

1.6 - Use AI responsibly with Azure AI Content Safety

- [Overview](#)
 - [Introduction](#)
 - [What is Content Safety](#)
 - [Trusting user-generated content](#)
 - [Content Safety in Azure AI Studio](#)
 - [How does Azure AI Content Safety work?](#)
 - [\(i\) Safeguarding text content](#)
 - [\(ii\) Safeguarding image content](#)
 - [\(iii\) Custom safety solutions](#)
 - [Limitations](#)
 - [Evaluating accuracy](#)
 - [When to use Azure AI Content Safety](#)
 - [Education](#)
 - [Social](#)
 - [Brands](#)
 - [E-Commerce](#)
 - [Gaming](#)
 - [Generative AI services](#)
 - [News](#)
 - [Other situations](#)
 - [Exercise - Implementing Azure AI Content Safety](#)
 - [Provision a _Content Safety_ resource](#)
 - [Use Azure AI Content Safety Prompt Shields](#)
 - [Configure your application](#)
 - [C#](#)
 - [Prerequisites](#)
 - [Setting up](#)
 - [Add code](#)
 - [Python](#)
 - [Prerequisites](#)
 - [Knowledge Check](#)
 - [Summary](#)
-

Overview

Azure AI Content Safety is a comprehensive tool designed to detect and manage harmful content in both user-generated and AI-generated materials. Learn how Azure AI Content Safety uses text and image APIs to help identify and filter out content related to violence, hate, sexual content, and self-harm.

Introduction

The amount of user-generated content being posted online is growing rapidly. We are also increasingly aware of the need to protect everyone from inappropriate or harmful content.

Azure AI Content Safety is an AI service designed to help developers include advanced content safety into their applications and services.

The challenges in maintaining safe and respectful online spaces are growing for developer teams responsible for hosting online discussions. Azure AI Content Safety identifies potentially unsafe content and helps organizations to comply with regulations and meet their own quality standards.

The need for improving online content safety has four main drivers:

- **Increase in harmful content:** There's been a huge growth in user-generated online content, including harmful and inappropriate content.
- **Regulatory pressures:** Government pressure to regulate online content.
- **Transparency:** Users need transparency in content moderation standards and enforcement.
- **Complex content:** Advances in technology are making it easier for users to post multimodal content and videos.

Note: Azure AI Content Safety replaces Azure Content Moderator, which was deprecated in February 2024 and will be retired by February 2027.

What is Content Safety

Azure AI Content Safety is a set of advanced content moderating features that can be incorporated into your applications and services. Azure AI Content Safety is available as a resource in the Azure portal.

Online content safeguarding is needed in a growing number of situations. Not only are we concerned with moderating content generated by people, but must also guard against the malicious use of AI.

Trusting user-generated content

Social interaction is increasingly a part of many digital spaces. Genuine user-generated content is seen as independent and trustworthy, and used alongside advertising and marketing. Different industries are encouraging their customers to connect with each other and their brand.

Harmful content has many negative effects. It damages trusted brands, discourages users from participating in online forums, and can have a devastating impact on individuals.

Azure AI Content Safety is designed to be used in applications and services to protect against harmful user-generated and AI-generated content.

Content Safety in Azure AI Studio

Azure AI Content Safety Studio is available as part of [Azure AI Studio](#), a unified platform that enables you to explore many different Azure AI services, including Content Safety.

← Content Safety

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

Filter text content

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 Preview

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

[Try it out](#)

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](#)

Filter image content

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 c... Preview

Run moderation tests on image and text combined contents. Assess the test results with detected severities.

[Try it out](#)

From the [Azure AI Studio](#) home page, scroll down to find Content Safety and then select **View all Content Safety capabilities**.

Azure AI Content Safety Studio enables you to explore and test Content Safety features for yourself. Select the feature you want to try, and then select *Try it out*. You can then use the user interface to test samples or your own material. Select *View code* to generate sample code in C#, Java, or Python. You can then copy and paste the sample code and amend the variables to use your own data.

Note: You can access Azure AI Content Studio either through [Azure AI Studio](#) or through [Content Safety Studio](#). For guided practice using Azure AI Content Studio, see [Moderate content and detect harm with Azure AI Content Safety Studio](#).

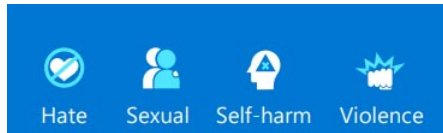
How does Azure AI Content Safety work?

Azure AI Content Safety works with text and images, and AI-generated content.

