**Implement a text moderation solution with Azure AI Content Safety**

The moderate text content tool considers several factors such as the type of content, the platform's policies, and the potential effect on users. You can run moderation tests on sample content. Configure filters to rerun and further refine the test results. And add specific terms to the blocklist that you want to detect and act on.

Whether you want to test a single sentence or an entire dataset, you can experiment with different sensitive levels.

1. Violence
2. Self-harm
3. Sexual
4. Hate

MS Learning : <https://learn.microsoft.com/en-us/shows/azure-ai-content-safety/>

Maximum length : 10000 characters

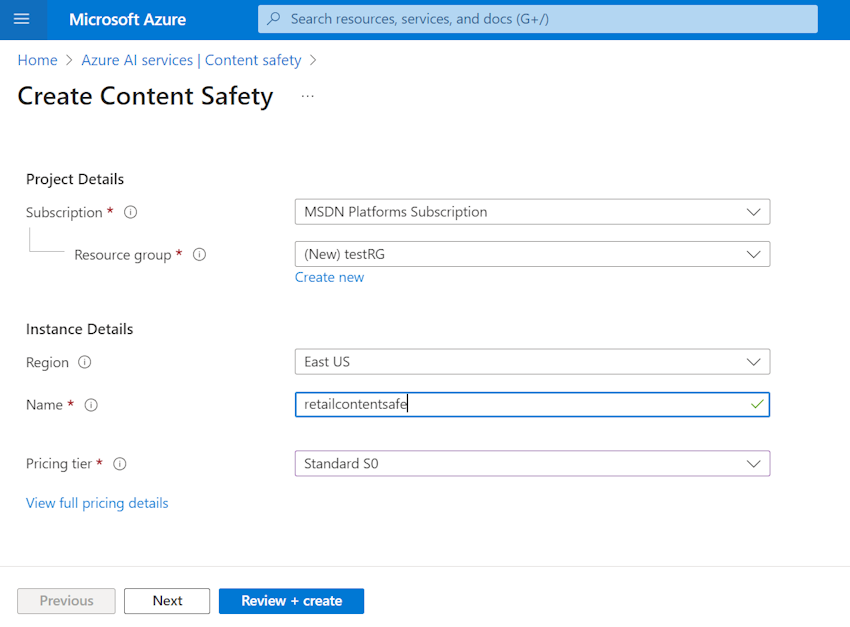
**Create Content Safety**

Step 1. Go to the [Azure](http://portal.azure.com/) portal and sign in with your Azure account.

Step 2. Search "Content Safety" in the search bar and select "Content Safety" from the search results.

Step 3. Click on the Create button for Content Safety.

Step 4. First, choose a subscription and then create a Resource Group named testRG, and type the resource name as retailcontentsafe.

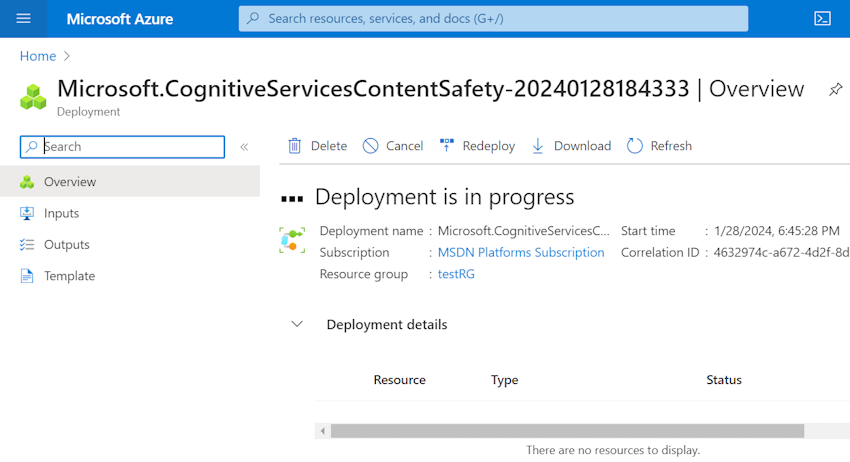
****

Step 5. Click the Review+Create button.

