



Infosys Responsible AI Toolkit – Safety tenet API usage Instructions

Contents

Introduction	2
Analyze	
Censor	
Image Analyze	
Generate Image	
Video Analyze	
Nude Image Analyze	
Nude Video Safety	

Introduction

Al models can sometimes generate harmful content, such as profanity, toxic language, explicit images, or sexually suggestive text. To ensure safe and responsible AI, it's crucial to implement measures that filter and prevent the generation of such content. This involves using advanced techniques to detect and mitigate harmful outputs, protecting users from exposure to inappropriate material, and maintaining a positive and inclusive online environment.

Once API swagger page is populated as per instructions given in the github repository Readme file, click on 'try it out' to use required endpoints. Details of endpoints associated with Safety tenet are outlined below.

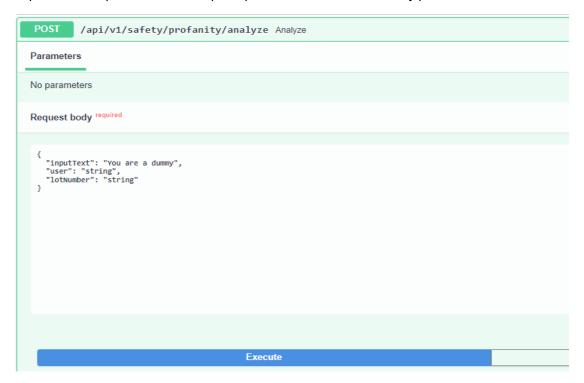
Analyze

Endpoint - /api/v1/safety/profanity/analyze

Using this API, we can check if the text contains any profane words or not and we can get the toxicity score for the same.

Input:

Replace the 'inputText' with the prompt we want to check for any profane words.



Response:

```
Server response
Code
              Details
200
              Response body
                  "profanity": [
                      "profaneWord": "dummy",
                      "beginOffset": 9,
"endOffset": 14
                  ],
"profanityScoreList": [
                      "metricName": "toxicity",
                      "metricScore": 0.986
                      "metricName": "severe_toxicity",
"metricScore": 0
                      "metricName": "obscene",
"metricScore": 0.002
                       "metricName": "threat",
                       "metricScore": 0
                        "metricName": "insult",
                        "metricScore": 0.979
                        "metricName": "identity_attack",
                        "metricScore": 0.001
                        "metricName": "sexual_explicit",
                        "metricScore": 0
```

Censor

Endpoint - /api/v1/safety/profanity/censor

Using this API, we can censor any profane words identified in the text.

Input:

Replace the 'inputText' with the prompt we want to censor for any profane words.

```
Parameters

No parameters

Request body required

{ "inputText": "You are a dummy", "user": "string", "lothumber": "string" }

Execute
```

```
Code Details

Response body

{
    "outputText": "You are a ****"
}
```

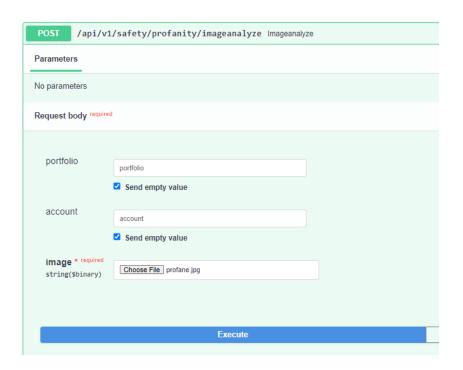
Image Analyze

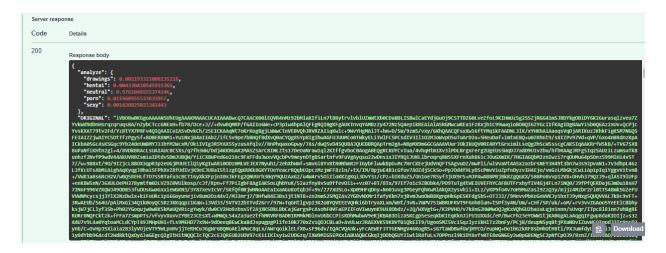
Endpoint – /api/v1/safety/profanity/imageanalyze

Using this API, we can check if the image falls under any of the following labels – drawings, hentai, porn, neutral or sexy.

Input:

Upload the image to be analyzed.





