

# Moderation

Kali Raushan IS-21-1

# Модерация

– это инструмент, который вы можете использовать для проверки того, являются ли текст или изображения потенциально опасными. После выявления вредоносного контента разработчики могут предпринять корректирующие действия, такие как фильтрация контента или вмешательство в учетные записи пользователей, создающие оскорбительный контент. Конечная точка модерации доступна для бесплатного использования.



# Для этой конечной точки доступны следующие модели:

01.

omni-moderation-latest: Эта модель и все моментальные снимки поддерживают больше возможностей категоризации и мультимодального ввода.

02.

text-moderation-latest (устаревшая): более старая модель, которая поддерживает только текстовый ввод и меньшее количество категорий ввода. Новые модели omni-moderation станут лучшим выбором для новых приложений.

# Модерация может использоваться для классификации как текста, так и изображений.

```
1  from openai import OpenAI
2  client = OpenAI()
3
4  response = client.moderations.create(
5      model="omni-moderation-latest",
6      input=[
7          {"type": "text", "text": "...text to classify goes here..."},
8          {
9              "type": "image_url",
10             "image_url": {
11                 "url": "https://example.com/image.png",
12                 # can also use base64 encoded image URLs
13                 # "url": "..."
14             }
15         },
16     ],
17 )
18
19 print(response)
```

# ***Типы контента которые можно обнаружить в API модерации***

## ***harassment***

Контент, который выражает, подстрекает или пропагандирует оскорбительные высказывания в отношении любой целевой аудитории.

## ***harassment/ threatening***

Контент оскорбительного характера, который также включает насилие или нанесение серьезного вреда любой цели.

## ***hate***

Контент, который выражает, подстрекает или пропагандирует ненависть по признаку расы, пола, этнической принадлежности, религии, национальности, сексуальной ориентации, инвалидности или касты. Контент, разжигающий ненависть и направленный на незащищенные группы (например, шахматистов), является домогательством.

## ***hate/threatening***

Разжигающий ненависть контент, который также включает насилие или серьезный вред целевой группе по признаку расы, пола, этнической принадлежности, религии, национальности, сексуальной ориентации, инвалидности или касты.

### ***illicit***

Контент, который поощряет планирование или выполнение ненасильственных правонарушений или дает советы или инструкции о том, как совершать незаконные действия. Фраза типа «как воровать в магазине» подойдет к этой категории.

### ***illicit/violent***

Те же типы контента, которые отмечены в *illicit* категории, но также включают упоминания насилия или приобретения оружия.

### ***self-harm***

Контент, который пропагандирует, поощряет или изображает акты членовредительства, такие как самоубийство, нанесение порезов и расстройства пищевого поведения.

### ***self-harm/intent***

Контент, в котором говорящий заявляет, что совершает или намеревается совершить акты самоповреждения, такие как самоубийство, нанесение порезов и расстройства пищевого поведения.

### ***self-harm/instructions***

Контент, поощряющий совершение актов самоповреждения, таких как самоубийство, нанесение себе порезов и расстройства пищевого поведения, или содержащий инструкции или советы о том, как совершать такие акты.

### ***sexual***

Контент, призванный вызвать сексуальное возбуждение, например, описание сексуальной активности или продвижение сексуальных услуг (за исключением полового воспитания и здорового образа жизни).

### ***violence/graphic***

Контент, в котором подробно описываются смерть, насилие или физические травмы.

### ***sexual/minors***

Сексуальный контент, включающий лицо моложе 18 лет.

### ***violence***

Контент, изображающий смерть, насилие или физические травмы.

# Как модерация может помочь моему проекту интерактивного дневника

Поскольку пользователи будут добавлять личные планы, эмоции и медиа, могут возникать ситуации, когда размещается запрещённый, оскорбительный или опасный контент. Модерация поможет фильтровать чрезмерно негативные или потенциально вредные сообщения (например, с суициальными мыслями) и направлять их в нужные службы поддержки.

Модерация, как ручная, так и автоматическая с помощью ИИ, может отслеживать такие случаи и принимать соответствующие меры.

**Thank you  
very much!**