What is Azure AI Content Safety?

Article • 12/05/2023

Azure AI Content Safety 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. We also have an interactive Content Safety Studio that allows you to view, explore and try out sample code for detecting harmful content across different modalities.

Content filtering software can help your app comply with regulations or maintain the intended environment for your users.

This documentation contains the following article types:

- Quickstarts are getting-started instructions to guide you through making requests to the service.
- How-to guides contain instructions for using the service in more specific or customized ways.
- Concepts provide in-depth explanations of the service functionality and features.

Where it's used

The following are a few scenarios in which a software developer or team would require a content moderation service:

- 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.
- K-12 education solution providers filtering out content that is inappropriate for students and educators.

> ⓘ **Important**
>
> You cannot use Azure AI Content Safety to detect illegal child exploitation images.

Product types

There are different types of analysis available from this service. The following table describes the currently available APIs.

Expand table

Type	Functionality
Analyze text API	Scans text for sexual content, violence, hate, and self harm with multi-severity levels.
Analyze image API	Scans images for sexual content, violence, hate, and self harm with multi-severity levels.
Jailbreak risk detection (new)	Scans text for the risk of a jailbreak attack on a Large Language Model. Quickstart
Protected material text detection (new)	Scans AI-generated text for known text content (for example, song lyrics, articles, recipes, selected web content). Quickstart

Content Safety Studio

Azure AI Content Safety Studio is an online tool designed to handle potentially offensive, risky, or undesirable content using cutting-edge content moderation ML models. It provides templates and customized workflows, enabling users to choose and build their own content moderation system. Users can upload their own content or try it out with provided sample content.

Content Safety Studio not only contains out-of-the-box AI models but also includes Microsoft's built-in terms blocklists to flag profanities and stay up to date with new trends. You can also upload your own blocklists to enhance the coverage of harmful content that's specific to your use case.

Studio also lets you set up a moderation workflow, where you can continuously monitor and improve content moderation performance. It can help you meet content requirements from all kinds of industries like gaming, media, education, E-commerce, and more. Businesses can easily connect their services to the Studio and have their content moderated in real-time, whether user-generated or AI-generated.

All of these capabilities are handled by the Studio and its backend; customers don't need to worry about model development. You can onboard your data for quick validation and monitor your KPIs accordingly, like technical metrics (latency, accuracy, recall), or business metrics (block rate, block volume, category proportions, language proportions, and more). With simple operations and configurations, customers can test different solutions quickly and find the best fit, instead of spending time experimenting with custom models or doing moderation manually.

Content Safety Studio

Content Safety Studio features

In Content Safety Studio, the following Azure AI Content Safety service features are available:

- Moderate Text Content : With the text moderation tool, you can easily run tests on text content. Whether you want to test a single sentence or an entire dataset, our tool offers a user-friendly interface that lets you assess the test results directly in the portal. You can experiment with different sensitivity levels to configure your content filters and blocklist management, ensuring that your content is always moderated to your exact specifications. Plus, with the ability to export the code, you can implement the tool directly in your application, streamlining your workflow and saving time.

- Moderate Image Content : With the image moderation tool, you can easily run tests on images to ensure that they meet your content standards. Our user-friendly interface allows you to evaluate the test results directly in the portal, and you can experiment with different sensitivity levels to configure your content filters. Once you've customized your settings, you can easily export the code to implement the tool in your application.

- Monitor Online Activity : The powerful monitoring page allows you to easily track your moderation API usage and trends across different modalities. With this feature, you can access detailed response information, including category and severity distribution, latency, error, and blocklist detection. This information provides you with a complete overview of your content moderation performance, enabling you to optimize your workflow and ensure that your content is always moderated to your exact specifications. With our user-friendly interface, you can quickly and easily navigate the monitoring page to access the information you need to make informed decisions about your content moderation strategy. You have the tools you need to stay on top of your content moderation performance and achieve your content goals.

Input requirements

The default maximum length for text submissions is 10K characters. If you need to analyze longer blocks of text, you can split the input text (for example, by punctuation or spacing) across multiple related submissions.

The maximum size for image submissions is 4 MB, and image dimensions must be between 50 x 50 pixels and 2,048 x 2,048 pixels. Images can be in JPEG, PNG, GIF, BMP, TIFF, or WEBP formats.

Security

Use Microsoft Entra ID or Managed Identity to manage access

For enhanced security, you can use Microsoft Entra ID or Managed Identity (MI) to manage access to your resources.

- Managed Identity is automatically enabled when you create a Content Safety resource.
- Microsoft Entra ID is supported in both API and SDK scenarios. Refer to the general AI services guideline of Authenticating with Microsoft Entra ID. You can also grant access to other users within your organization by assigning them the roles of **Cognitive Services Users** and **Reader**. To learn more about granting user access to Azure resources using the Azure portal, refer to the Role-based access control guide.

Encryption of data at rest

Learn how Azure AI Content Safety handles the encryption and decryption of your data. Customer-managed keys (CMK), also known as Bring Your Own Key (BYOK), offer greater flexibility to create, rotate, disable, and revoke access controls. You can also audit the encryption keys used to protect your data.

Pricing

Currently, Azure AI Content Safety has an **F0 and S0** pricing tier.

Service limits

Language support

Content Safety models have been specifically trained and tested in the following languages: English, German, Japanese, Spanish, French, Italian, Portuguese, and Chinese. However, the service can work in many other languages, but the quality might vary. In all cases, you should do your own testing to ensure that it works for your application.

For more information, see Language support.

Region/location

To use the Content Safety APIs, you must create your Azure AI Content Safety resource in the supported regions. Currently, it is available in the following Azure regions:

- Australia East
- Canada East
- Central US
- East US
- East US 2
- France Central
- Japan East
- North Central US
- South Central US
- Switzerland North
- UK South
- West Europe
- West US 2
- Sweden Central

Private preview features, such as jailbreak risk detection and protected material detection, are available in the following Azure regions:

- East US
- West Europe

Feel free to contact us if you need other regions for your business.

Query rates

⛶ Expand table

Pricing Tier	Requests per 10 seconds (RPS)
F0	1000
S0	1000

If you need a faster rate, please contact us to request.

Contact us

If you get stuck, email us or use the feedback widget on the upper right of any docs page.

Next steps

Follow a quickstart to get started using Azure AI Content Safety in your application.

Content Safety quickstart