Content Safety vision capabilities are powered by **Microsoft's Florence foundation model**, which has been **trained with billions of text-image pairs**.

Text analysis uses natural language processing techniques, giving a better understanding of nuance and context. Azure AI Content Safety is multilingual and can detect harmful content in both short form and long form. It's currently available in English, German, Spanish, French, Portuguese, Italian, and Chinese.

Azure AI Content Safety classifies content into four categories:



A severity level for each category is used to determine whether content should be blocked, sent to a moderator, or auto approved.

Azure AI Content Safety features include:

(i) Safeguarding text content

- **Moderate text** scans text across four categories: **violence, hate speech, sexual content, and self-harm**. A severity level from 0 to 6 is returned for each category. This level helps to prioritize what needs immediate attention by people, and how urgently. You can also create a blocklist to scan for terms specific to your situation.
- **Prompt shields** is a unified API to identify and block jailbreak attacks from inputs to LLMs. It includes both user input and documents. These attacks are prompts to LLMs that attempt to bypass the model's in-built safety features. User prompts are tested to ensure the input to the LLM is safe. **Documents are tested to ensure they don't contain unsafe instructions embedded within the text.**
- **Protected material detection** checks AI-generated text for protected text such as recipes, copyrighted song lyrics, or other original material.
- **Groundedness detection** protects against inaccurate responses in AI-generated text by LLMs. Public LLMs use data available at the time they were trained. However, data can be introduced after the original training of the model or be built on private data. **A grounded response is one where the model's output is based on the source information. An ungrounded response is one where the model's output varies from the source information.** Groundedness detection includes a *reasoning* option in the API response. This adds a *reasoning* field that explains any ungroundedness detection. However, reasoning increases processing time and costs.

(ii) Safeguarding image content

- **Moderate images** scans for inappropriate content across four categories: violence, self-harm, sexual, and hate. A severity level is returned: safe, low, or high. You then set a threshold level of low, medium, or high. The combination of the severity and threshold level determines whether the image is allowed or blocked for each category.
- **Moderate multimodal content** scans both images and text, including text extracted from an image using optical character recognition (OCR). Content is analyzed across four categories: violence, hate speech, sexual content, and self-harm.

(iii) Custom safety solutions

- **Custom categories** enables you to create your own categories by providing positive and negative examples, and training the model. Content can then be scanned according to your own category definitions.
- **Safety system message** helps you to write effective prompts to guide an AI system's behavior.

Limitations

Azure AI Content Safety uses AI algorithms, and so may not always detect inappropriate language. And on occasions it might block acceptable language because it relies on algorithms and machine learning to detect problematic language.

Azure AI Content Safety should be tested and evaluated on real data before being deployed. And once deployed, you should continue to monitor the system to see how accurately it's performing.

Evaluating accuracy

When evaluating how accurately Azure AI Content Safety is for your situation, compare its performance against four criteria:

- **True positive** - correct identification of harmful content.
- **False positive** - incorrect identification of harmful content.
- **True negative** - correct identification of harmless content.
- **False negative** - harmful content isn't identified.

Azure AI Content Safety works best to support human moderators who can resolve cases of incorrect identification. When people add content to a site, they don't expect posts to be removed without reason. Communicating with users about why content is removed or flagged as inappropriate helps everyone to understand what is permissible and what isn't.

When to use Azure AI Content Safety

Many online sites encourage users to share their views. People trust other people's feedback about products, services, brands, and more. These comments are often frank, insightful, and seen to be free of marketing bias. But not all content is well intended.

Azure AI Content Safety is an AI service designed to provide a more comprehensive approach to content moderation. Azure AI Content Safety helps organizations to prioritize work for human moderators in a growing number of situations:

Education

The number of learning platforms and online educational sites is growing rapidly, with more and more information being added all the time. Educators need to be sure that students aren't being exposed to inappropriate content, or inputting harmful requests to LLMs. In addition, both educators and students want to know that the content they're consuming is correct and close to the source material.

Social

Social media platforms are dynamic and fast moving, requiring real-time moderation. Moderation of user-generated content includes posts, comments, and images. Azure AI Content Safety helps moderate content that is nuanced and multi-lingual to identify harmful material.

Brands

Brands are making more use of **chat rooms and message forums** to encourage loyal customers to share their views. However offensive material can damage a brand, and discourage customers from contributing. They want to be assured that inappropriate material can be quickly identified and removed. Brands are

also adding generative AI services to help people to communicate with them, and therefore need to guard against bad actors attempting to exploit large language models (LLMs).

