

Infosys Responsible AI Toolkit – Safety tenet API usage Instructions

Contents

About Safety	2
Dependencies	2
Features & API End Points	3
Safety APIs	3
Add Profane Words	3
Analyze	4
Censor	5
Image Analyze	6
Generate Image	7
Video Analyze	8
Nude Image Analyze	9
Nude Video Safety	10
Malicious URL Detection	11
CSV File Analyze	12
Endpoints Usage Flow	13



About Safety

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.

Dependencies

The imageGenerate API depend on responsible-ai-llm repository. Please follow the setup instructions in the README files of both repositories to configure them. Ensure that both services are up and running before interacting with the imageGenerate API. Other APIs do not have any dependencies.



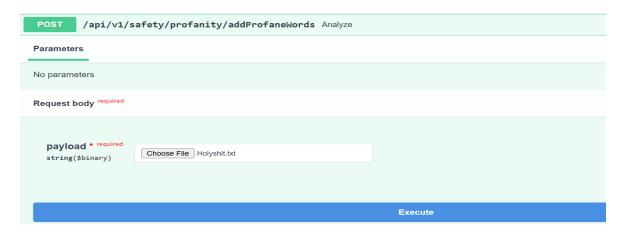
Features & API End Points

Safety APIs

Add Profane Words

Endpoint: /api/v1/safety/profanity/addProfaneWords Profane words can be added to the existing profane words list.

Input: Give the profane words .txt document as input



Response: Returns "success" if the profane words are added successfully





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.

```
POST /api/v1/safety/profanity/analyze Analyze

Parameters

No parameters

Request body required

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

Response: Returns the list of profane words detected with positions and the label scores for toxicity, severe_toxicity, obscene, threat,insult, identity_attack and sexual_explicit.



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.

```
Post /api/v1/safety/profanity/censor Censor

Parameters

No parameters

Request body required

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

Response: Mask the words identified as profane words.

```
Code Details

Response body

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

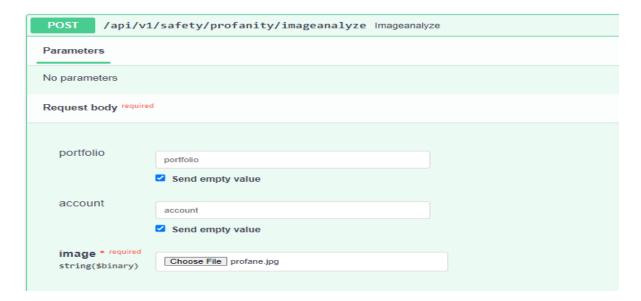


Image Analyze

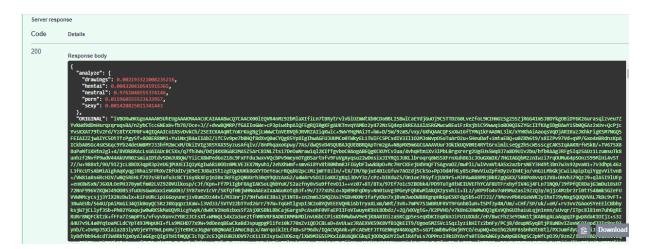
Endpoint - /api/v1/safety/profanity/imageanalyze

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.



Response: Analyzes the image and returns the score for labels - drawings, hentai, neutral, porn and sexy for the image and base64-encoded format for image.



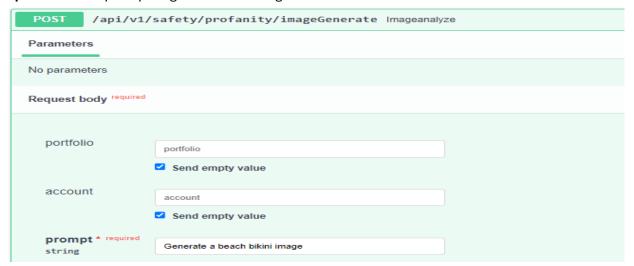


Generate Image

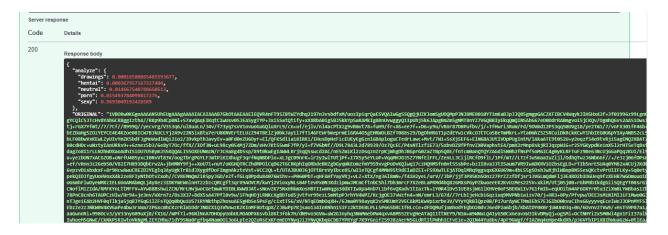
Endpoint – /api/v1/safety/profanity/imageGenerate

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



Response: Analyzes an image for content types (e.g., hentai, porn, sexy, neutral, drawings) and returns a classification score along with the base64-encoded image data.



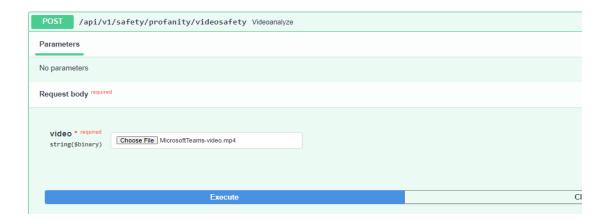


Video Analyze

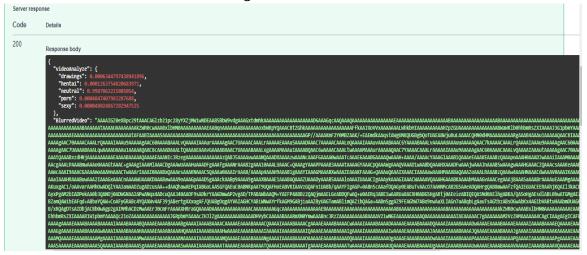
Endpoint - /api/v1/safety/profanity/videosafety

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: Analyzes a video for content types (e.g., hentai, porn, sexy, neutral, drawings) and returns a classification score along with the base64-encoded video data.





Nude Image Analyze

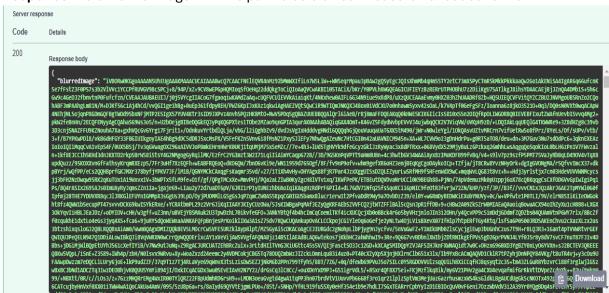
Endpoint - /api/v1/safety/profanity/nudanalyze

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.



Response: Returns the image with the part to be blurred in base-64 encoded format.



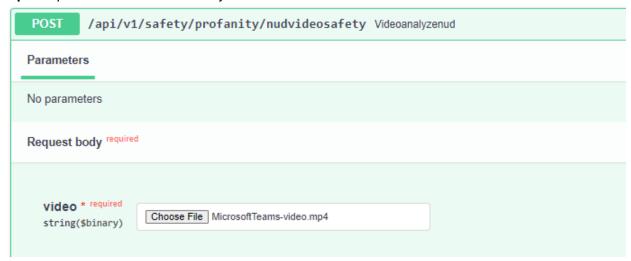


Nude Video Safety

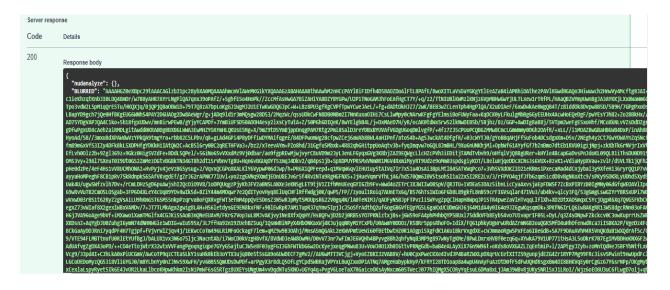
Endpoint - /api/v1/safety/profanity/nudanalyze

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: Returns the video with the part to be blurred in base-64 encoded format.





Malicious URL Detection

Endpoint: /api/v1/safety/profanity/maliciousUrl

Detects the malicious url and returns passed or unmoderated.

Input: Enter the link to be verified in the 'inputText' field

