

CREATING RESPONSIBLE AI SOLUTIONS USING AZURE AI CONTENT SAFETY

Naveen Kumar M



ABOUT ME

Naveen Kumar M



- **Data Engineer**
- **13 + Years of Experience**
- **2 x C# Corner MVP**
- **4 x Azure Certified**
- **Technical Blogger**
- **Speaker & Mentor**



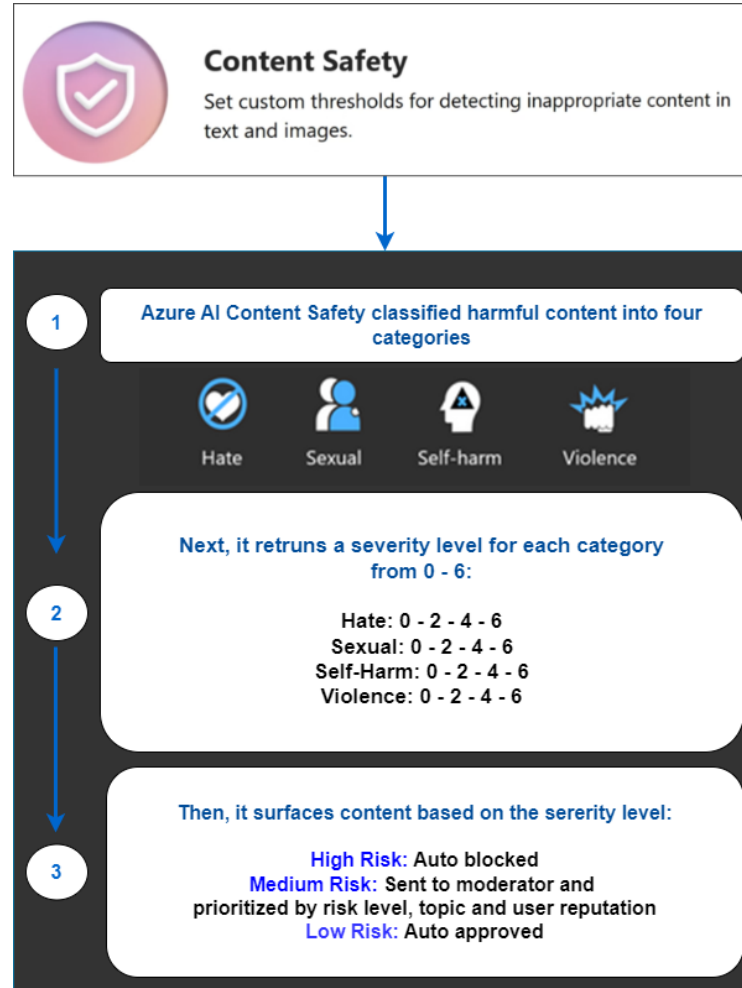
AGENDA

- Regulate Text Content
- Regulate Image Content
- Regulate Multimodal Content
- Detect Protected Material in Text
- Detect Protected Material in Code
- Utilize Prompt Shields
- Monitor Online Activities
- Implement Responsible AI in Content Safety Studio

AZURE AI CONTENT SAFETY

- Azure AI Content Safety service detects harmful user-generated and AI-generated content in applications and services.
- It provides text and image APIs that allow you to detect harmful or inappropriate material.

AZURE AI CONTENT SAFETY



AZURE AI CONTENT SAFETY API'S

Text Detection API	Scans text for sexual content, violence, hate, and self-harm with multi-severity levels.
Image Detection API	Scans images for sexual content, violence, hate, and self-harm with multi-severity levels.
Text Blocklist Management API	You can create blocklists of terms to use with the Text API.

CONTENT SAFETY STUDIO

Azure AI | 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.

[Try it out](#)



Groundedness detection

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

[Private preview - sign up.](#)



Protected material detection for text

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

[Try it out](#)



Protected material detection for code

PREVIEW

Run tests on code generated by LLM and identify whether the code already exists in GitHub repo.

