Moderation

Copy page

Identify potentially harmful content in text and images.

Use the moderations endpoint to check whether text or images are potentially harmful. If harmful content is identified, you can take corrective action, like filtering content or intervening with user accounts creating offending content. The moderation endpoint is free to use.

You can use two models for this endpoint:

- `omni-moderation-latest` : This model and all snapshots support more categorization options and multi-modal inputs.
- `text-moderation-latest` **(Legacy)**: Older model that supports only text inputs and fewer input categorizations. The newer omni-moderation models will be the best choice for new applications.

Quickstart

Use the tabs below to see how you can moderate text inputs or image inputs, using our official SDKs and the omni-moderation-latest model:

Moderate text inputs | Moderate images and text

Get classification information for a text input — python

```python
from openai import OpenAI
client = OpenAI()

response = client.moderations.create(
    model="omni-moderation-latest",
    input="...text to classify goes here...",
)

print(response)
```

Here's a full example output, where the input is an image from a single frame of a war movie. The model correctly predicts indicators of violence in the image, with a `violence` category score of greater than 0.8.

```
{
  "id": "modr-970d409ef3bef3b70c73d8232df86e7d",
  "model": "omni-moderation-latest",
```

```
  "results": [
    {
      "flagged": true,
      "categories": {
        "sexual": false,
        "sexual/minors": false,
        "harassment": false,
        "harassment/threatening": false,
        "hate": false,
        "hate/threatening": false,
        "illicit": false,
        "illicit/violent": false,
        "self-harm": false,
        "self-harm/intent": false,
        "self-harm/instructions": false,
        "violence": true,
        "violence/graphic": false
      },
      "category_scores": {
        "sexual": 2.34135824776394e-7,
        "sexual/minors": 1.6346470245419304e-7,
        "harassment": 0.0011643905680426018,
        "harassment/threatening": 0.0022121340080906377,
        "hate": 3.1999824407395835e-7,
        "hate/threatening": 2.4923252458203563e-7,
        "illicit": 0.0005227032493135171,
        "illicit/violent": 3.682979260160596e-7,
        "self-harm": 0.0011175734280627694,
        "self-harm/intent": 0.0006264858507989037,
        "self-harm/instructions": 7.368592981140821e-8,
        "violence": 0.8599265510337075,
        "violence/graphic": 0.37701736389561064
      },
      "category_applied_input_types": {
        "sexual": [
          "image"
        ],
        "sexual/minors": [],
        "harassment": [],
        "harassment/threatening": [],
        "hate": [],
        "hate/threatening": [],
        "illicit": [],
        "illicit/violent": [],
        "self-harm": [
          "image"
        ],
        "self-harm/intent": [
          "image"
        ],
        "self-harm/instructions": [
          "image"
        ],
        "violence": [
          "image"
        ],
```

```
            "violence/graphic": [
              "image"
            ]
          }
        }
      ]
    }
```

The output has several categories in the JSON response, which tell you which (if any) categories of content are present in the inputs, and to what degree the model believes them to be present.

OUTPUT CATEGORY	DESCRIPTION
`flagged`	Set to `true` if the model classifies the content as potentially harmful, `false` otherwise.
`categories`	Contains a dictionary of per-category violation flags. For each category, the value is `true` if the model flags the corresponding category as violated, `false` otherwise.
`category_scores`	Contains a dictionary of per-category scores output by the model, denoting the model's confidence that the input violates the OpenAI's policy for the category. The value is between 0 and 1, where higher values denote higher confidence.
`category_applied_input_types`	This property contains information on which input types were flagged in the response, for each category. For example, if the both the image and text inputs to the model are flagged for "violence/graphic", the `violence/graphic` property will be set to `["image", "text"]`. This is only available on omni models.

ⓘ We plan to continuously upgrade the moderation endpoint's underlying model. Therefore, custom policies that rely on `category_scores` may need recalibration over time.

Content classifications

The table below describes the types of content that can be detected in the moderation API, along with which models and input types are supported for each category.

CATEGORY	DESCRIPTION	MODELS	INPUTS
`harassment`	Content that expresses, incites, or promotes harassing language towards any target.	All	Text only

`harassment/threatening`	Harassment content that also includes violence or serious harm towards any target.	All	Text only
`hate`	Content that expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste. Hateful content aimed at non-protected groups (e.g., chess players) is harassment.	All	Text only
`hate/threatening`	Hateful content that also includes violence or serious harm towards the targeted group based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.	All	Text only
`illicit`	Content that gives advice or instruction on how to commit illicit acts. A phrase like "how to shoplift" would fit this category.	Omni only	Text only
`illicit/violent`	The same types of content flagged by the `illicit` category, but also includes references to violence or procuring a weapon.	Omni only	Text only
`self-harm`	Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.	All	Text and images
`self-harm/intent`	Content where the speaker expresses that they are engaging or intend to engage in acts of self-harm, such as suicide, cutting, and eating disorders.	All	Text and images
`self-harm/instructions`	Content that encourages performing acts of self-harm, such as suicide, cutting, and eating disorders, or that gives instructions or advice on how to commit such acts.	All	Text and images
`sexual`	Content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness).	All	Text and images
`sexual/minors`	Sexual content that includes an individual who is under 18 years old.	All	Text only
`violence`	Content that depicts death, violence, or physical injury.	All	Text and images

`violence/graphic`	Content that depicts death, violence, or physical injury in graphic detail.	All	Text and images