Moderation

Learn how to build moderation into your AI applications.

Overview

The moderations endpoint is a tool you can use to check whether text is potentially harmful. Developers can use it to identify content that might be harmful and take action, for instance by filtering it.

The models classifies the following categories:

CATEGORY	DESCRIPTION
`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.
`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.
`harassment`	Content that expresses, incites, or promotes harassing language towards any target.
`harassment/threatening`	Harassment content that also includes violence or serious harm towards any target.
`self-harm`	Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.
`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.
`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.
`sexual`	Content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness).
`sexual/minors`	Sexual content that includes an individual who is under 18 years old.

CATEGORY	DESCRIPTION
`violence`	Content that depicts death, violence, or physical injury.
`violence/graphic`	Content that depicts death, violence, or physical injury in graphic detail.

The moderation endpoint is free to use for most developers. For higher accuracy, try splitting long pieces of text into smaller chunks each less than 2,000 characters.

> ⓘ We are continuously working to improve the accuracy of our classifier. Our support for non-English languages is currently limited.

Quickstart

To obtain a classification for a piece of text, make a request to the moderation endpoint as demonstrated in the following code snippets:

Example: Getting moderations

curl

```
curl https://api.openai.com/v1/moderations \
  -X POST \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -d '{"input": "Sample text goes here"}'
```

Below is an example output of the endpoint. It returns the following fields:

- `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 raw 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. The scores should not be interpreted as probabilities.

```
{
    "id": "modr-XXXXX",
    "model": "text-moderation-007",
    "results": [
        {
            "flagged": true,
            "categories": {
                "sexual": false,
                "hate": false,
                "harassment": false,
                "self-harm": false,
                "sexual/minors": false,
                "hate/threatening": false,
                "violence/graphic": false,
                "self-harm/intent": false,
                "self-harm/instructions": false,
                "harassment/threatening": true,
                "violence": true
            },
            "category_scores": {
                "sexual": 1.2282071e-6,
                "hate": 0.010696256,
                "harassment": 0.29842457,
                "self-harm": 1.5236925e-8,
                "sexual/minors": 5.7246268e-8,
                "hate/threatening": 0.0060676364,
                "violence/graphic": 4.435014e-6,
                "self-harm/intent": 8.098441e-10,
                "self-harm/instructions": 2.8498655e-11,
                "harassment/threatening": 0.63055265,
                "violence": 0.99011886
            }
        }
    ]
}
```

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