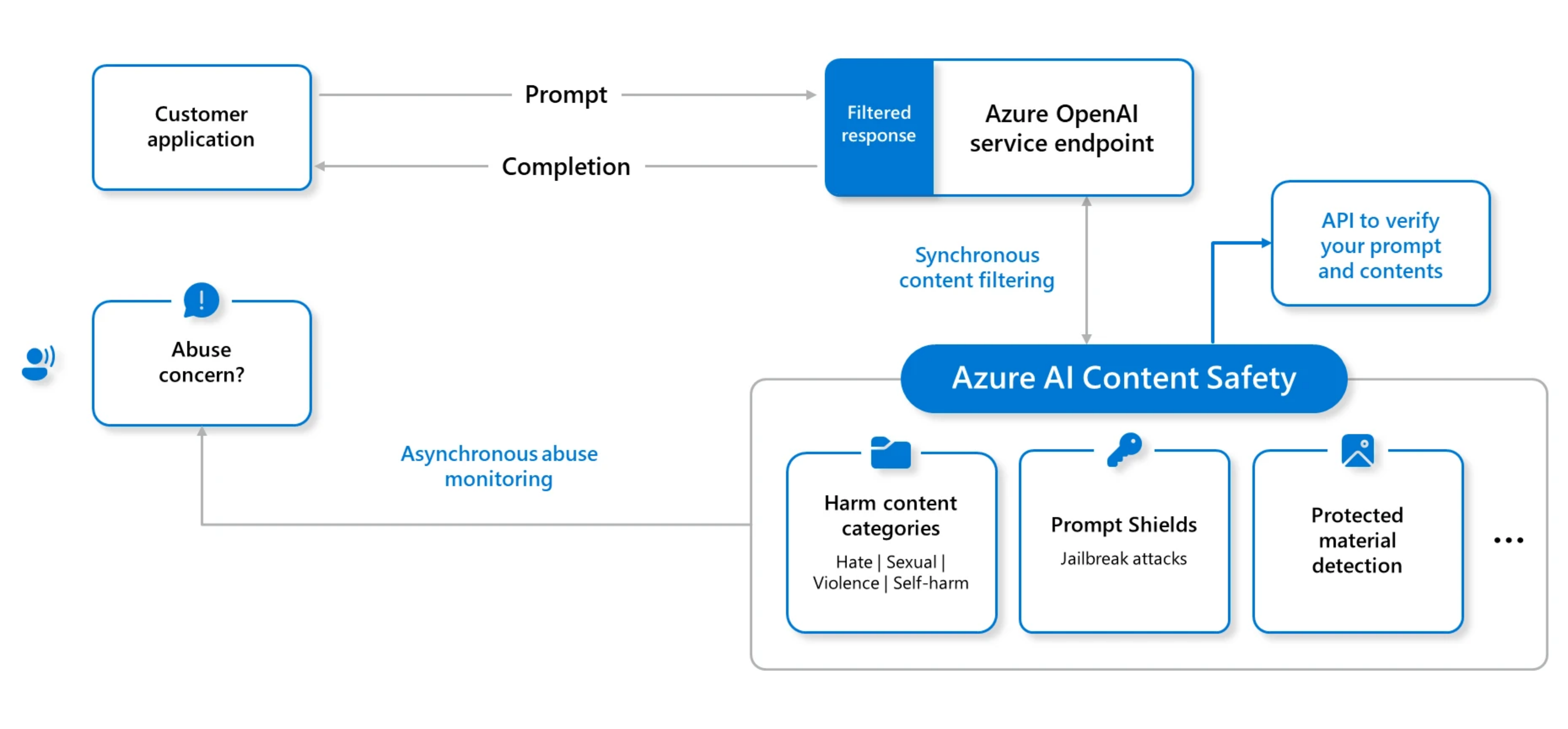
Azure AI Content Safety is an AI service that detects harmful user-generated and AI-generated content in applications and services. Azure AI Content Safety includes text and image APIs that allow you to detect material that is harmful.

- Content filtering software can help your AI app to comply with regulations or maintain the intended environment for your users
- Azure OpenAI Service also provides default content filtering for Models as a Service for OpenAI core models (GPT model series, DALL-E 2/3, for example)
- The "interactive" Content Safety Studio allows you to view, explore, and try out sample code for detecting harmful content across different modalities

Azure AI Content Safety - Key Features

- Harm categories
- Custom Categories
- Prompt Shields
- Groundedness detection (preview)
- Protected material detection (preview)

How it works?