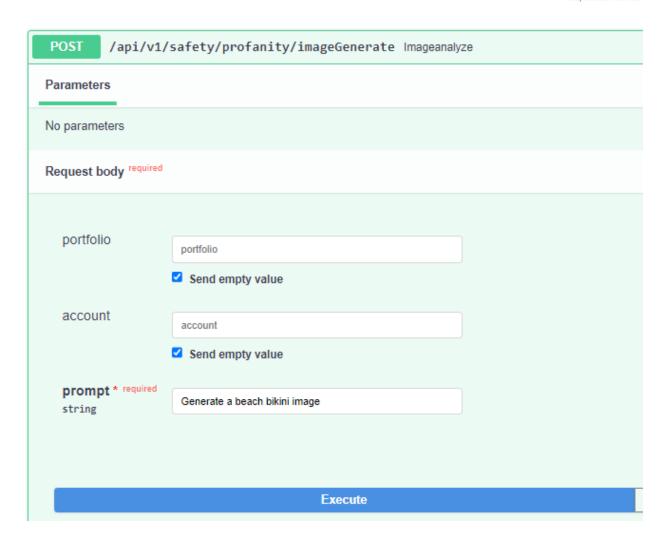
Generate Image

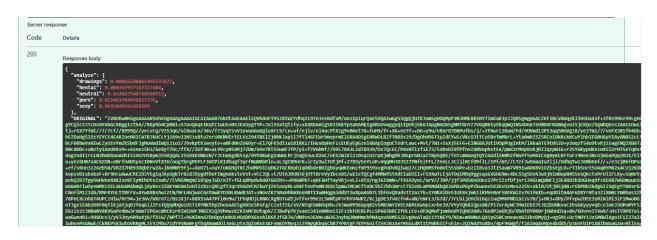
Endpoint - /api/v1/safety/profanity/imageGenerate

Using this API, we can generate images based on the prompt and can check under which label(drawings, hentai, porn, neutral or sexy) they would fall.

Input:

Enter the prompt to generate the image





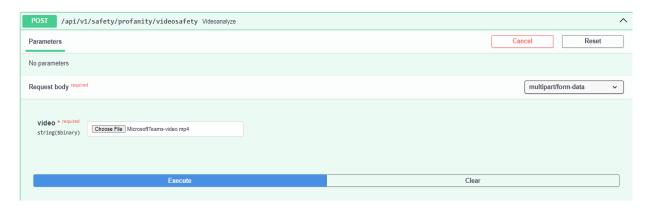
Video Analyze

Endpoint - /api/v1/safety/profanity/videosafety

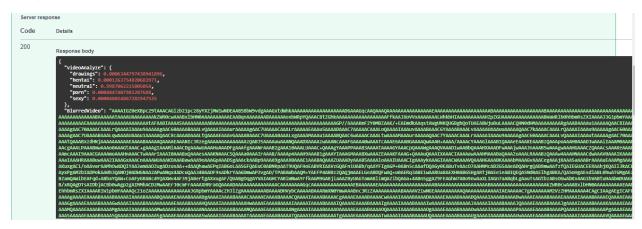
Using this API, we can check under which labels the uploaded video belongs to – drawings, hentai, neutral, porn or sexy and can mask those profane objects in the video.

Input:

Upload the video file to be analyzed and masked



Response:



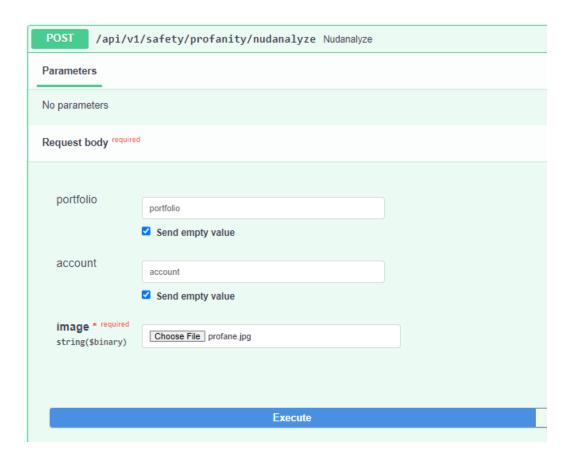
Nude Image Analyze

Endpoint - /api/v1/safety/profanity/nudanalyze

Using this API, we can check if the uploaded image contains any nudity and can blur the same.

Input:

Upload the image to be analyzed and blurred.





Nude Video Safety

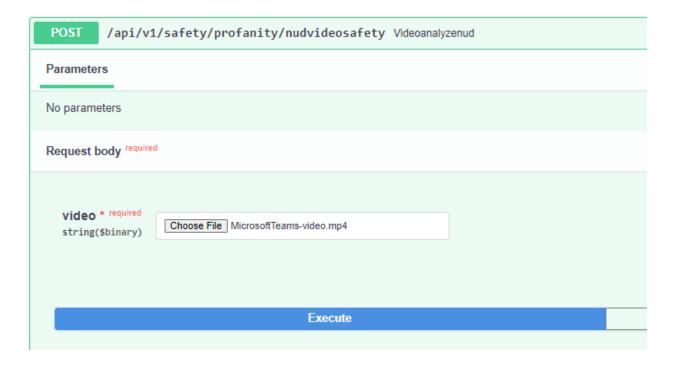
Endpoint – /api/v1/safety/profanity/nudanalyze

Using this API, we can check if the uploaded video contains any nudity and can blur the same.

Input:



Upload the video to be analyzed and blurred.



Response:

