CREATING RESPONSIBLE AI SOLUTIONS USING AZURE AI CONTENT SAFETY

Naveen Kumar M

ABOUT ME

LinkedIn

Naveen Kumar M

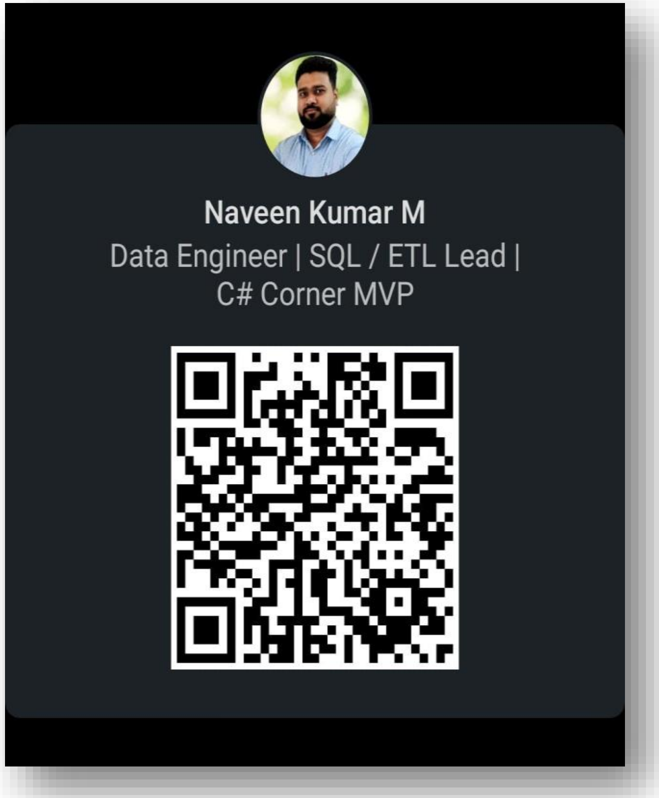
- Data Engineer
- 13 + Years of Experience
- 2 x C# Corner MVP
- 4 x Azure Certified
- Technical Blogger
- Speaker & Mentor

C# Corner Most Valuable Professional

Microsoft CERTIFIED

FABRIC ANALYTICS ENGINEER

ASSOCIATE



Naveen Kumar M

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

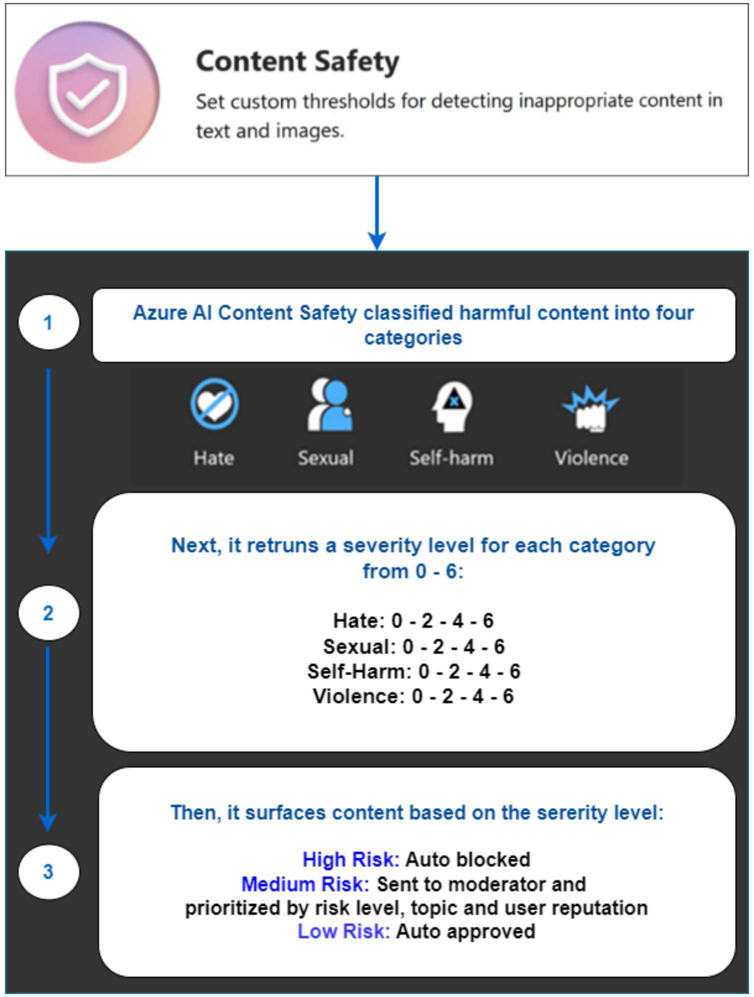
AGENDA

- Moderate Text Content
- Moderate Image Content
- Moderate Multimodal Content
- Protected Material Detection for Text
- Protected Material Detection for Code
- 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



Content Safety

Set custom thresholds for detecting inappropriate content in text and images.

