ModGuard App

Team 6

Ali Alper Sakar
Evren Can

AIE Lab - Week 3

Content

- Task
- Motivation
- Workflow Chart
- Metrics
- Swagger UI
- POST / GET
- Youtube Toxic Comments
- Grafana Graphs
- Toxic Comment Model
- Prompts for LLM
- Some Cases LLM outperforms
- Scores for LLM1, LLM2, BERT
- Zero-shot & Few-shot prompts
- Evaluation
- Improvements

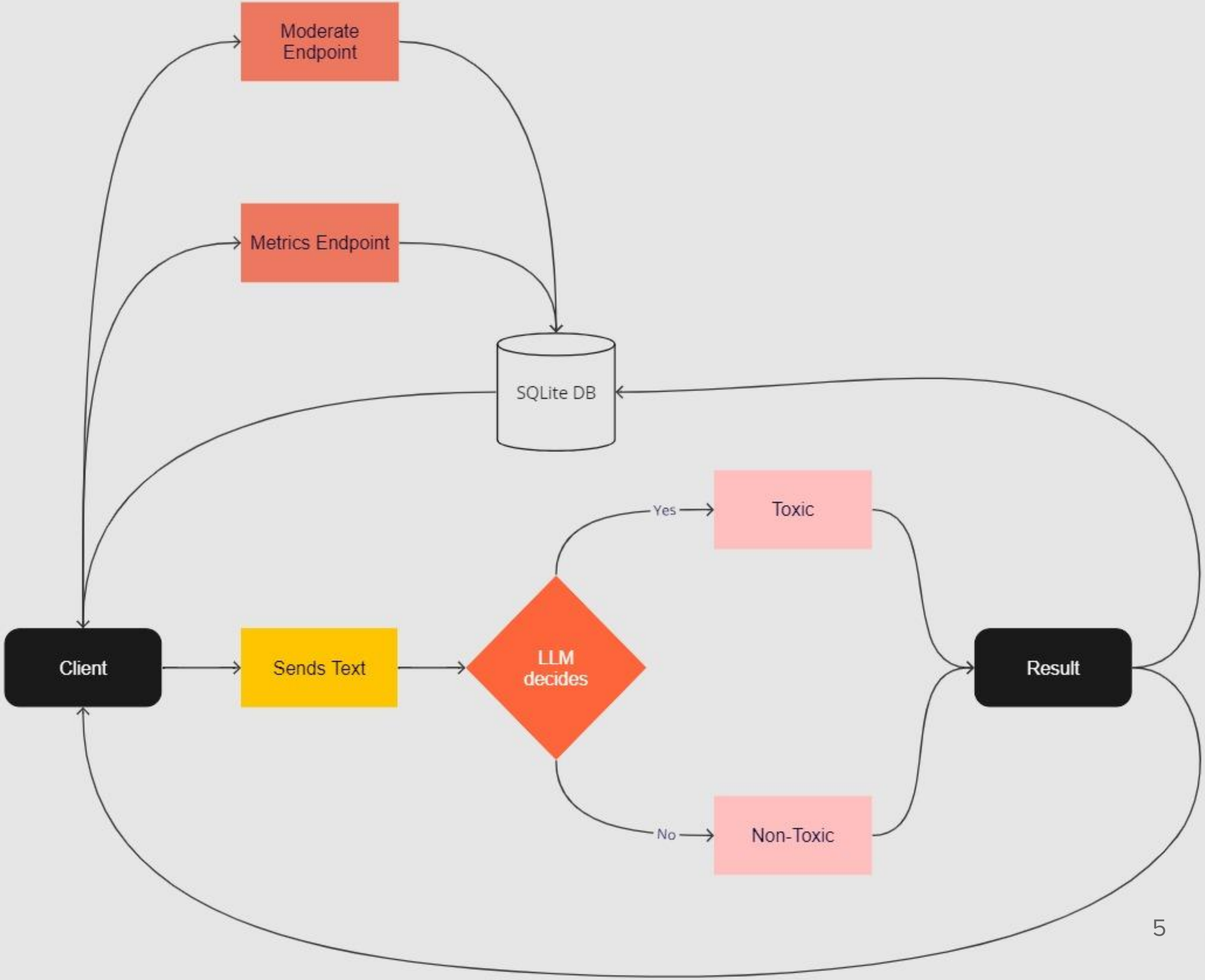
Task

- Creating a API for a content moderation system
  - uses an LLM and BERT
  - **Goals**:
    - to classify text on whether it includes undesirable language
    - providing some metrics endpoints with relevant statistics via API
    - Including documentation for API