E-Commerce

User content is generated by reviewing products and discussing products with other people. This material is powerful marketing, but when inappropriate content is posted it damages consumer confidence. In addition, regulatory and compliance issues are increasingly important. Azure AI Content Safety helps screen product listings for fake reviews and other unwanted content.

Gaming

Gaming is a challenging area to moderate due to its highly visual and often violent graphics. Gaming has strong communities where people are enthusiastic about sharing progress and their experiences. Supporting human moderators to keep gaming safe includes monitoring avatars, usernames, images, and text-based materials. Azure AI Content Safety has advanced AI vision tools to help moderate gaming platforms to detect misconduct.

Generative AI services

Organizations are increasingly using generative AI services to enable internal data to be accessed more easily. To maintain the integrity and safety of internal data, both user prompts and AI-generated outputs need to be checked to prevent malicious use of these systems.

News

News websites need to moderate user comments to prevent the spread of misinformation. Azure AI Content Safety can identify language that includes hate speech and other harmful content.

Other situations

There are many other situations where content needs to be moderated. Azure AI Content Safety can be customized to identify problematic language for specific cases.

Exercise - Implementing Azure AI Content Safety

Provision a *Content Safety* resource

If you don't already have one, you'll need to provision a **Content Safety** resource in your Azure subscription.


1. Open the Azure portal at `https://portal.azure.com`, and sign in using the Microsoft account associated with your Azure subscription.
2. Select **Create a resource**.
3. In the search field, search for **Content Safety**. Then, in the results, select **Create** under **Azure AI Content Safety**.
4. Provision the resource using the following settings:
 - **Subscription:** *Your Azure subscription.*
 - **Resource group:** *Choose or create a resource group.*

- **Region:** Select **East US**
 - **Name:** *Enter a unique name.*
 - **Pricing tier:** Select **F0** (*free*), or **S** (*standard*) if F0 is not available.
5. Select **Review + create**, then select **Create** to provision the resource.
 6. Wait for deployment to complete, and then go to the resource.
 7. Select **Access Control** in the left navigation bar, then select **+ Add** and **Add role assignment**.
 8. Scroll down to choose the **Cognitive Services User** role and select **Next**.
 9. Add your account to this role, and then select **Review + assign**.
 10. Select **Resource Management** in the left hand navigation bar and select **Keys and Endpoint**. Leave this page open so you can copy the keys later.

Use Azure AI Content Safety Prompt Shields

In this exercise you will use Azure AI Studio to test Content Safety Prompt **Shields** with two sample inputs. One simulates a **user prompt**, and the other simulates a **document with potentially unsafe text embedded into it**.

1. In another browser tab, open the Content Safety page of [Azure AI Studio](#) and sign in.
2. Under **Moderate text content** select **Try it out**.
3. On the **Moderate text content** page, under **Azure AI Services** select the Content Safety resource you created earlier.
4. Select **Multiple risk categories in one sentence**. Review the document text for potential issues.
5. Select **Run test** and review the results.
6. Optionally, alter the threshold levels and select **Run test** again.
7. On the left navigation bar, select **Protected material detection for text**.
8. Select **Protected lyrics** and note that these are the lyrics of a published song.
9. Select **Run test** and review the results.
10. On the left navigation bar, select **Moderate image content**.
11. Select **Self-harm content**.
12. Notice that all images are blurred by default in AI Studio. You should also be aware that the sexual content in the samples is very mild.
13. Select **Run test** and review the results.
14. On the left navigation bar, select **Prompt shields**.
15. On the **Prompt shields page**, under **Azure AI Services** select the Content Safety resource you created earlier.
16. Select **Prompt & document attack content**. Review the user prompt and document text for potential issues.
17. Select **Run test**.
18. In **View results**, verify that Jailbreak attacks were detected in both the user prompt and the document.

 **Code is available for all of the samples in AI Studio.**

19. Under **Next steps**, under **View the code** select **View code**. The **Sample code** window is displayed.
20. Use the down arrow to select either Python or C# and then select **Copy** to copy the sample code to the clipboard.
21. Close the **Sample code** screen.

Configure your application

You will now create an application in either C# or Python.

C#

Prerequisites

- [Visual Studio Code](#) on one of the [supported platforms](#).
- [.NET 8](#) is the target framework for this exercise.
- The [C# extension](#) for Visual Studio Code.

Setting up

Perform the following steps to prepare Visual Studio Code for the exercise.