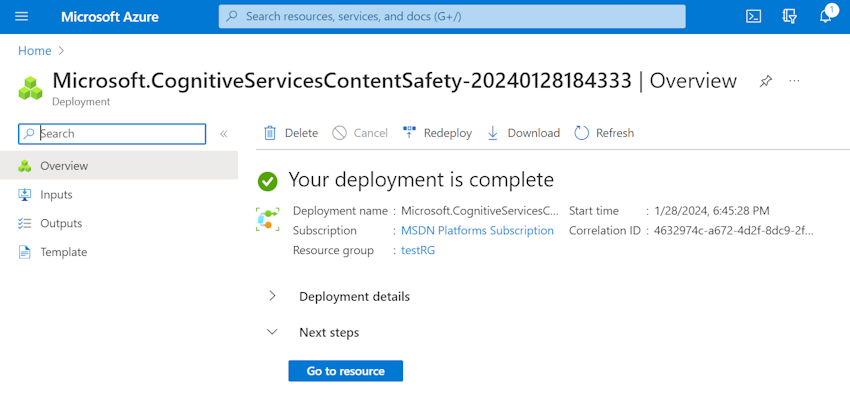
Step 6. Finally, Click the Create button.

****

Step 7. Deployment started in progress in a minute or two deployment will become successful.

****

Step 8. You can see deployment is completed successfully.

****

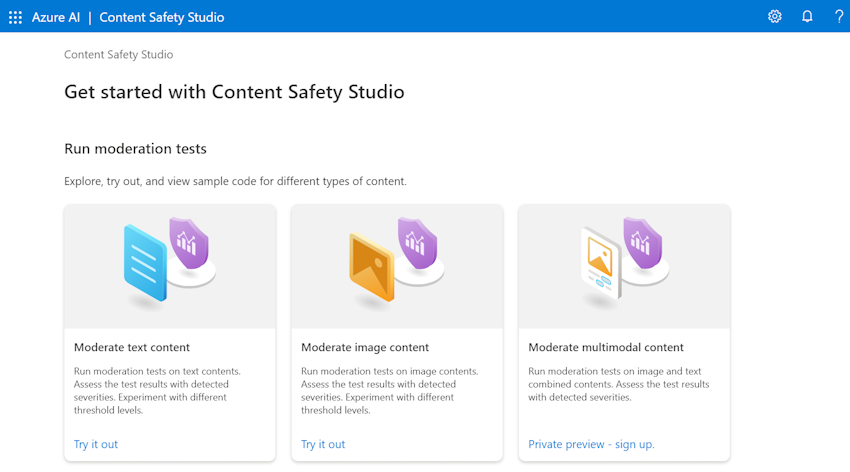
**Getting Started with Content Safety Studio**

**Safe Content Scenario**

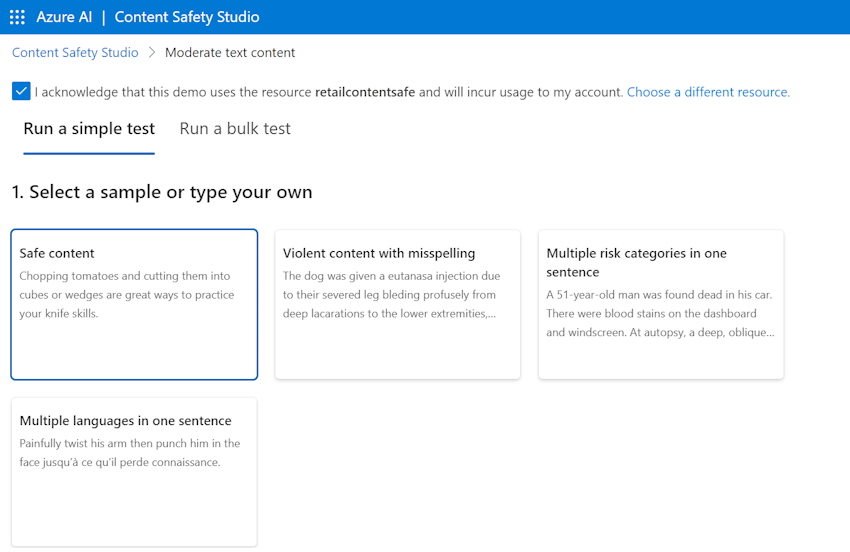
Step 1. Click Content Safety Studio from the Azure Portal.

Step 2. Now you are going to run a moderation test related to the Text.

Step 3. Click Try it out for Moderate Text Content.