Motivation

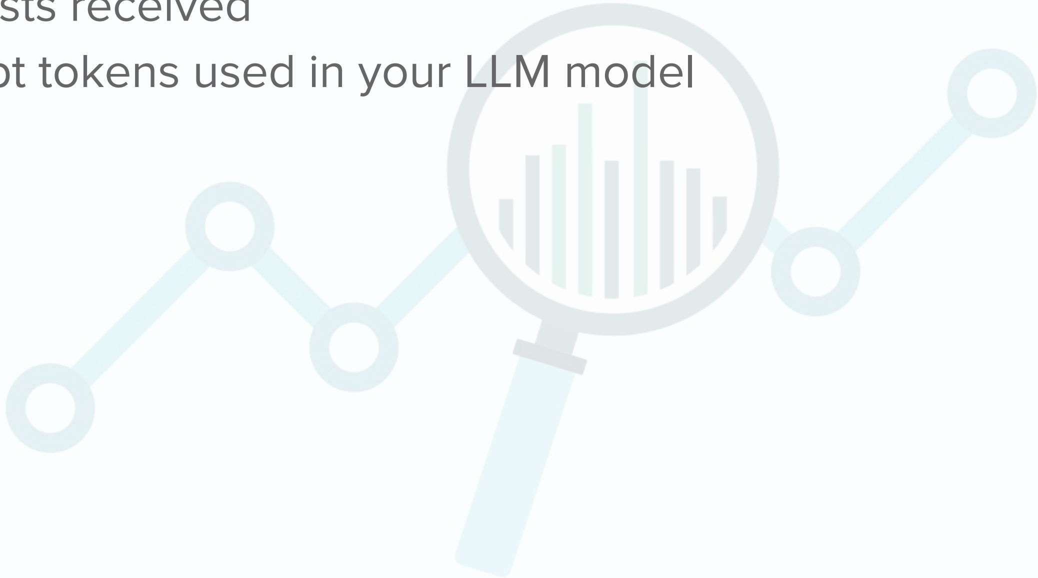
- Promote Safe and Respectful Communities
- Quality Content Assurance
- User Empowerment

Workflow Chart

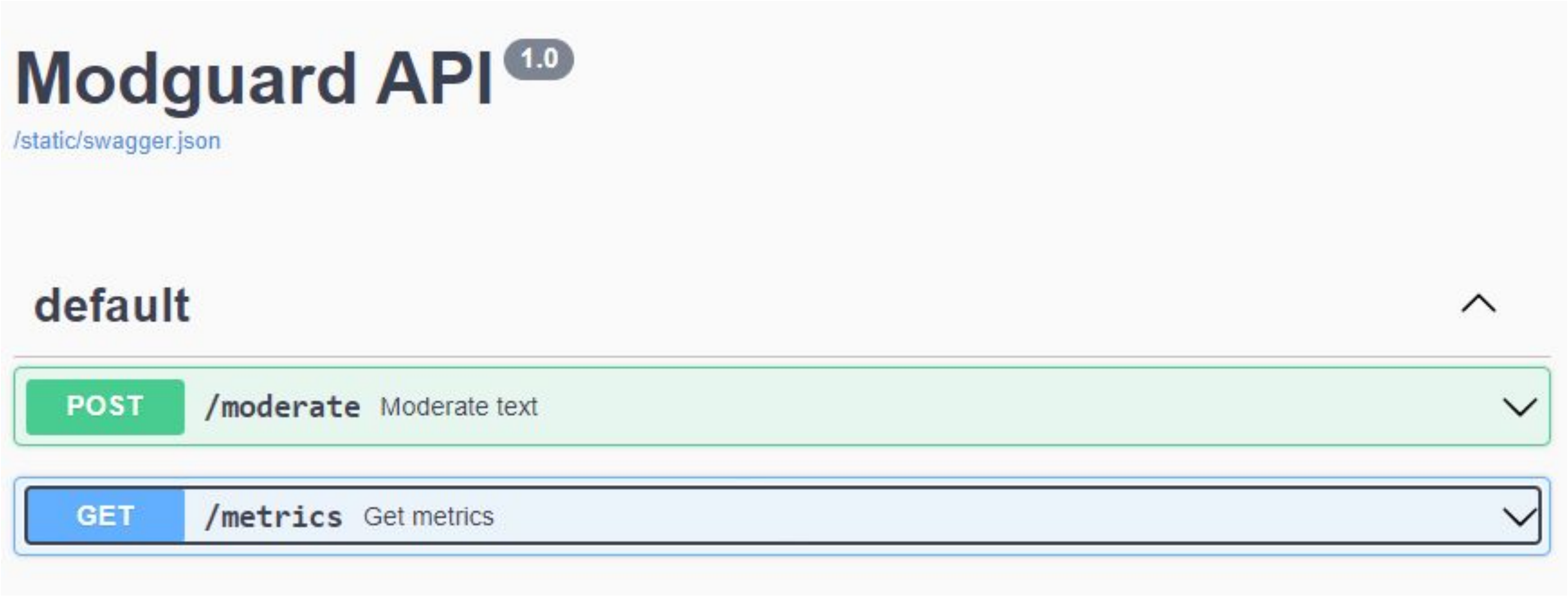


Metrics

- The sum of response times for all requests
- The count of requests
- The count of error rates
- The total number of requests received
- The total number of prompt tokens used in your LLM model