1. Start Visual Studio Code and in the Explorer view, click **Create .NET Project** selecting **Console App**.
2. Select a folder on your computer, and give the project a name. Select **Create project** and acknowledge the warning message.
3. In the Explorer pane, expand Solution Explorer and select **Program.cs**.
4. Build and run the project by selecting **Run -> Run without Debugging**.
5. Under Solution Explorer, right-click the C# project and select **Add NuGet Package**.
6. Search for **Azure.AI.TextAnalytics** and select the latest version.
7. Search for a second NuGet Package: **Microsoft.Extensions.Configuration.Json 8.0.0**. The project file should now list two NuGet packages.

Add code

1. Paste the sample code you copied earlier under the **ItemGroup** section.
2. Scroll down to find *Replace with your own subscription_key and endpoint*.
3. In the Azure portal, on the Keys and Endpoint page, copy one of the Keys (1 or 2). Replace **** with this value.
4. In the Azure portal, on the Keys and Endpoint page, copy the Endpoint. Paste this value into your code to replace ****.
5. In **Azure AI Studio**, copy the **User prompt** value. Paste this into your code to replace ****.
6. Scroll down to **** and delete this line of code.
7. In **Azure AI Studio**, copy the **Document** value.
8. Scroll down to **** and paste your document value.
9. Select **Run -> Run without Debugging** and verify that an attack was detected.

Python

Prerequisites

- [Visual Studio Code](#) on one of the [supported platforms](#).
- The [Python extension](#) is installed for Visual Studio Code.
- The [requests module](#) is installed.

1. Create a new Python file with a **.py** extension and give it a suitable name.
2. Paste the sample code you copied earlier.

3. Scroll down to find the section titled *Replace with your own subscription_key and endpoint*.
 4. In the Azure portal, on the Keys and Endpoint page, copy one of the Keys (1 or 2). Replace **** with this value.
 5. In the Azure portal, on the Keys and Endpoint page, copy the Endpoint. Paste this value into your code to replace ****.
 6. In **Azure AI Studio**, copy the **User prompt** value. Paste this into your code to replace ****.
 7. Scroll down to **** and delete this line of code.
 8. In **Azure AI Studio**, copy the **Document** value.
 9. Scroll down to **** and paste your document value.
 10. From the integrated terminal for your file, run the program, eg:
 - `.\prompt-shield.py`
 11. Validate that an attack is detected.
 12. Optionally, you can experiment with different test content and document values.
-

Knowledge Check

1. Which feature of Azure AI Content Safety helps protect large language models from document injection attacks? *

☒ Prompt Shields

✓ Correct. Prompt shields block jailbreak attacks from inputs to LLMs including user input prompts and document injection attacks.

☐ Groundedness detection

☐ Custom categories

2. What is the purpose of the Groundedness detection feature in Azure AI Content Safety? *

☒ To verify AI-generated text is based on provided source materials.

✓ Correct. Groundedness detection protects against inaccurate responses in AI-generated text by LLMs.

☐ To detect harmful images.

☐ To moderate multimodal content

3. Which social media issues does Azure AI Content Safety address? *

☒ The growth of inappropriate online content including bullying and hate speech.

✓ Correct. Azure AI Content Safety helps moderate inappropriate content including bullying and hate speech.

☐ The popularity of online gaming.

☐ The need to train AI models with more accurate content.

4. How does Azure AI Content Safety help businesses to protect their brand image? *

☒ By moderating comments and messages from customers.

✓ Correct. Brands often encourage customers to leave comments and reviews, which need to be moderated to prevent inappropriate content.

☐ By detecting a brand's violent graphics.

☐ To ensure students are consuming accurate content.

5. What is a benefit of Azure AI Content Safety? *

☒ Reducing the amount of psychologically damaging material that human moderators are exposed to.

✓ Correct. Human moderators can reduce the amount of harmful material they have to read by automatically assigning a severity score to content.

☐ Providing an easy-to-implement service for users to leave comments on e-commerce sites.

☐ Allowing e-commerce companies to create custom categories for their products and services.

Summary

The proliferation of user-generated content makes it near-impossible for human moderators to effectively manage online platforms. Yet as the amount of user-generated content grows, so does the importance of online safety.

Azure AI Content Safety uses AI models to automatically detect violent, sexual, self-harm, or hateful language in real time. It allocates a severity level, so that human moderators can focus on high-priority cases and be exposed to a smaller amount of disturbing content. Azure AI Content Safety includes features to moderate both people-generated and AI-generated material.

In this module, you've seen how the features of Azure AI Content Safety can help e-commerce brands, gaming companies, and educators to provide safer spaces for users.

👉 Compiled by [Kenneth Leung](#) (2025)