Customer application

Prompt

Completion

Filtered response

Azure OpenAI service endpoint

Synchronous content filtering

API to verify your prompt and contents

Abuse concern?

Asynchronous abuse monitoring

Azure AI Content Safety

Harm content categories

Hate | Sexual | Violence | Self-harm

Prompt Shields

Jailbreak attacks

Protected material detection

...

Picture: Microsoft

Multi-Category Filtering

Category	Description
Hate	Content that promotes discrimination, prejudice, or animosity towards individuals or groups based on race, religion, gender, or other identity defining characteristics.
Sexual	Explicit or suggestive content, including but not limited to, nudity and intimate media.
Self-Harm	Material that depicts, glorifies, or suggests acts of self-injury or suicide.
Violence	Content displaying or advocating for physical harm, threats, or violent actions against oneself or others.

Language support: Azure AI Content Safety models have been specifically trained and tested on the following languages: Chinese, English, French, German, Spanish, Italian, Japanese, Portuguese. However, these features can work in many other languages, but the quality might vary.

The Azure AI Content Safety models for protected material, groundedness detection, and custom categories work with English only.

Severity levels

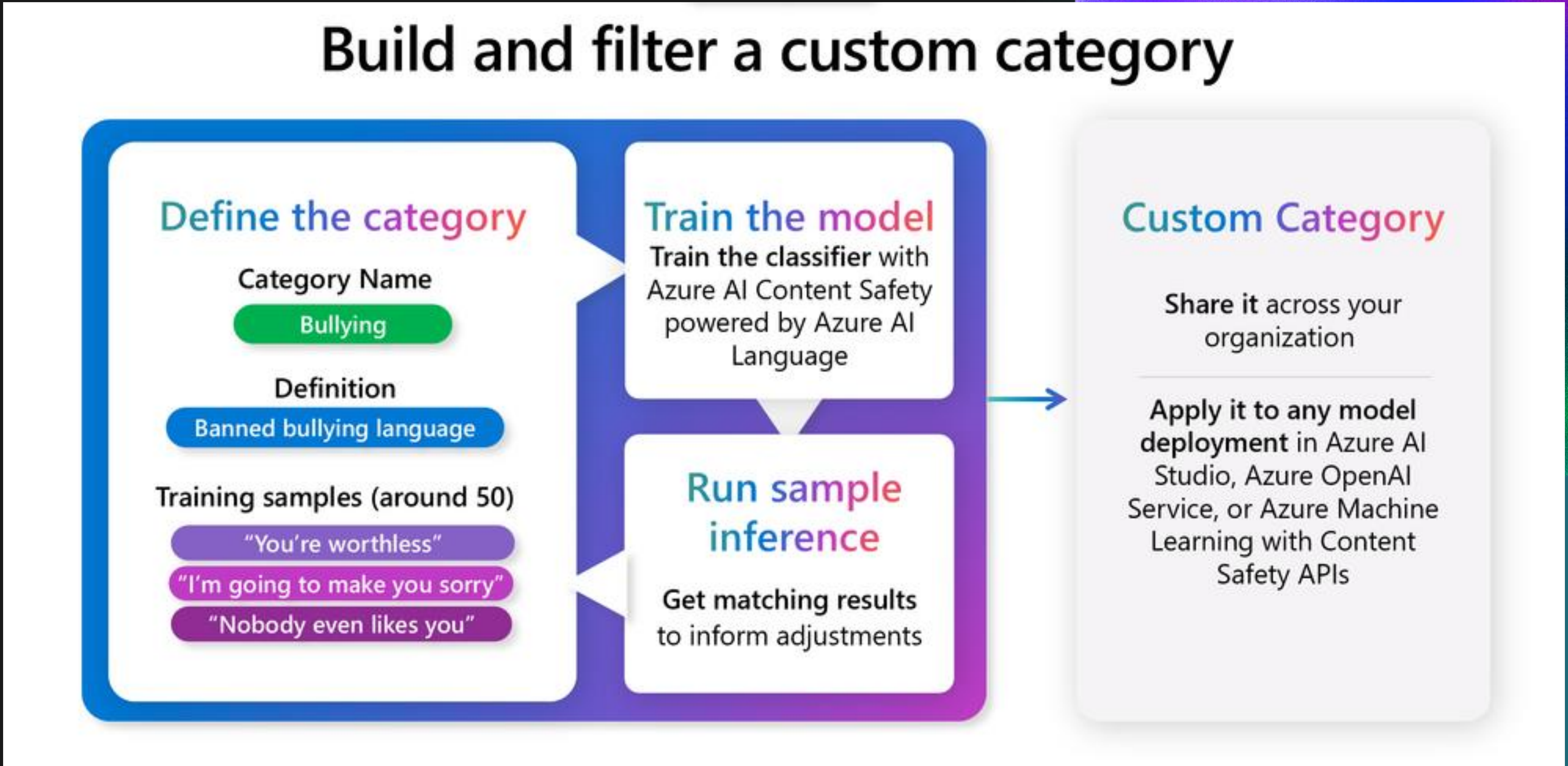
- Every category comes with a severity level rating
- Severity level is meant to indicate the severity of the consequences of showing the flagged content
  - Text model supports the scale of 0-7
  - Image model supports only severities 0, 2, 4 and 6
  - Multimodal model supports the full 0-7 severity scale