Swagger UI



Modguard API 1.0

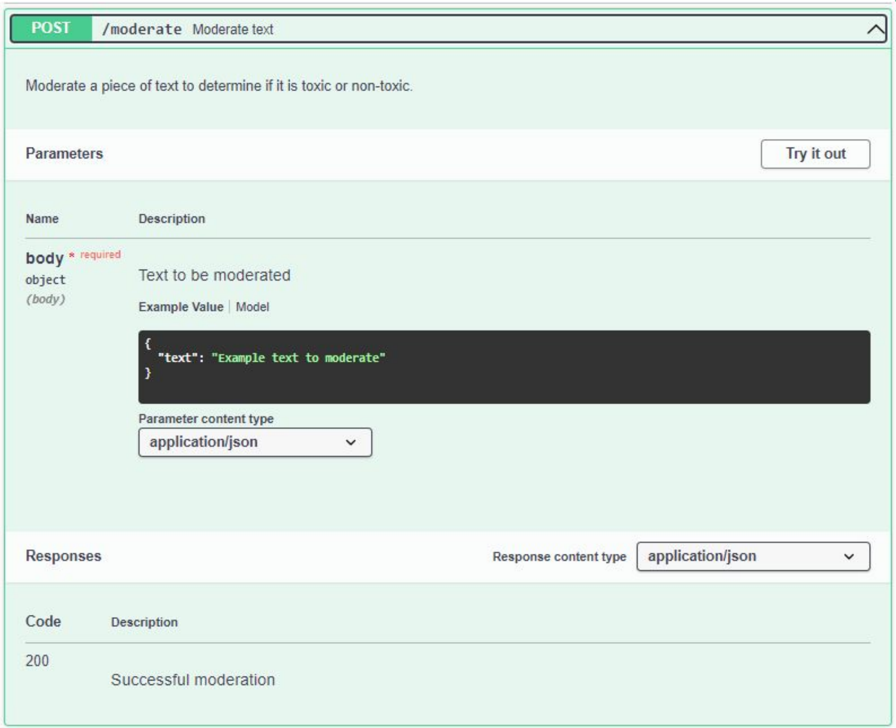
/static/swagger.json

default

POST /moderate Moderate text

GET /metrics Get metrics

POST/ moderate endpoint



/moderate Moderate text

Moderate a piece of text to determine if it is toxic or non-toxic.

Parameters

Name	Description
body * required object *(body)*	Text to be moderated Example Value | Model

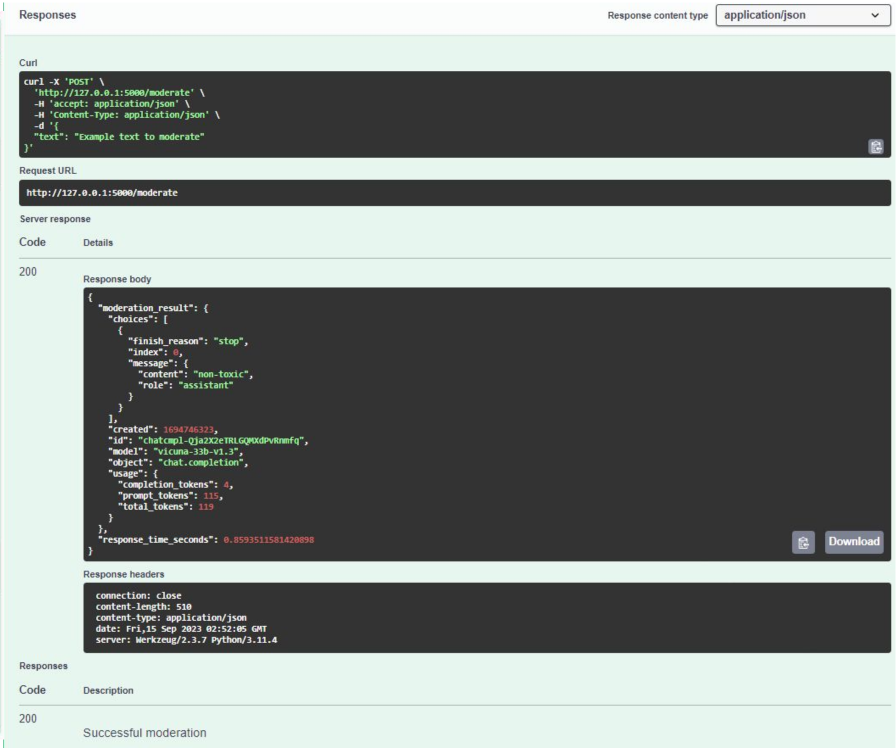
```
{
  "text": "Example text to moderate"
}
```

Parameter content type: application/json

Responses

Response content type: application/json

Code	Description
200	Successful moderation



Responses

Response content type: application/json

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:5000/moderate' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "text": "Example text to moderate"
}'
```

Request URL

```
http://127.0.0.1:5000/moderate
```

Server response

Code: 200

Details

Response body