[Try it out](#)



Prompt Shields

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

[Try it out](#)

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.



Moderate multimodal content

PREVIEW

Run moderation tests on image and text

SEVERITY LEVELS

 **Configure filters**  Use blocklist

 **View code**

Set the severity thresholds for each category and select Run test to see how the results change.

[Reset to default](#)

Severity ⓘ

SAFE

LOW

MEDIUM

HIGH

Violence ⓘ



Self-harm ⓘ



Sexual ⓘ



Hate ⓘ



Configure filters Use blocklist

 **View code**

Set the Severity thresholds for each category. Content with a severity level less than the threshold will be allowed. [Learn more about categories and threshold](#)

Category

Threshold level



Violence

Medium



Allow Low / Block Medium and High



Self-harm

Medium



Allow Low / Block Medium and High



Sexual

Medium



Allow Low / Block Medium and High



Hate

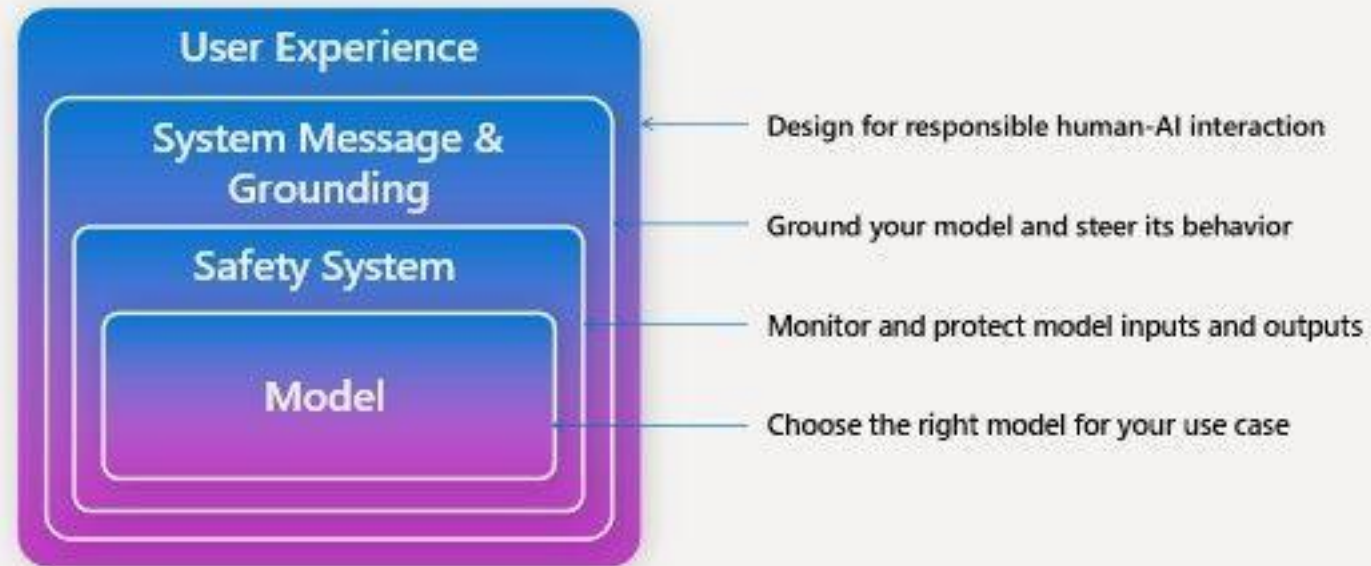
Medium



Allow Low / Block Medium and High

RESPONSIBLE AI FOR AZURE AI FOUNDRY

Risk mitigation layers



PREREQUISITES



Python 3.7
library

Azure
Subscription

Azure AI
Content Safety
Resource

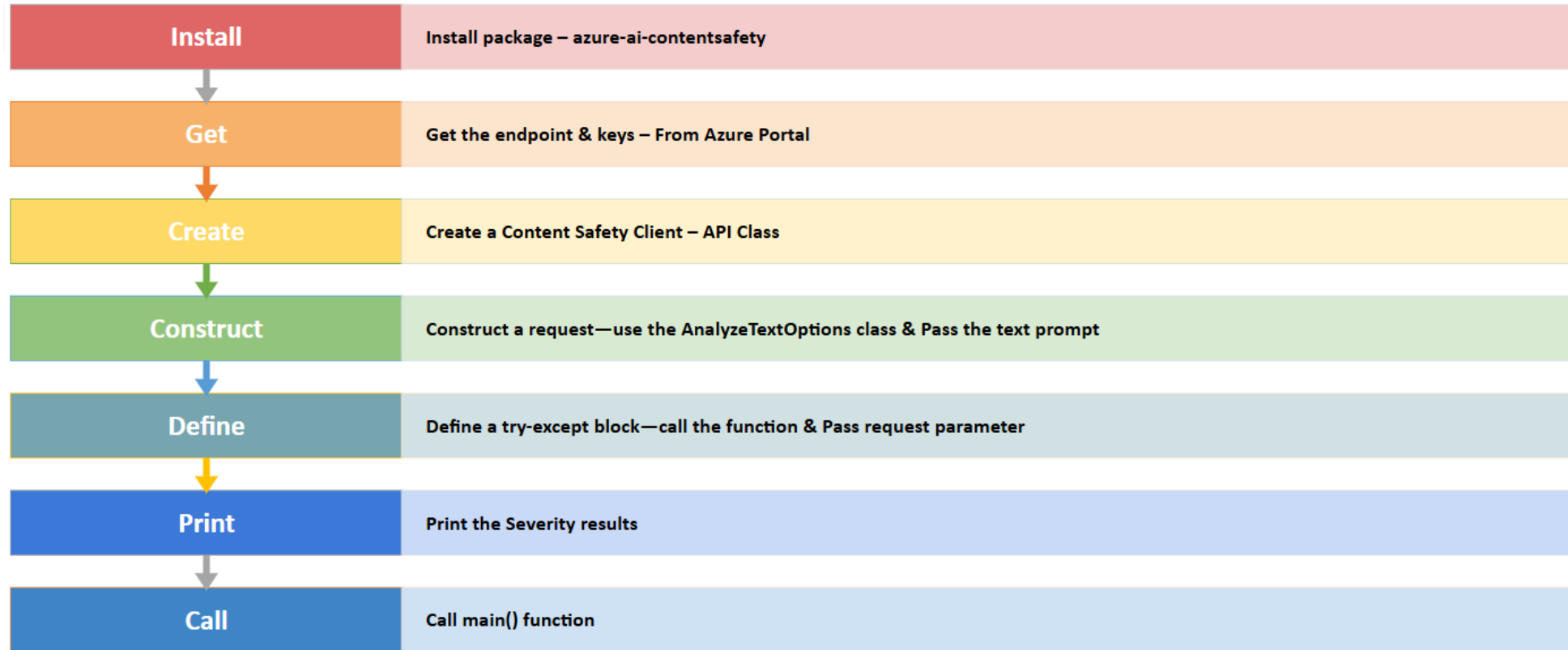
CLIENT LIBRARY FOR PYTHON

- Azure AI Content Safety Client Library for Python