Custom categories

- Custom categories enables you to create your own customized classifier based on your specific needs for content filtering and AI safety whether you want to detect sensitive content, moderate user-generated content, or comply with local regulations.

- Use Custom categories to train and deploy your own custom content filter with ease and flexibility.

- Two deployment options:
  - Standard or Rapid



Build and filter a custom category

Define the category

Category Name

Bullying

Definition

Banned bullying language

Training samples (around 50)

"You're worthless"

"I'm going to make you sorry"

"Nobody even likes you"

Train the model

Train the classifier with Azure AI Content Safety powered by Azure AI Language

Run sample inference

Get matching results to inform adjustments

Custom Category

Share it across your organization

Apply it to any model deployment in Azure AI Studio, Azure OpenAI Service, or Azure Machine Learning with Content Safety APIs

Picture: Microsoft

Content Safety Prompt Shields

- If a prompt is detected as potentially harmful or likely to lead to policy-violating outputs, the shield blocks the prompt and alerts the user to modify their input
- Shield targets User Prompt injection attacks, where users deliberately tries to exploit system vulnerabilities to elicit unauthorized behavior from the AI app
- It can also be used to protect the AI app from document attacks, where attackers might embed hidden instructions in external documents to gain unauthorized control over the AI app session

Groundedness Detection (Preview)

- The Groundedness detection API detects whether the text responses of large language models (LLMs) are grounded in the source materials provided by the users

- The API includes a correction feature that automatically corrects any detected ungroundedness in the text based on the provided grounding sources
    - When the correction feature is enabled, the response includes a corrected text field that presents the corrected text aligned with the grounding sources

Protected material detection (Preview)

- Protected material text and code APIs flags known content (song lyrics, articles, known GitHub repositories, software libraries etc.) that might be output by LLM models

- By detecting and preventing the display of protected material, organizations can ensure compliance with intellectual property laws, maintain content originality and protect the reputation

Note! The content safety service's code scanner/indexer is only current through November 6, 2021. Code that was added to GitHub after this date will not be detected.

Content Safety pricing

Instance	Features	Price
Free – Web	Text Prompt Shields Protected material detection Groundedness detection	5000 text records per month
	Image	5000 images per month
Standard – Web	Text Prompt Shields Protected material detection Groundedness detection	0.34 € per 1000 text records
	Image	0.68 € per 1000 images

Azure AI Content Safety use-cases

- User prompts submitted to a generative AI service
- Content produced by generative AI models
- Online marketplaces that moderate product catalogs and other user-generated content
- Gaming companies that moderate user-generated game artifacts and chat rooms
- Social messaging platforms that moderate images and text added by their users
- Enterprise media companies that implement centralized moderation for their content
- Education solution providers filtering out content that is inappropriate for students and educators

Content Safety Overview

From Content Safety Studio & Azure AI Studio

Final thoughts

- Set the threshold values according to the business use case of the AI application
- Experiment carefully with the risk categories before applying them into the production environment
- The custom category (standard) model training requires time, don’t rush with it and make sure you have a balanced dataset with both positive and negative examples
- Recommend to start with smaller PoC first, and then apply it to larger extent
- Don’t forget other AI security features, when utilizing AI in your organization
  - For example, AI security posture management

Q/A

Thank you!

ZURE