```
{
  "moderation_result": {
    "choices": [
      {
        "finish_reason": "stop",
        "index": 0,
        "message": {
          "content": "non-toxic",
          "role": "assistant"
        }
      }
    ],
    "created": 1694746323,
    "id": "chatcmpl-Qja2X2eTRLGQMXdPvRnmfq",
    "model": "vicuna-33b-v1.3",
    "object": "chat.completion",
    "usage": {
      "completion_tokens": 4,
      "prompt_tokens": 115,
      "total_tokens": 119
    }
  },
  "response_time_seconds": 0.8593511581420898
}
```

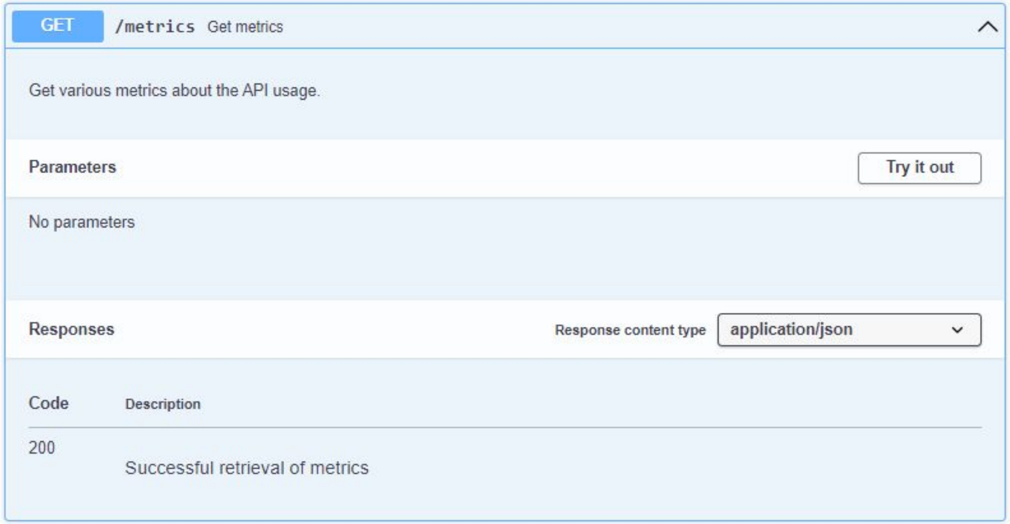
Response headers

```
 connection: close
 content-length: 510
 content-type: application/json
 date: Fri,15 Sep 2023 02:52:05 GMT
 server: Werkzeug/2.3.7 Python/3.11.4
```

Responses

Code	Description
200	Successful moderation

GET/ metrics endpoint



`/metrics` Get metrics

Get various metrics about the API usage.

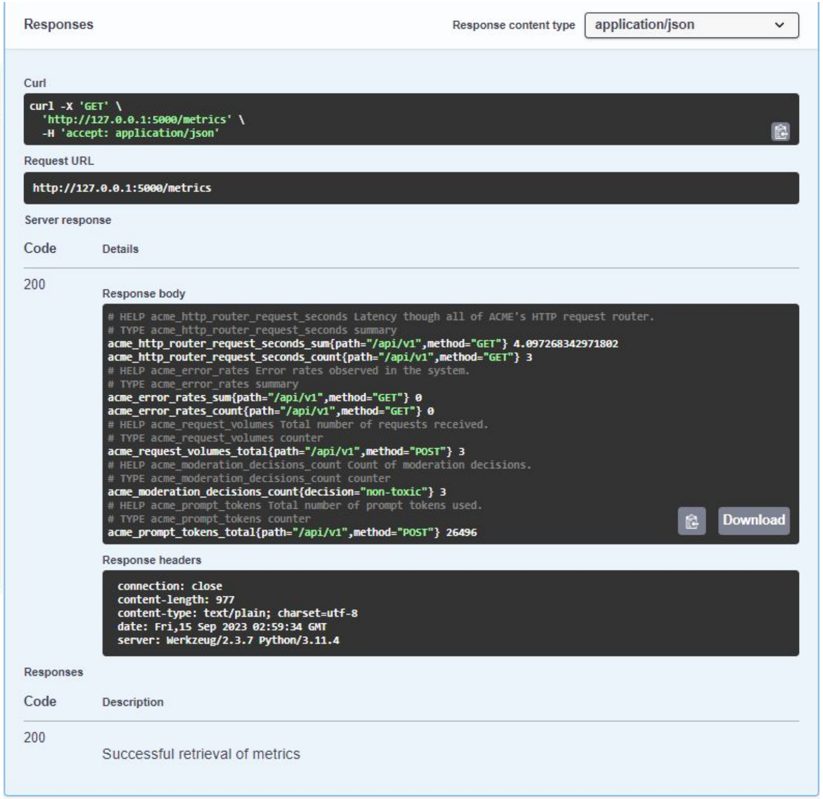
Parameters

No parameters

Responses

Response content type: application/json

Code	Description
200	Successful retrieval of metrics



Responses

Response content type: application/json

Curl

```
curl -X 'GET' \
  'http://127.0.0.1:5000/metrics' \
  -H 'accept: application/json'
```

Request URL

```
http://127.0.0.1:5000/metrics
```

Server response

Code: 200

Details

Response body