1 Azure AI Content Safety classified harmful content into four categories

Hate
Sexual
Self-harm
Violence

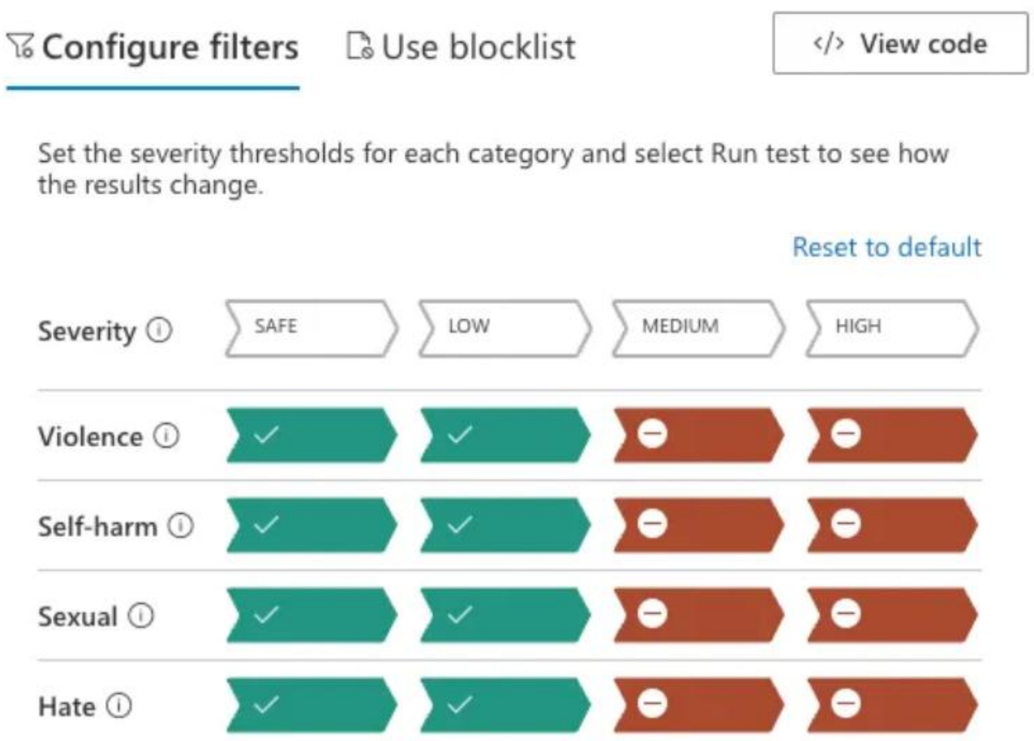
2 Next, it retruns a severity level for each category from 0 - 6:

Hate: 0 - 2 - 4 - 6
Sexual: 0 - 2 - 4 - 6
Self-Harm: 0 - 2 - 4 - 6
Violence: 0 - 2 - 4 - 6

3 Then, it surfaces content based on the sererity level:

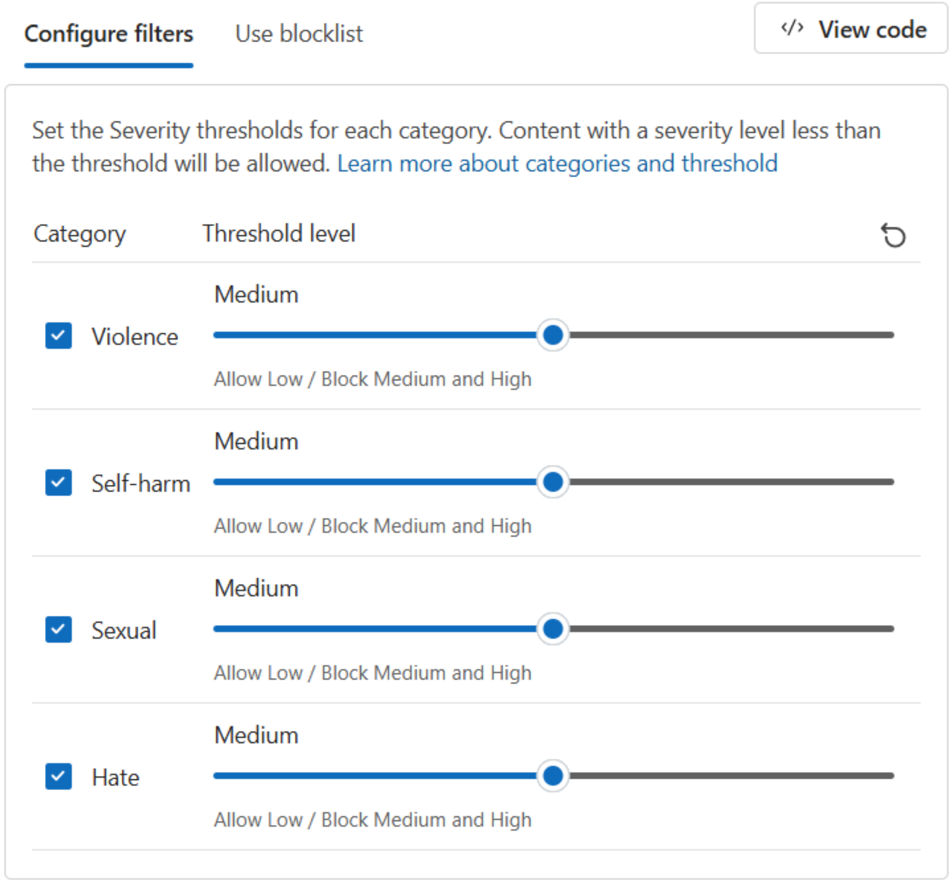
High Risk: Auto blocked
Medium Risk: Sent to moderator and prioritized by risk level, topic and user reputation
Low Risk: Auto approved

SEVERITY LEVELS



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

Severity ⓘ	SAFE	LOW	MEDIUM	HIGH
Violence ⓘ	✓	✓	⊖	⊖
Self-harm ⓘ	✓	✓	⊖	⊖
Sexual ⓘ	✓	✓	⊖	⊖
Hate ⓘ	✓	✓	⊖	⊖



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

AZURE AI CONTENT SAFETY API’S

Text Detection API

Scans text for sexual content, violence, hate, and self-harm with multi-severity levels.

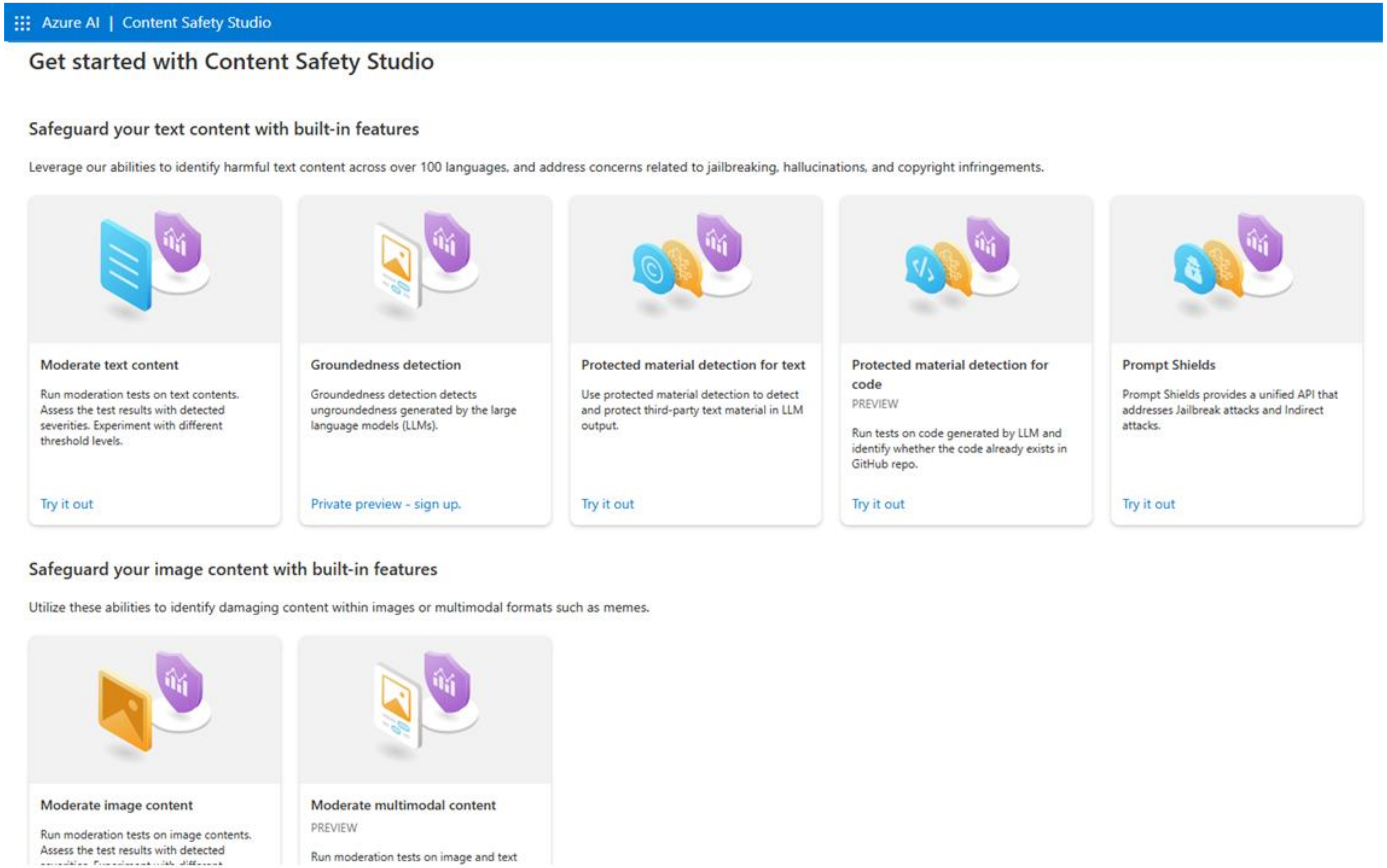
Image Detection API

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

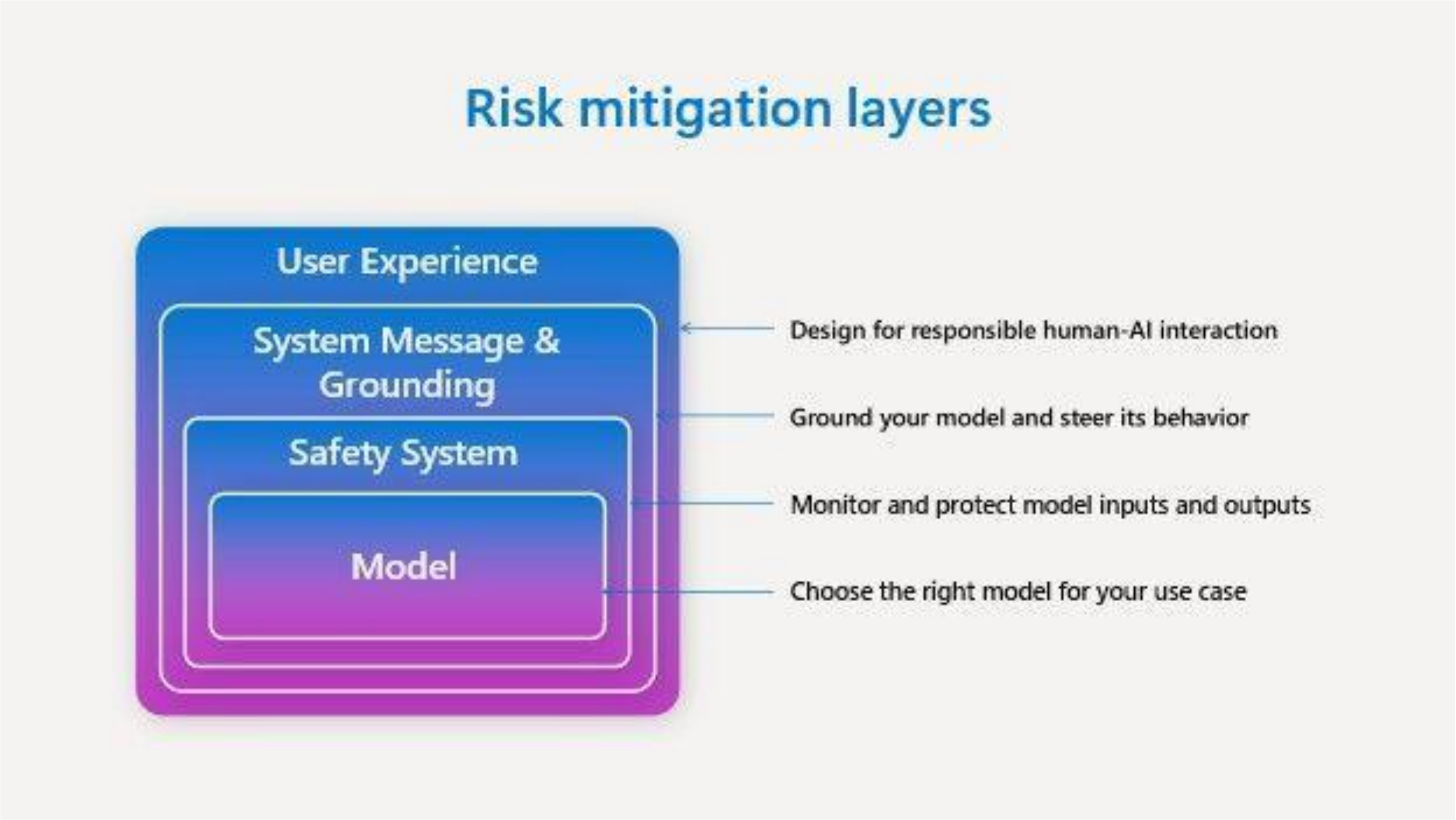
Run moderation tests on image contents. Assess the test results with detected severities. Experiment with different

Moderate multimodal content

PREVIEW

Run moderation tests on image and text

RESPONSIBLE AI FOR AZURE AI FOUNDRY



Risk mitigation layers

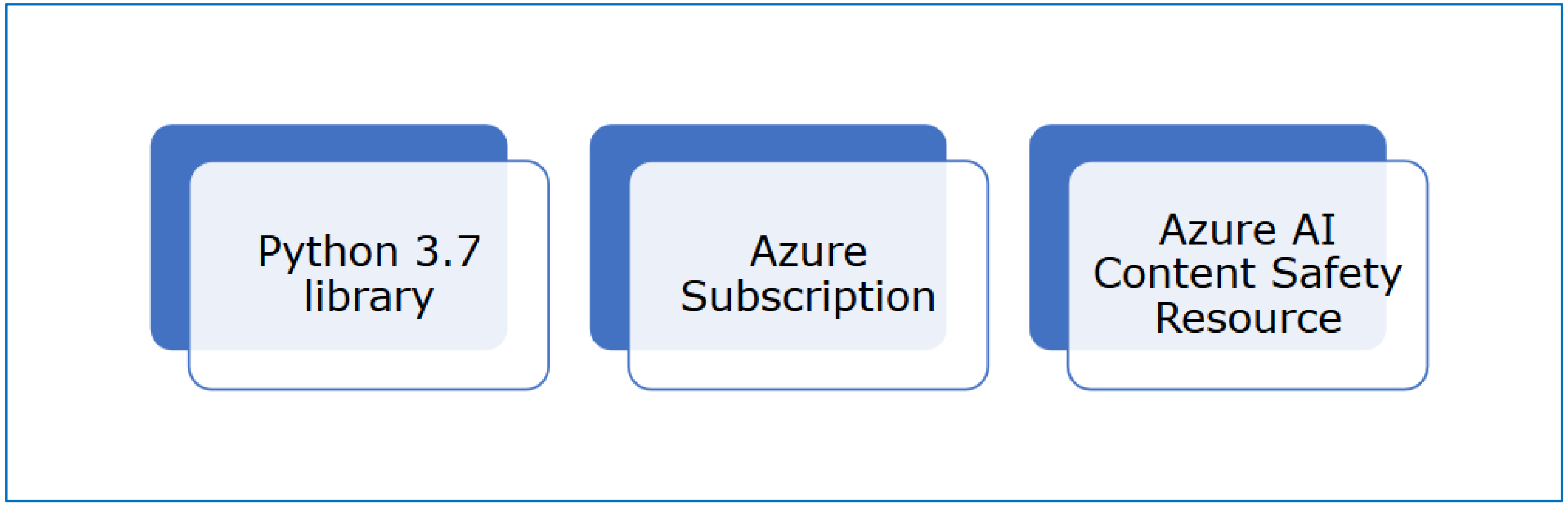
Design for responsible human-AI interaction

Ground your model and steer its behavior

Monitor and protect model inputs and outputs

Choose the right model for your use case

PREREQUISITES



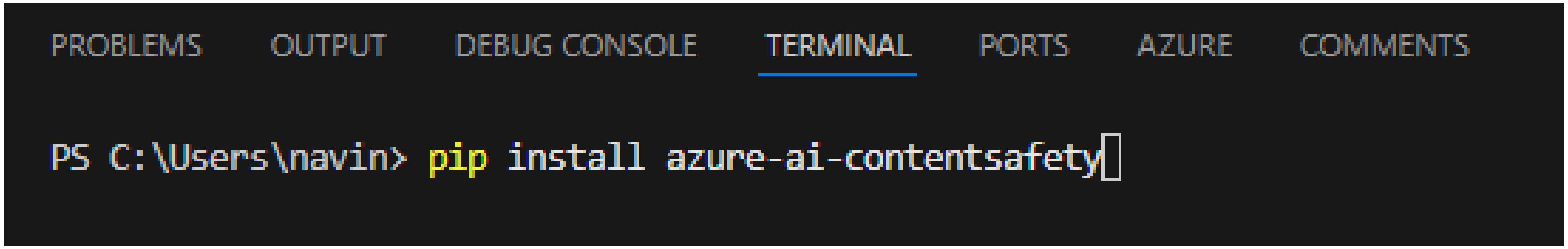
Python 3.7 library

Azure Subscription

Azure AI Content Safety Resource

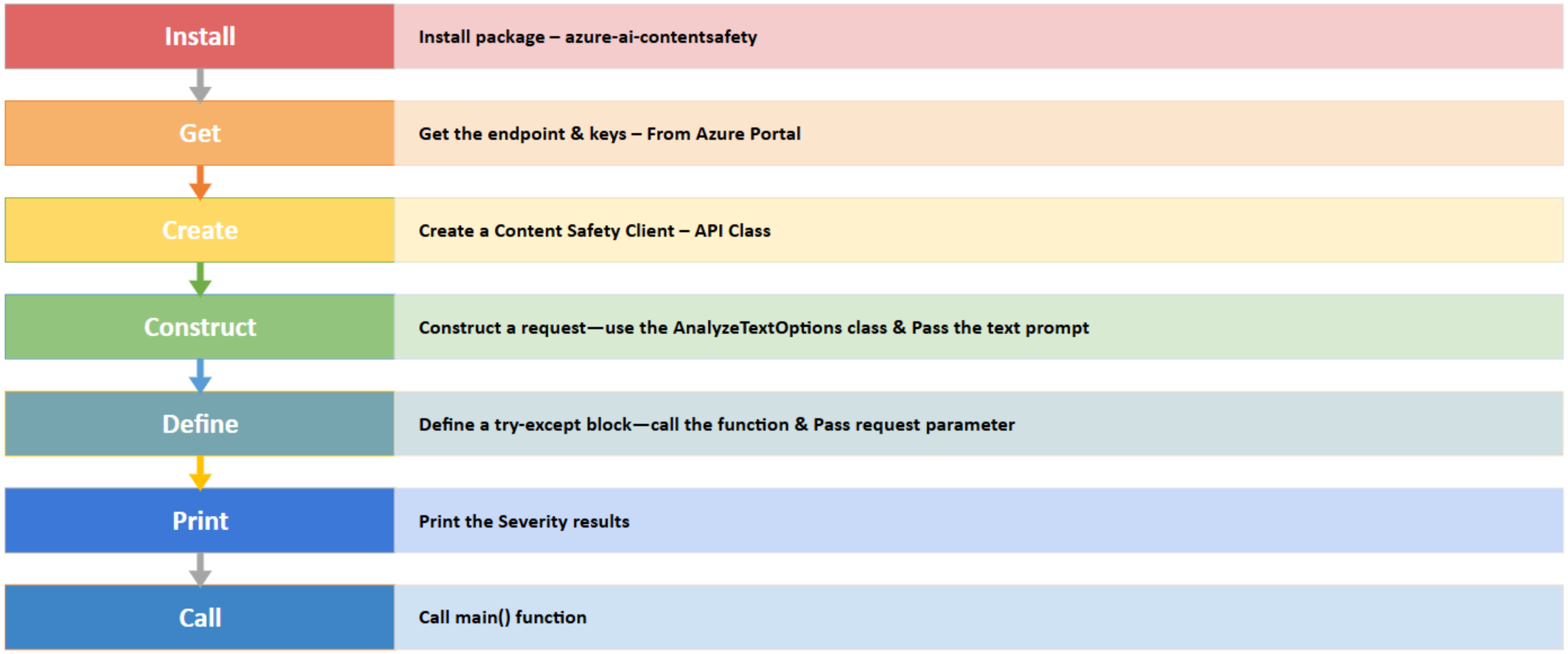
CLIENT LIBRARY FOR PYTHON

- Azure AI Content Safety Client Library for Python



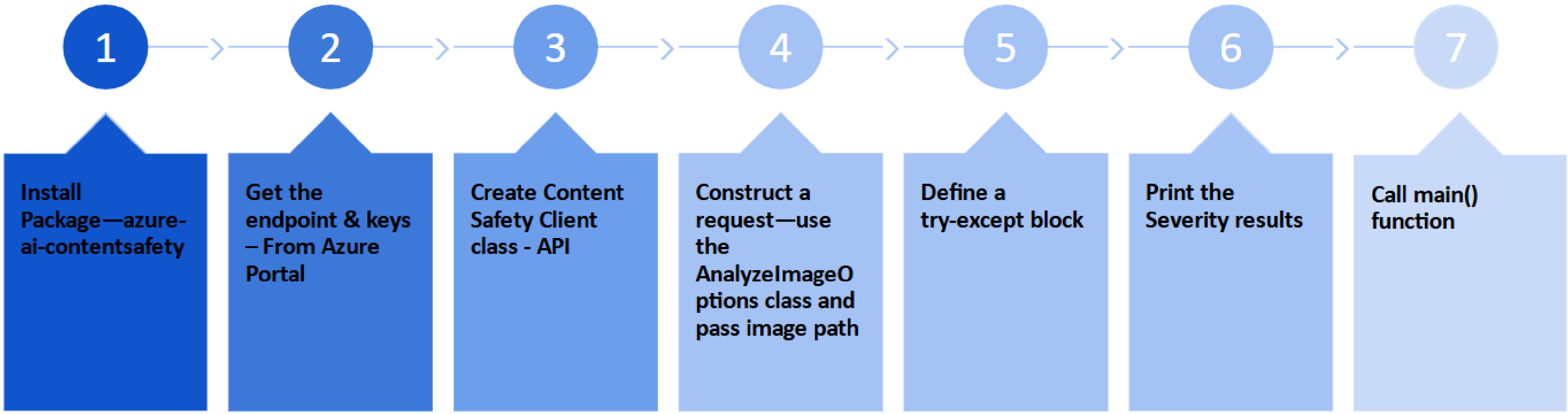
```
PS C:\Users\navin> pip install azure-ai-contentsafety
```