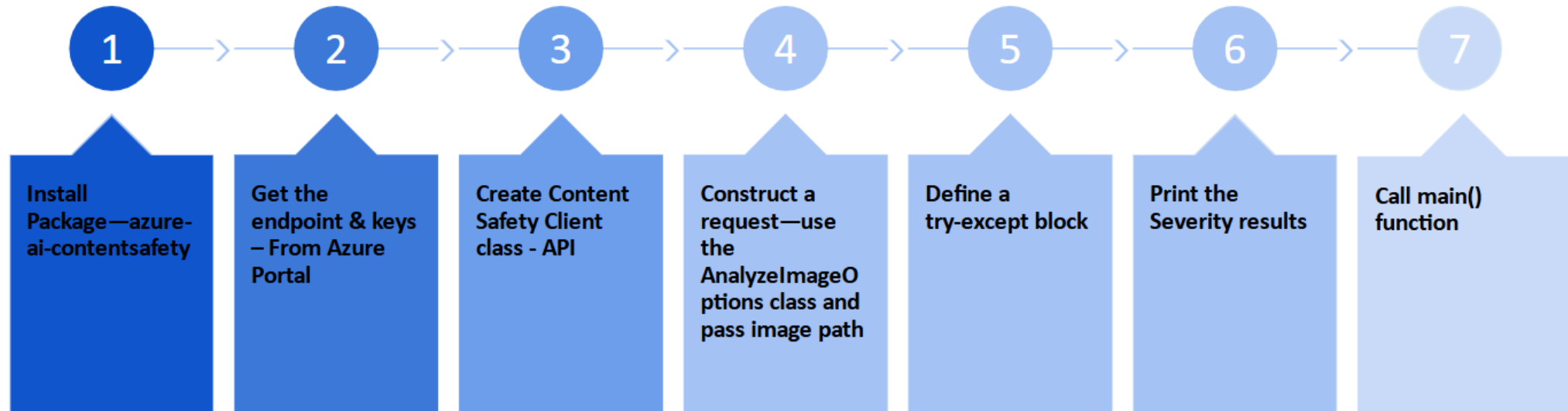
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS AZURE COMMENTS

```
PS C:\Users\navin> pip install azure-ai-contentssafety
```

CREATING TEXT MODERATION



CREATING IMAGE MODERATION



CONTENT SAFETY USE CASES



Online
Marketplaces



Gaming



Social Messaging
Platforms



Enterprise Media
Companies



Education
Solution Providers

CONTENT SAFETY USE CASES



Online marketplaces: Moderate product catalogs and other user-generated content.



Gaming companies: Moderate user-generated game artifacts and chat rooms.

Social messaging platforms: Moderate images and text added by their users.



Enterprise media: Implement centralized moderation for their content.



K-12 education: Filtering out content that is inappropriate for students and educators.

PRICING

Region:

East US

▼

Currency:

United States – Dollar (\$) USD

▼

Instance	Features	Price
Free – Web	Text Prompt Shields Protected material detection Groundedness detection	5,000 text records per month ¹
	Image Multimodal	5,000 images per month
Standard – Web	Text Prompt Shields Protected material detection Groundedness detection	\$0.38 per 1,000 text records ¹
	Image Multimodal	\$0.75 per 1,000 images