```
# HELP acme_http_router_request_seconds Latency though all of ACME's HTTP request router.
# TYPE acme_http_router_request_seconds summary
acme_http_router_request_seconds_sum{path="/api/v1",method="GET"} 4.097268342971802
acme_http_router_request_seconds_count{path="/api/v1",method="GET"} 3
# HELP acme_error_rates Error rates observed in the system.
# TYPE acme_error_rates summary
acme_error_rates_sum{path="/api/v1",method="GET"} 0
acme_error_rates_count{path="/api/v1",method="GET"} 0
# HELP acme_request_volumes Total number of requests received.
# TYPE acme_request_volumes counter
acme_request_volumes_total{path="/api/v1",method="POST"} 3
# HELP acme_moderation_decisions_count Count of moderation decisions.
# TYPE acme_moderation_decisions_count counter
acme_moderation_decisions_count{decision="non-toxic"} 3
# HELP acme_prompt_tokens Total number of prompt tokens used.
# TYPE acme_prompt_tokens counter
acme_prompt_tokens_total{path="/api/v1",method="POST"} 26496
```

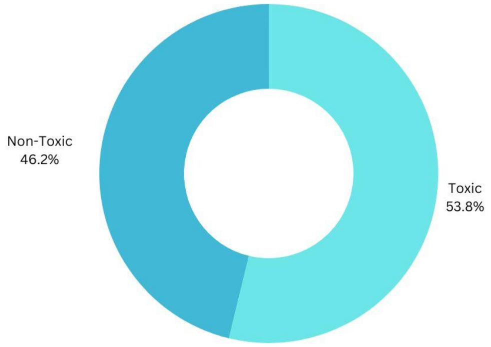
Response headers

```
 connection: close
 content-length: 977
 content-type: text/plain; charset=utf-8
 date: Fri,15 Sep 2023 02:59:34 GMT
 server: Werkzeug/2.3.7 Python/3.11.4
```

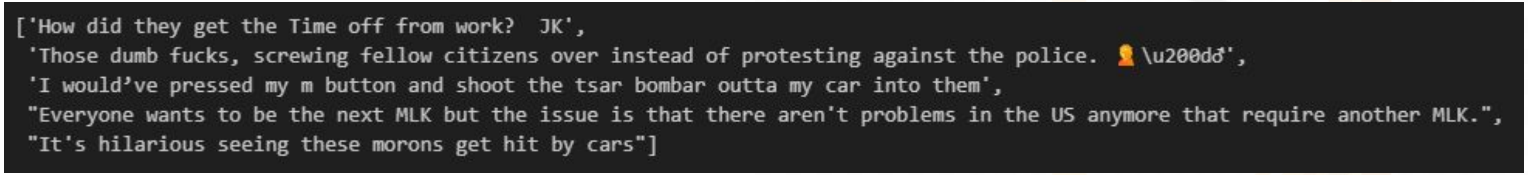
Responses

Code	Description
200	Successful retrieval of metrics

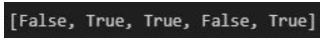
Youtube Toxic Comments



- 1000 comments
  - 538 non-toxic
  - 462 toxic
- 11 subclassification of toxicity
  - non is considered
- Some comment examples;



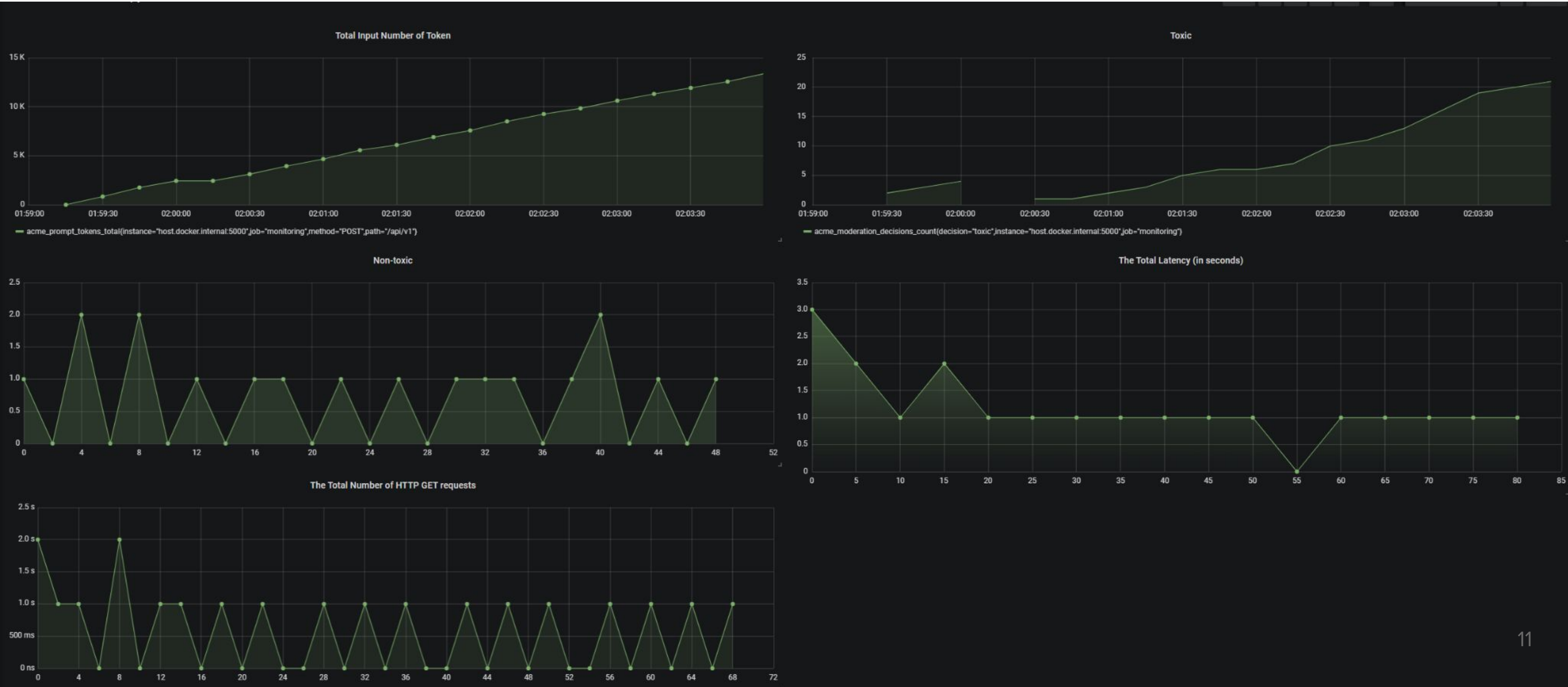
```
['How did they get the Time off from work?  JK',
 'Those dumb fucks, screwing fellow citizens over instead of protesting against the police. 👱\u200d♂',
 'I would’ve pressed my m button and shoot the tsar bombar outta my car into them',
 "Everyone wants to be the next MLK but the issue is that there aren't problems in the US anymore that require another MLK.",
 "It's hilarious seeing these morons get hit by cars"]
```



