**Azure AI Content Safety**

Azure AI Content Safety is an AI service that **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. The interactive Content Safety Studio allows you to view, explore, and try out sample code for detecting harmful content across different modalities.

**Where it's used**

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

* User prompts submitted to a generative AI service.
* Content produced by generative AI models.
* 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.

**Product features**

This service makes several different types of analysis available. The following table describes the currently available APIs.

A screenshot of a computer

Description automatically generated

**Content Safety Studio**

[Azure AI Content Safety Studio](https://contentsafety.cognitive.azure.com/) 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 content 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.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Content Safety Studio features**

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

* [**Moderate Text Content**](https://contentsafety.cognitive.azure.com/text): 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**](https://contentsafety.cognitive.azure.com/image): 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**](https://contentsafety.cognitive.azure.com/monitor): 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.

**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](https://learn.microsoft.com/en-us/azure/ai-services/authentication?tabs=powershell#authenticate-with-azure-active-directory). 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](https://learn.microsoft.com/en-us/azure/role-based-access-control/quickstart-assign-role-user-portal).

**Pricing**

* Currently, Azure AI Content Safety has an **F0** and **S0** pricing tier.
* <https://azure.microsoft.com/en-us/pricing/details/cognitive-services/content-safety/>

**Input requirements**

See the following list for the input requirements for each feature.

* **Analyze text API**:
  + Default maximum length: 10K characters (split longer texts as needed).
* **Analyze image API**:
  + Maximum image file size: 4 MB
  + Dimensions between 50 x 50 and 7200 x 7200 pixels.
  + Images can be in JPEG, PNG, GIF, BMP, TIFF, or WEBP formats.
* **Analyze multimodal API (preview)**:
  + Default maximum text length: 1K characters.
  + Maximum image file size: 4 MB
  + Dimensions between 50 x 50 and 7200 x 7200 pixels.
  + Images can be in JPEG, PNG, GIF, BMP, TIFF, or WEBP formats.
* **Prompt Shields API**:
  + Maximum prompt length: 10K characters.
  + Up to five documents with a total of 10K characters.
* **Groundedness detection API (preview)**:
  + Maximum length for grounding sources: 55,000 characters (per API call).
  + Maximum text and query length: 7,500 characters.
* **Protected material detection APIs**:
  + Default maximum length: 1K characters.
  + Default minimum length: 110 characters (for scanning LLM completions, not user prompts).
* **Custom categories (standard) API (preview)**:
  + Maximum inference input length: 1K characters.

**Language support**

* The Azure AI Content Safety models for protected material, groundedness detection, and custom categories (standard) work with English only.
* Other Azure AI Content Safety models have been specifically trained and tested on the following languages: Chinese, English, French, German, Spanish, Italian, Japanese, Portuguese. However, these features 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.
* <https://learn.microsoft.com/en-us/azure/ai-services/content-safety/language-support>

**Region availability**

* To use the Content Safety APIs, you must create your Azure AI Content Safety resource in a supported region. Currently, the Content Safety features are available in the following Azure regions with different API versions:

A screenshot of a survey

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Query rates**

Content Safety features have query rate limits in requests-per-second (RPS) or requests-per-10-seconds (RP10S) . See the following table for the rate limits for each feature.