¹A text record in the S tier contains up to 1,000 characters as measured by Unicode code points. If a text input into the Content Safety API is more than 1,000 characters, it counts as one text record for each unit of 1,000 characters. For instance, if a text input sent to the API contains 7,500 characters, it will count as 8 text records. If a text input sent to the API contains 500 characters, it will count as 1 text record.






Naveen Kumar M

Data Engineer | SQL / ETL Lead |
C# Corner MVP







TECHNOLOGY INSIGHTS

Tech Time
WITH NAVEEN

[SUBSCRIBE NOW](#)



Tech Time with Naveen

@ttwithnaveen · 204 subscribers · 115 videos

Tech Time with Naveen: Your Ultimate Destination for New Technologies and Career Growth...more

[linkedin.com/in/naveenkumarm17](#) and 2 more links

[Customize channel](#) [Manage videos](#)

Home Videos Shorts Playlists Community

Latest Popular Oldest



Festive Tech Calendar 2024: How to Build Personalized Chatbots Using Azure AI

19 views · 1 month ago



Create Your Own Chatbot With Azure AI Service | Hands-On Demo #chatbot #azure...

141 views · 2 months ago



Crafting Custom Copilots with Azure AI Studio | Hands-On Demo #copilot #azure #ai

71 views · 2 months ago



Introducing The DP-700 Fabric Data Engineer Certification - A Game-changer!

275 views · 3 months ago

@ttwithnaveen



 Azure AI

Thank you!

For exploring Azure AI Content Safety



AI

AI

TATIE AI

AZURE AI