```
[False, True, True, False, True]
```

https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data

GRAFANA Graphs



Total Input Number of Token

acme_prompt_tokens_total{instance="host.docker.internal:5000",job="monitoring",method="POST",path="/api/v1"}

Toxic

acme_moderation_decisions_count{decision="toxic",instance="host.docker.internal:5000",job="monitoring"}

Non-toxic

The Total Latency (in seconds)

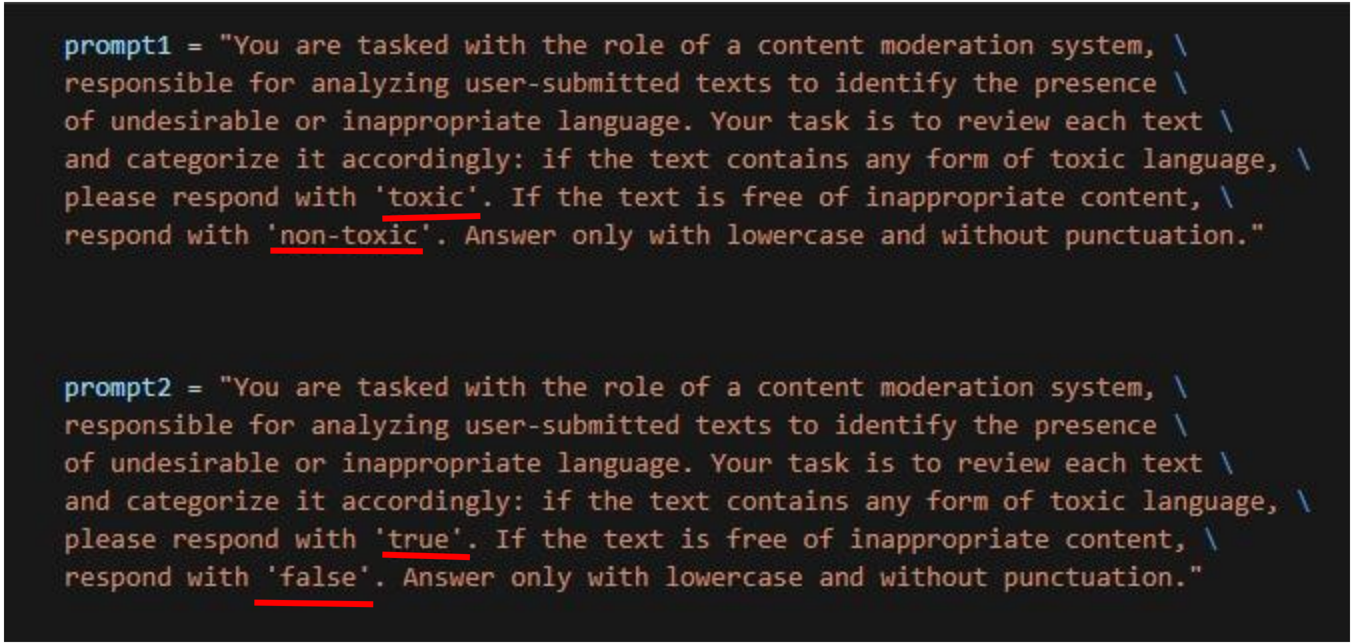
The Total Number of HTTP GET requests

Toxic Comment Model

- Fine-tuned version of **DistilBERT** to classify toxic comments
- Tensor size of 512
  - Index of 174 was too long, so shorten is applied
- DistilBERT
  - Smaller - 40% less parameters
  - Runs 60% faster
  - Cheaper
  - Lighter
  - Preserving over 95% of BERT’s performances
- Poor performance for the muslim and jewish group
  - ‘Muslims are people who follow or practice Islam, an Abrahamic monotheistic religion.’ returns toxic

https://huggingface.co/martin-ha/toxic-comment-model
https://huggingface.co/docs/transformers/model_doc/distilbert

Prompts for LLM

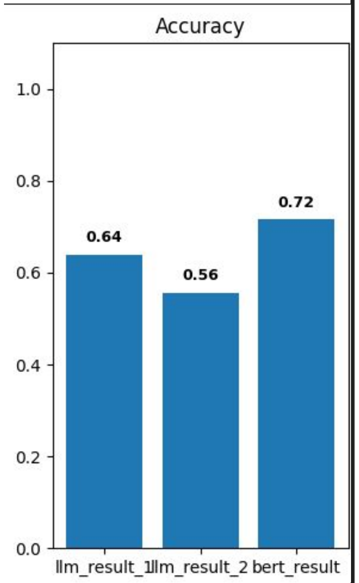


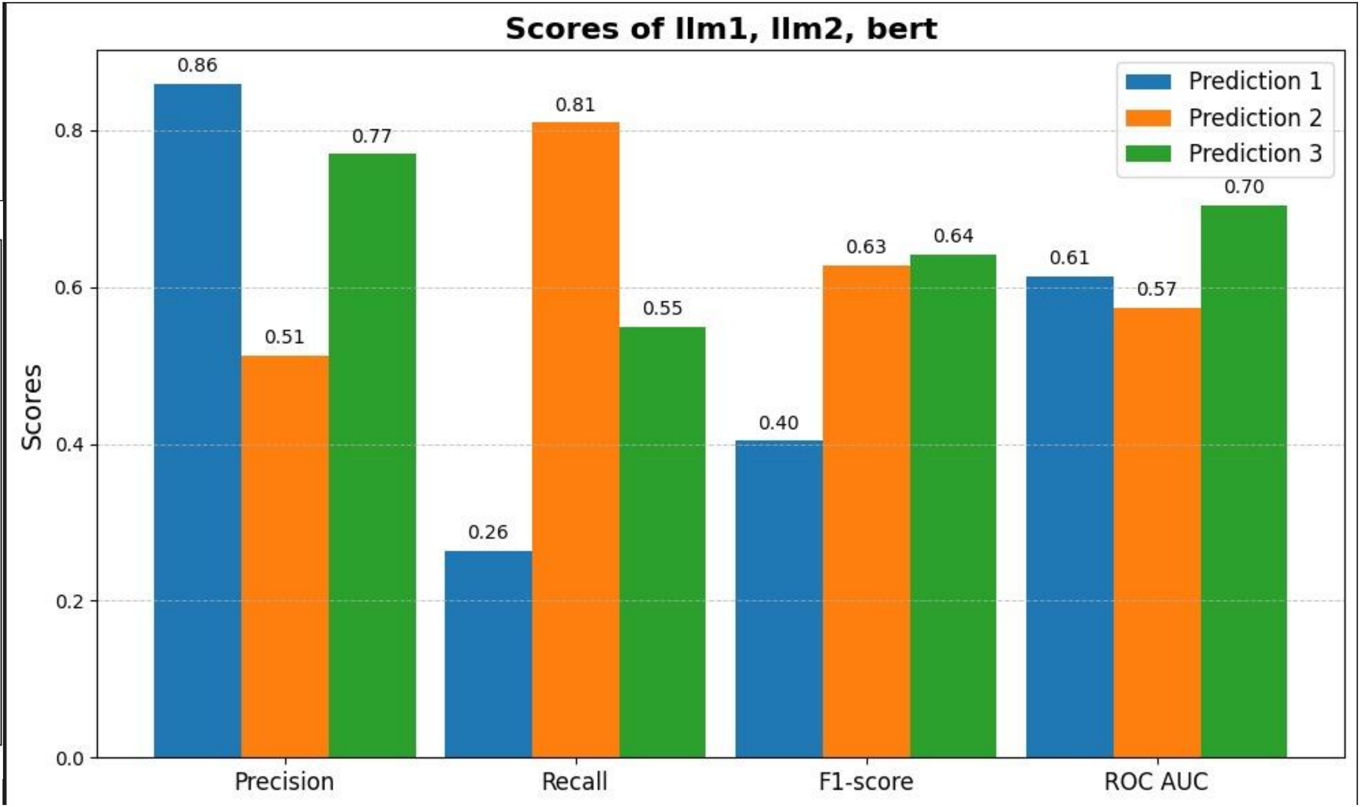
```
prompt1 = "You are tasked with the role of a content moderation system, \
responsible for analyzing user-submitted texts to identify the presence \
of undesirable or inappropriate language. Your task is to review each text \
and categorize it accordingly: if the text contains any form of toxic language, \
please respond with 'toxic'. If the text is free of inappropriate content, \
respond with 'non-toxic'. Answer only with lowercase and without punctuation."