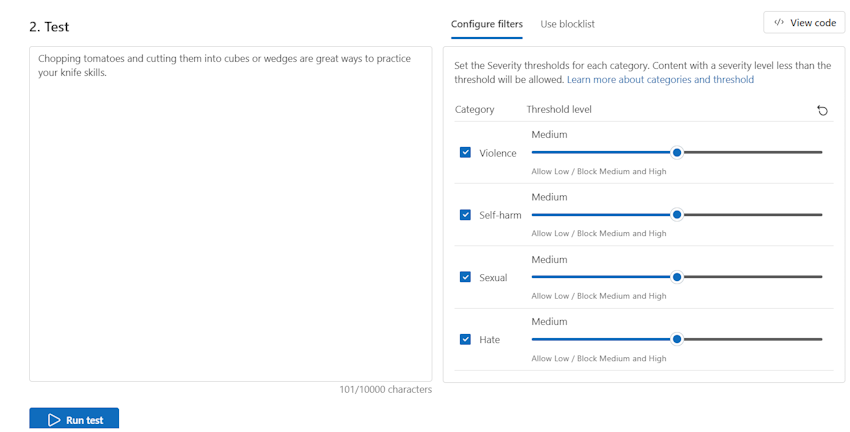
****

Step 4. Click Run a Simple Test and choose the Safe Content option from the moderate text content.

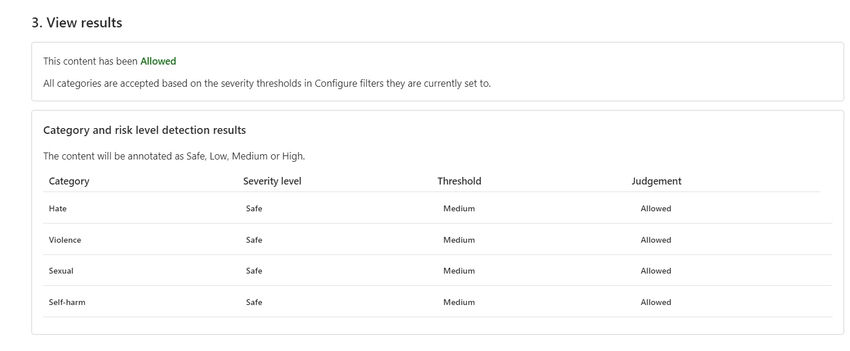
****

Step 5. Prompt - "Chopping tomatoes and cutting them into cubes or wedges are great ways to practice your knife skills."

Step 6. Click the run test button and it's going to be determined whether the status of the prompt is allowed or blocked in the Judgement option.

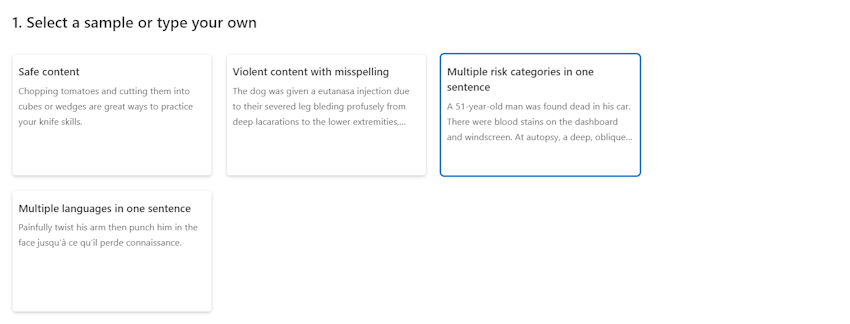
****

Step 7. In the View results section, Prompt content has been allowed and also you can see the category and risk level detection results below.

****

**Multiple risk categories in one sentence scenario**

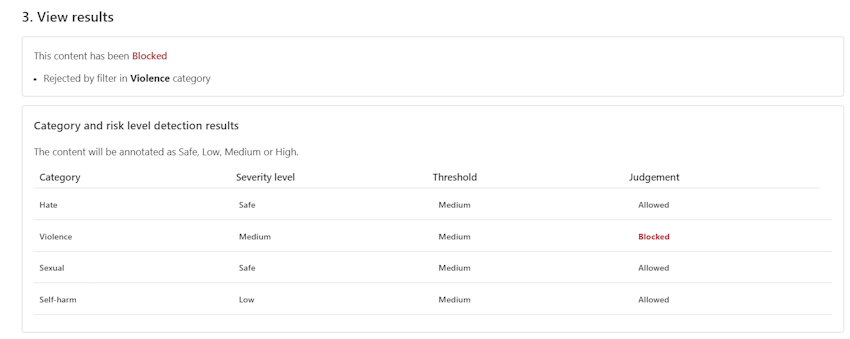
Step 1. Click Run a Simple Test and choose the multiple risk categories in one sentence option from the moderate text content.