```
Parameters

No parameters

Request body required

{
    "inputText": "https://example.com",
    "maliciousThreshold": 0.5,
    "user": "string",
    "lockNumber": "string"
}
```

Response: Returns the response as PASSED (No malicious URL detected).

```
Code Details

Response body

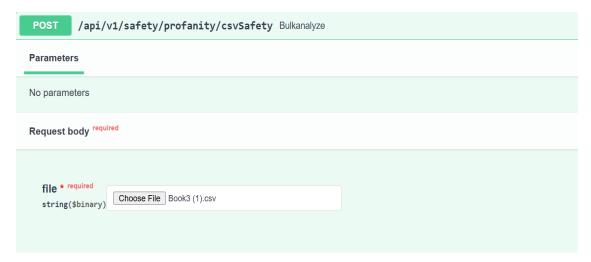
{
    "prompt": "https://example.com",
    "scoreList": [],
    "result": "PASSED",
    "threshold": 0.5,
    "time": "0.185s"
}
```

CSV File Analyze

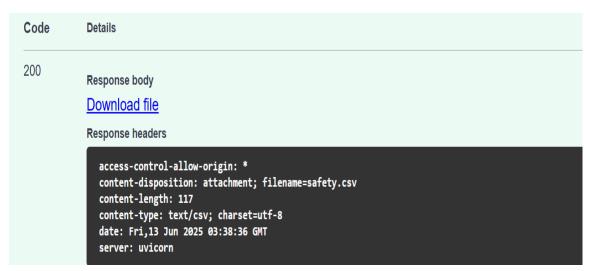
Endpoint: /api/v1/safety/profanity/csvSafety -

Analyze the csv file content and remove the malicious content from it.

Input: Upload CSV file that needs to be analyzed



Response: Returns a "Download File" link that can be downloaded.





Endpoints Usage Flow

Endpoint	Description
/api/v1/safety/profanity/addProfaneWords	Profane words can be added to the existing profane words list.
/api/v1/safety/profanity/analyze	Detect the profane words and return the words with the toxic scores.
/api/v1/safety/profanity/censor	Masks identified profane words in output.
/api/v1/safety/profanity/imageanalyze	Detects NSFW image based on parameters (porn, sexy, neutral, drawings, hentai) in input image.
/api/v1/safety/profanity/imageGenerate	Generates an image based on the prompt given as an input.
/api/v1/safety/profanity/videosafety	Detects NSFW video based on parameters (porn, sexy, neutral, drawings, hental) in input video.
/api/v1/safety/profanity/nudanalyze	Detects the specific parts of nudity in the given image.
/api/v1/safety/profanity/nudvideosafety	Detects the specific parts of nudity in the given video.
/api/v1/safety/profanity/maliciousUrl	Detects the malicious url and returns passed or unmoderated
/api/v1/safety/profanity/csvSafety	Analyze the csv file content and remove the malicious content from it.