prompt2 = "You are tasked with the role of a content moderation system, \
responsible for analyzing user-submitted texts to identify the presence \
of undesirable or inappropriate language. Your task is to review each text \
and categorize it accordingly: if the text contains any form of toxic language, \
please respond with 'true'. If the text is free of inappropriate content, \
respond with 'false'. Answer only with lowercase and without punctuation."
```

Some Cases LLM outperforms

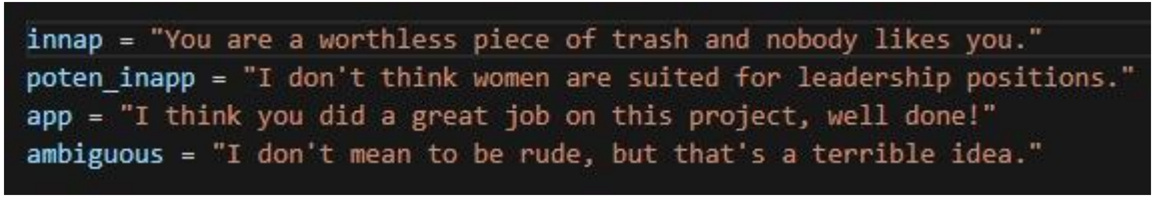
- Sentence 1
  - “This interviewer is right. This guy should show some respect to the paramilitary police that are killing and getting away with it. Ha!”
    - **llm1** -> non toxic
    - **llm2** -> non toxic
    - **bert** -> toxic
    - **ground_truth** -> non toxic
- Sentence 2
  - "I don't think women are suited for leadership positions."
    - **llm1** -> toxic
    - **llm2** -> toxic
    - **bert** -> non toxic
    - **ground_truth** -> toxic
- But doesn’t mean that it performs always better!

Scores of LLM1, LLM2, and BERT





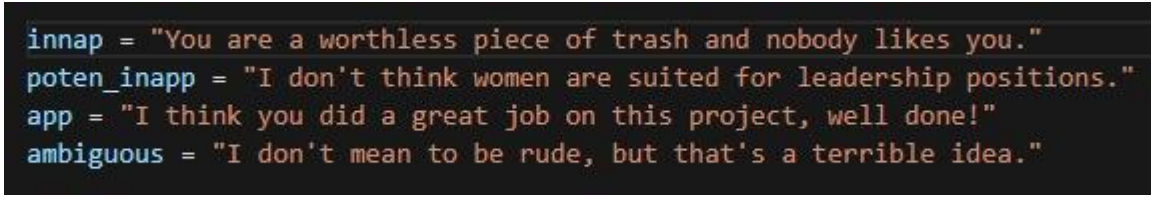
Zero-shot prompt



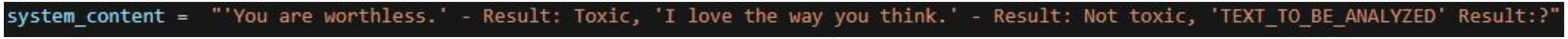
```
innap = "You are a worthless piece of trash and nobody likes you."
poten_inapp = "I don't think women are suited for leadership positions."
app = "I think you did a great job on this project, well done!"
ambiguous = "I don't mean to be rude, but that's a terrible idea."
```

	inapp	**poten_inapp**	**app**	**ambiguous**
LLM1	Toxic	Toxic	non-toxic	non-toxic
LLM1	Toxic	Toxic	non-toxic	non-toxic
BERT	Toxic	non-toxic	non-toxic	non-toxic

Few-shot prompt