CREATING TEXT MODERATION



Install	Install package – azure-ai-contentsafety
Get	Get the endpoint & keys – From Azure Portal
Create	Create a Content Safety Client – API Class
Construct	Construct a request—use the AnalyzeTextOptions class & Pass the text prompt
Define	Define a try-except block—call the function & Pass request parameter
Print	Print the Severity results
Call	Call main() function

CREATING IMAGE MODERATION



1. Install Package—azure-ai-contentsafety
2. Get the endpoint & keys – From Azure Portal
3. Create Content Safety Client class - API
4. Construct a request—use the AnalyzeImageOptions class and pass image path
5. Define a try-except block
6. Print the Severity results
7. Call main() function

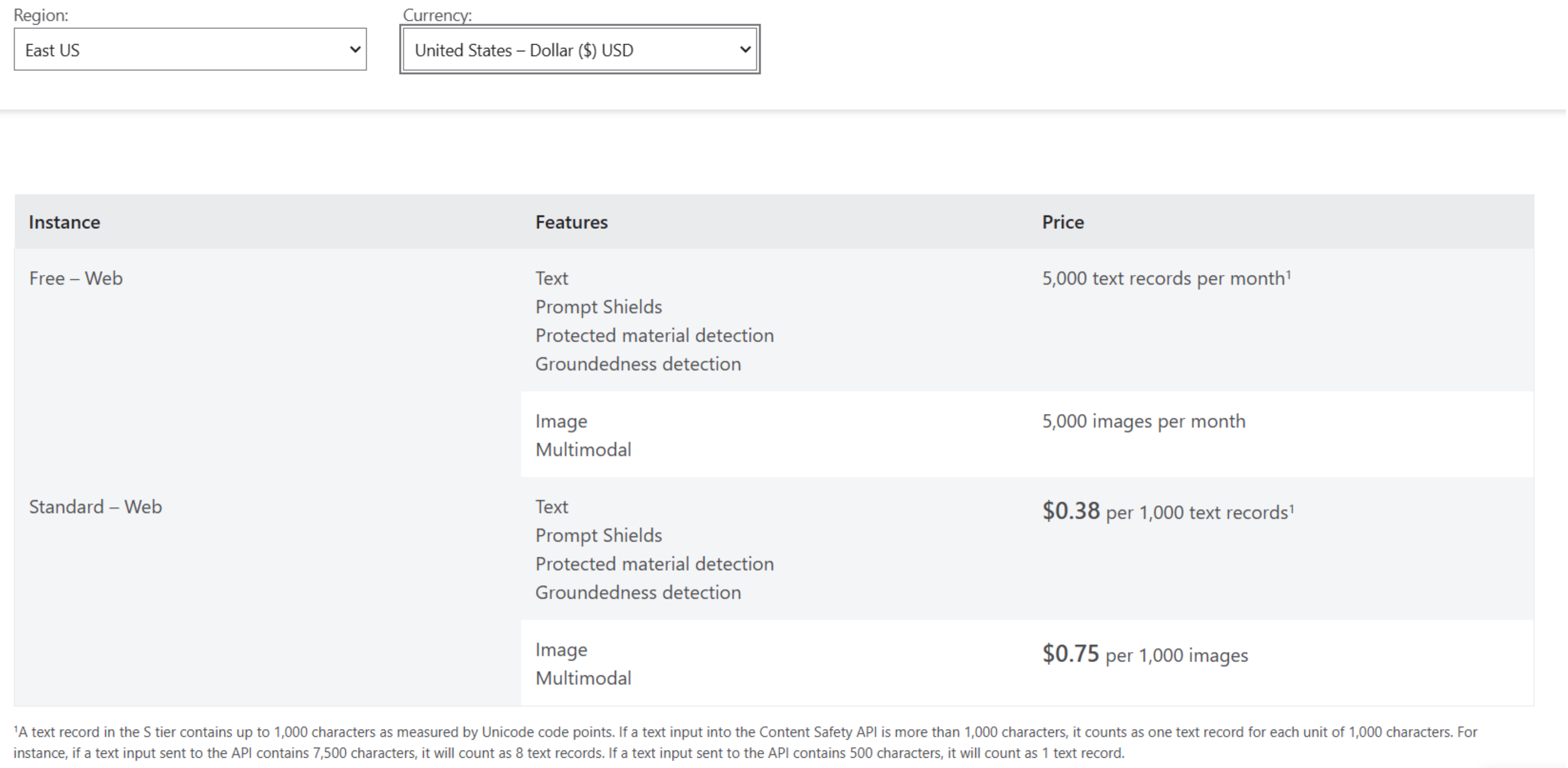
CONTENT SAFETY USE CASES

- 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



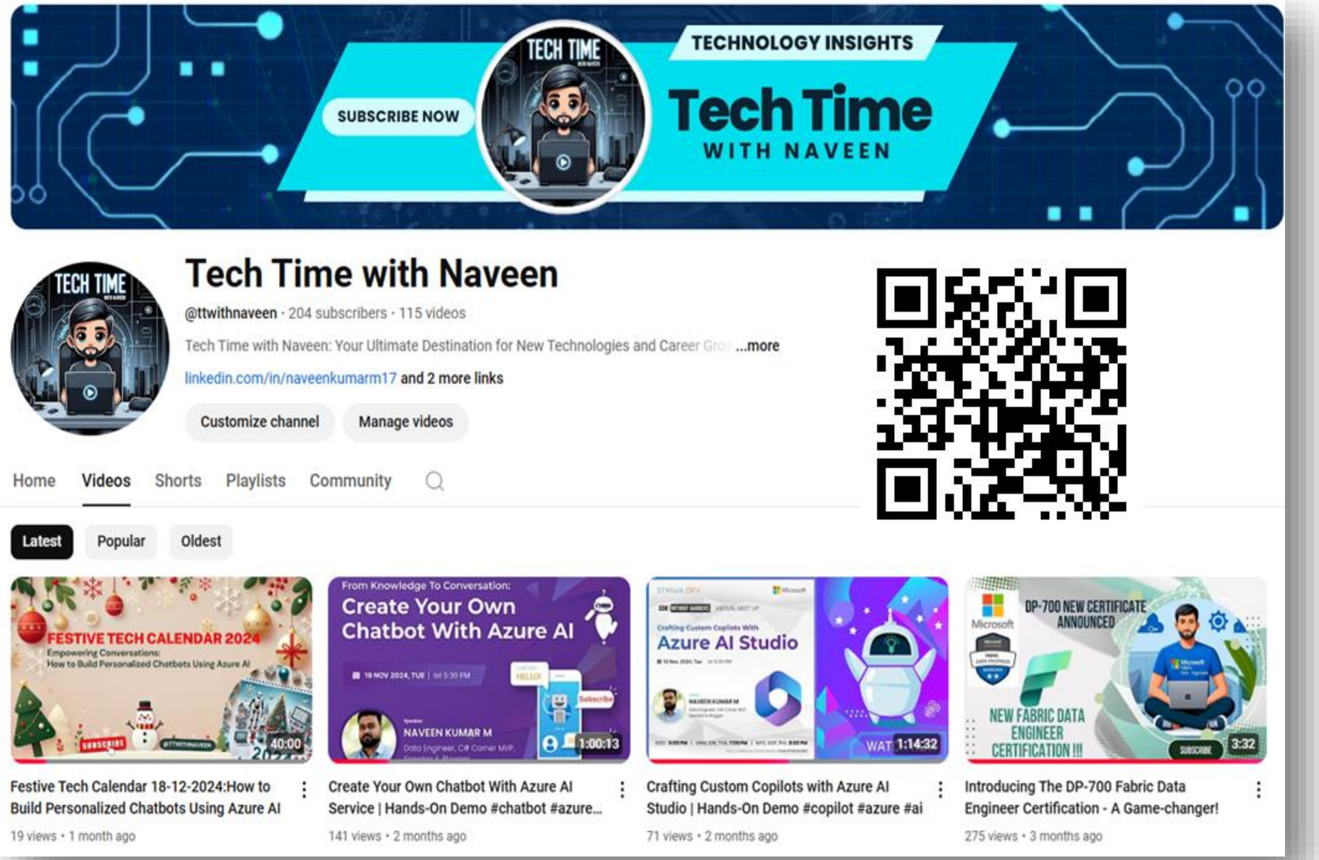
Region: East US

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

[1] A 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.

YouTube



TECHNOLOGY INSIGHTS
SUBSCRIBE NOW
TECH TIME
Tech Time
WITH NAVEEN

Tech Time with Naveen

@ttwithnaveen · 204 subscribers · 115 videos

Tech Time with Naveen: Your Ultimate Destination for New Technologies and Career Gro ...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
Empowering Conversations:
How to Build Personalized Chatbots Using Azure AI
40:00
From Knowledge To Conversation:
Create Your Own Chatbot With Azure AI
NAVEEN KUMAR M
1:00:13
Azure AI Studio
1:14:32
DP-700 NEW CERTIFICATE ANNOUNCED
Microsoft
NEW FABRIC DATA ENGINEER CERTIFICATION !!!
3:32

Festive Tech Calendar 18-12-2024: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

GitHub

Azure AI

Thank you!

For exploring Azure AI Content Safety