Omni-moderation-2024-09-26 Information for developers

June 2, 2026

0.0.1 Intro

Omni-moderation-2024-09-26 is OpenAI's Moderation API model for identifying potentially harmful content. Its primary purpose is to help developers protect their users and shield their applications from possible harms. The model is available for free, with a rate limit, as part of OpenAI's long standing commitment to making the AI ecosystem safer.

Omni-moderation-2024-09-26 accepts text inputs and image inputs, and returns a top-level binary flagged prediction, per-category binary flags, per-category scores, and, when relevant, the input type(s) that triggered each category.

0.0.2 Model Data and Training

Training data sources Like OpenAI's other models, the Omni-moderation-2024-09-26 model was trained on diverse datasets, including information that is publicly available on the internet, information that we partner with third parties to access, and information that our users or human trainers and researchers provide or generate. To ensure that our moderation system performs well in the context of our production use cases, while maintaining strong data quality and privacy safeguards, Omni-moderation-2024-09-26's training data is curated using an iterative, production-informed data collection and labeling process. At a high level, dataset construction follows a continuous improvement loop that combines de-identified real-world data, active learning, expert annotation, and targeted synthetic data generation. Further details on data generation and annotation can be found in the technical paper and our latest blog post.

0.0.3 Policy and taxonomy definitions

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

harassment/threatening	Harassment content that also includes violence or serious harm towards any target.	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.	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.	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.	Text only
illicit/violent	The same types of content flagged by the illicit category, but also includes references to violence or procuring a weapon.	Text only
self-harm	Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.	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.	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.	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).	Text and images
sexual/minors	Sexual content that includes an individual who is under 18 years old.	Text only
violence	Content that depicts death, violence, or physical injury.	Text and images
violence/graphic	Content that depicts death, violence, or physical injury in graphic detail.	Text and images

Source: https://platform.openai.com/docs/guides/moderation

0.0.4 Evaluations

Performance Evaluations Our model is trained and tested on both production and public data. We report the model performance on a different test dataset containing only samples from public data[1], as well as several publicly available datasets on undesired content detection. These evaluations were run on available datasets at the time of the release, on September 26, 2024, and do not cover more recent datasets or models.

As indicated in the technical paper, we compare the performance of our model with Perspective API as a baseline on our test dataset, TweetEval (Barbieri et al. 2020), Stormfront hate speech dataset (de Gibert et al. 2018), a subset of Reddit comments with noisy labels on erotic content processed according to Barrientos et al. (2020) and a downsampled Jigsaw toxic comments test dataset (Jigsaw 2018).

[1] https://github.com/openai/moderation-api-release

Table 2

Category	AUPRC (text only)
harassment/threatening	0.637
harassment	0.832
sexual/minors	0.872
sexual	0.950
hate/threatening	0.711
hate	0.790
illicit/violent	0.725
illicit	0.747
self-harm/instructions	0.673
self-harm/intent	0.826
self-harm	0.841
violence/graphic-violent	0.822
violence	0.875

Safety Evaluations

Fairness and Bias As detailed in the technical paper, similar to other existing NLP models, our models may also suffer from counterfactual bias towards certain demographic attributes as bias commonly exists in the training data. We mitigated the issue by curating a synthetic dataset with templates that tend to lead to hateful predictions, e.g., "[subject] is selfish/foolish/narrow-minded.". The [subject] could either be filled with real demographic attributes (e.g., Latino) or random object names (e.g., "black blanket"), which forms hateful and safe samples respectively. We observe that the curated dataset not only mitigates bias to some degree, but also helps improve the model performance. For instance, the average AUPRC on hateful content of an earlier version of the model was improved from 0.417 to 0.551 by adding 69k curated synthetic examples. More details can be found in the technical paper.

Our latest update of the moderation model also improved accuracy across forty languages compared to the previous model.

0.1 Implementation guidance

Omni-moderation-2024-09-26 is available via OpenAI's API and is free for use with certain rate limits ensuring a fair and reliable access to the model.

Consult our API documentation for up to date rate limits, and our implementation guide to use the moderation endpoint.