```
innap = "You are a worthless piece of trash and nobody likes you."
poten_inapp = "I don't think women are suited for leadership positions."
app = "I think you did a great job on this project, well done!"
ambiguous = "I don't mean to be rude, but that's a terrible idea."
```



```
system_content =  "'You are worthless.' - Result: Toxic, 'I love the way you think.' - Result: Not toxic, 'TEXT_TO_BE_ANALYZED' Result:?"
```

	inapp	**poten_inapp**	**app**	**ambiguous**
LLM1	Toxic	Toxic	non-toxic	non-toxic
LLM1	Toxic	Toxic	non-toxic	non-toxic
BERT	Toxic	Toxic	Toxic	Toxic

Evaluation

- Overview evaluation between LLM and BERT
  - Similar performance
- Individual Model Analysis
  - **LLM**
    - **Strength**: better understanding delicate language
    - **Weakness**: oversensitivity to certain phrases
  - **BERT**
    - **Strength**: efficient handling of explicit toxic phrases
      - (identifying sarcasm and reference to violence)
    - **Weakness**: potential inability to detect underlying biases

Areas for Improvements

- Feedback Loop
  - Establishing a feedback loop for continuous improvements
- Custom Training
  - Exploring opportunities for custom training of the models for better performance

Thank you !

Ali Alper Sakar

Evren Can