****

Step 2. Prompt: "A 51-year-old man was found dead in his car. There were blood stains on the dashboard and windscreen. At autopsy, a deep, oblique, long incised injury was found on the front of the neck. It turns out that he died by suicide."

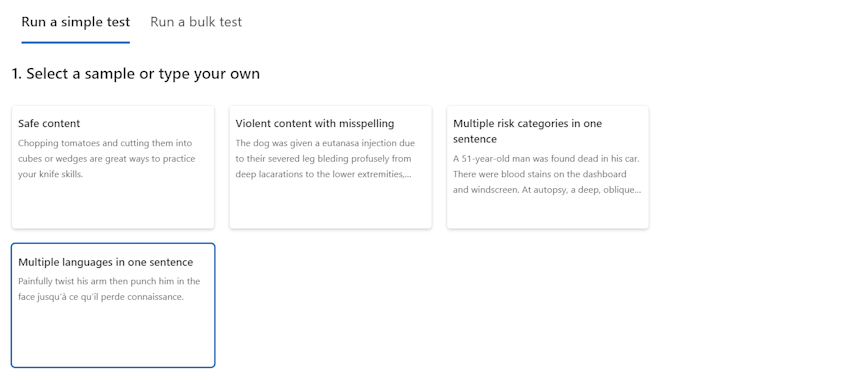
****

Step 3. Click the run test button and it's going to be determined whether the status of the prompt is allowed or blocked in the Judgement option.

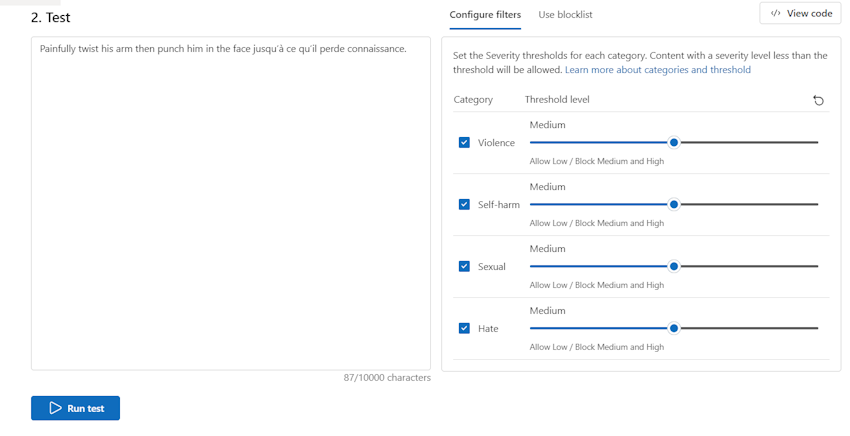
****

**Multiple languages in one sentence Scenario**

Step 1. Click Run a Simple Test and choose the Multiple languages in one sentence option from the moderate text content.

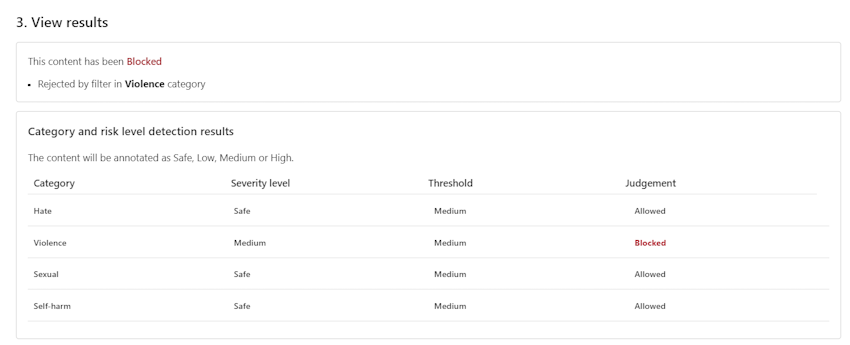
****

Step 2. Prompt: "Painfully twist his arm then punch him in the face jusqu' a` ce qui'l perde connaissance."

****

Step 3. Click the run test button and it going to be determined whether the status of the prompt is allowed or blocked in the Judgement option.

Step 4.  In the View results section, Prompt content has been Blocked and also you can see the category and risk level detection results below.

****

**Other Test Scenarios**

1. Test with medium level with normal text.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Test by changing **Hate** threshold level to low.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Test by removing **Hate** threshold.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Test with violence text.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Test with incorrect spelling words.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Test with correcting **blood** word.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Test by removing blood word sentence.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated