

Adventure time! A journey for flagging dangerous multimodal content using LLMs

Michał Mikołajczak
Tech Lead & CEO, datarabbit

5th Apr 2024

Short Bio/quick introduction

- Michal Mikolajczak
- Worked mainly in companies combining the fields of IT and medicine – primarily on stuff connected to machine learning and (medical) image processing
- Lot of startups involvement (back in the days CTO of startup acquired by NASDAQ company), hence a lot of exposure to other aspects of products (infrastructure/data/architecture)
- Currently owner and tech lead in **Datarabbit** – ML/cloud-focused software house



Agenda

1. Business context: content moderation – what is it and why is it needed?
2. Automating the problem.
3. LLMs to the rescue!
4. Moving beyond text data.
5. Final solution pipeline.
6. Lessons learned/future steps.
7. Q&A.

What is content moderation?

Content moderation is a practice of monitoring and regulating user-generated content on digital platforms (e.g. such as social media websites, forums, online marketplaces).

The primary goal of it is to ensure that content complies with different standards respective to the particular platform.



Why is it needed?

- Preventing spam and scams.
- Protecting users from harmful content.
- Safeguarding brand reputation.
- Complying with legal/regulatory requirements (e.g. DSA in the EU).

The Content Moderation Report

According to respondents from Business Insider Intelligence's 2019 Digital Trust Survey:



Respondents believed



were **least likely** to show **deceptive content**

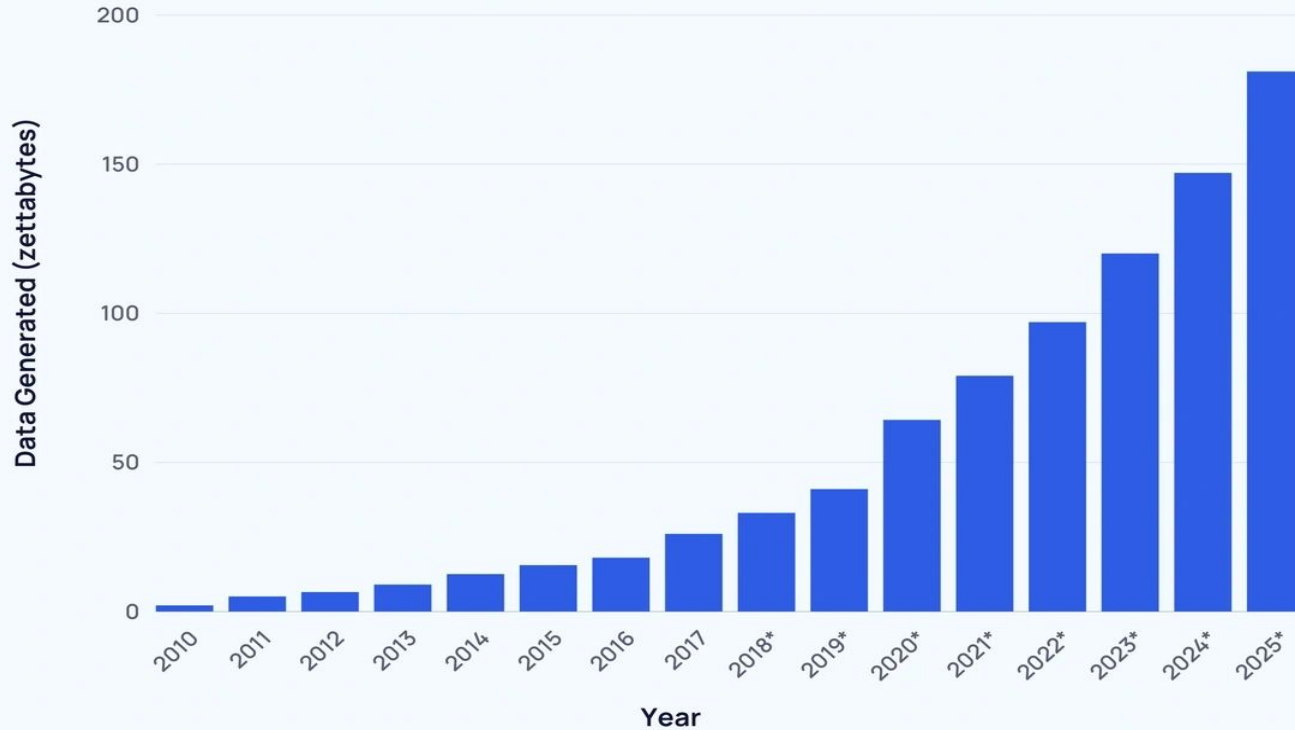
27%

said they'd **stop using** a **social platform** if it continued to **allow harmful content**

70%

believe stakeholders other than social companies should have **final say** in determining what **content is permitted**

Global Data Generated Annually



Social media account to ~13% of it (video excluded)

For human moderators responsible for reviewing all that data – not the brightest perspective!



Our (target) use case

- Communication platform, can't reveal specific industry, but can be think of as social media with specific audience.
- Audience includes minors – that are the primary priority/concern.
- Automation/assistance is required – human moderators are to constrained resource to handle everything.
- Currently limited data and labels available.
- Some “standard” harmful content types to detect (e.g. sexual, vulgarity), but also very non-standard categories e.g. region/culture specific or related to the ultimate target which is...



ONLINE PREDATOR

Out of the box services?



amazon Rekognition

sightengine

Out of the box services?

There are solutions available – but each had posed a number of problems:

- Lack of support for every modality we're interested in (text, images, videos later in the future).
- **There were some performance issues during internal tests for officially supported content categories.**
- Lack of some categories that we wanted/needed to support due to client constraints – we needed some very specific, region-based categories (imagine symbols/clothing forbidden in some cultures).
- Scaling/quotas problems were present in some of the tools.



Labels detected:
None

Expected: guns



Labels detected:
Class: nudity, Value: 0.99

Expected: violence

Out of the box services?

There are solutions available – but each had posed a number of problems:

- **Lack of support for every modality we're interested in (text, images, videos later in the future).**
- There were some performance issues during internal tests for officially supported content categories.
- **Lack of some categories that we wanted/needed to support due to client constraints – we needed some very specific, region-based categories (imagine symbols/clothing forbidden in some cultures).**
- Scaling/quotas problems were present in some of the tools.

What about training
custom models?

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

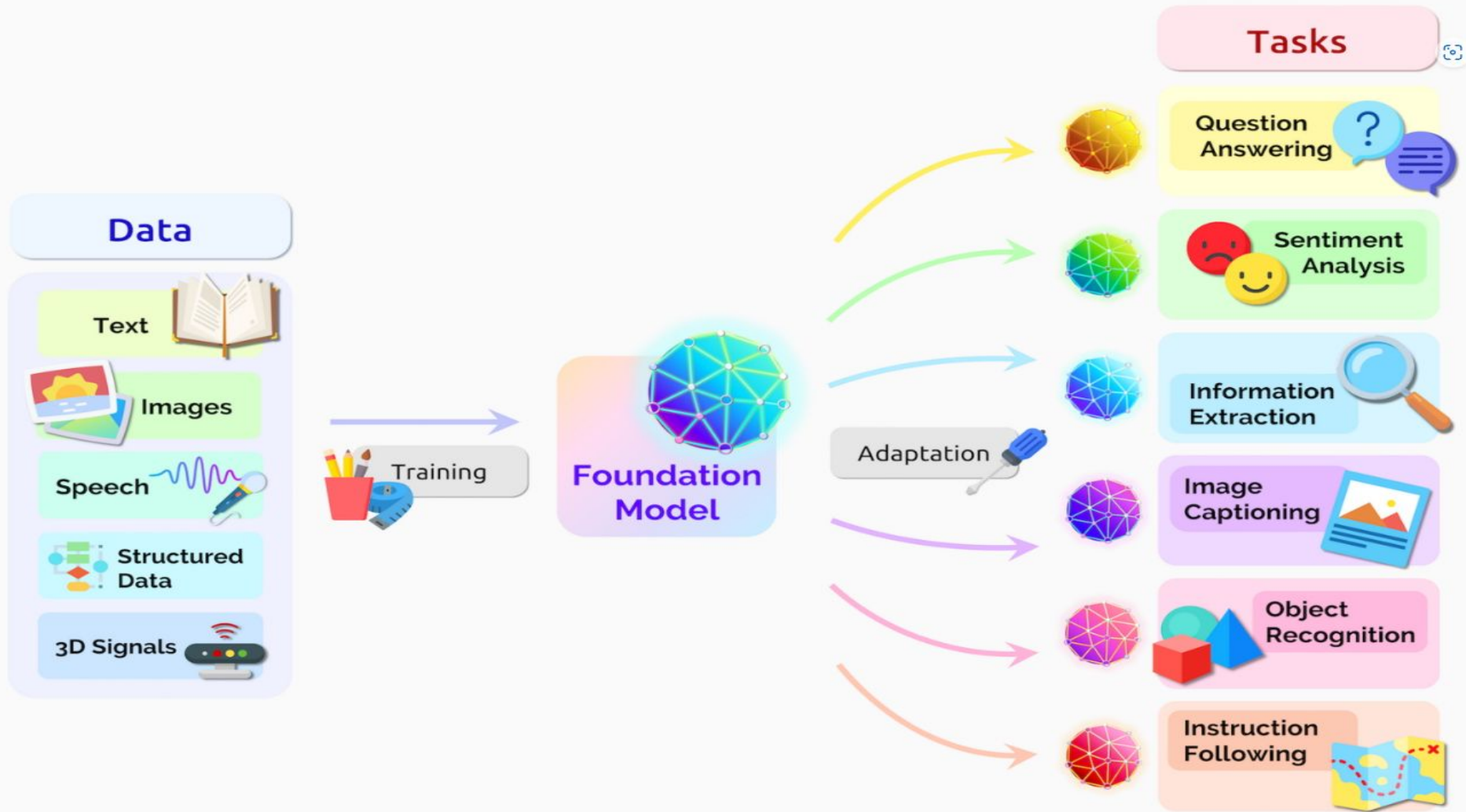
WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.

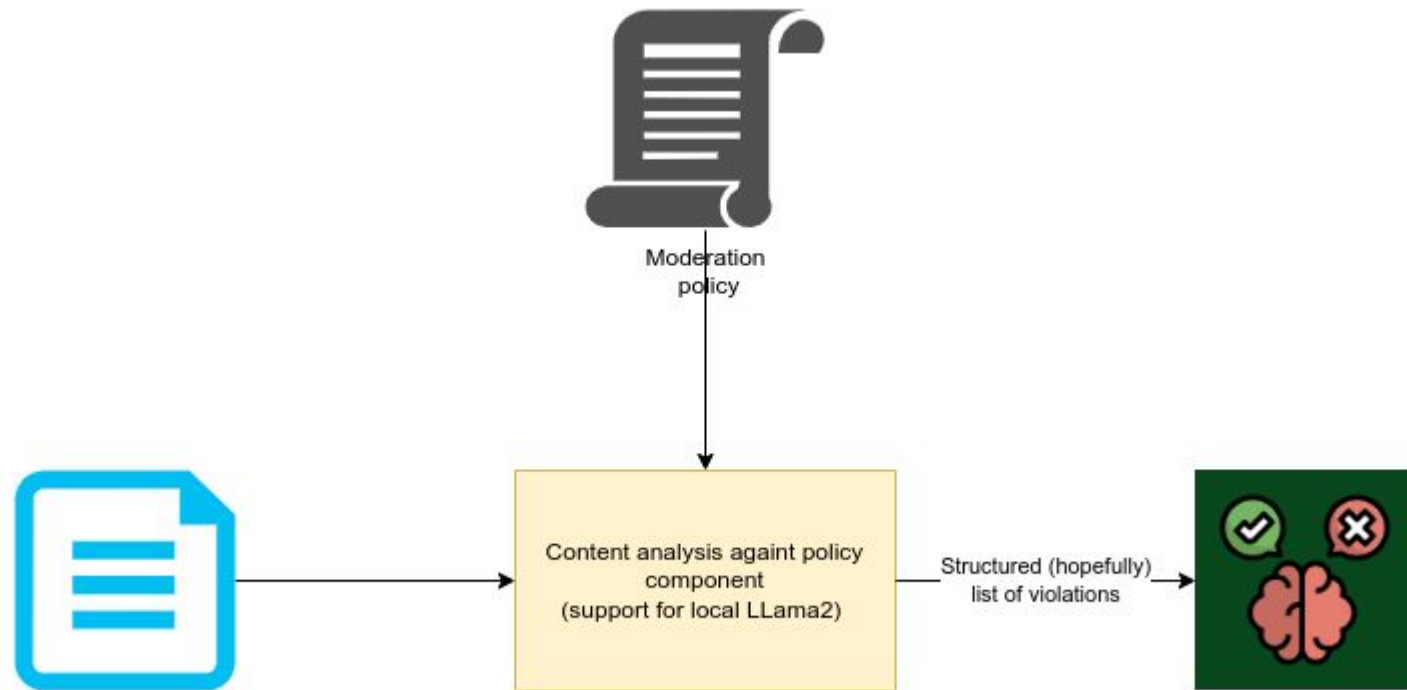




IN TERMS OF DATA - WE HAVE NO DATA



Initial pipeline



Example (image) moderation policy

```
MODERATION_POLICY_MINORS = """
```

1. Explicit Nudity: content contains Nudity, Graphic Male Nudity, Graphic Female Nudity, Sexual Activity, Illustrated Explicit Nudity and Adult Toys.
2. Suggestive: content contains Female Swimwear Or Underwear, Male Swimwear Or Underwear, Partial Nudity, Barechested Male, Revealing Clothes and Sexual Situations.
3. Violence: content contains Graphic Violence Or Gore, Physical Violence, Weapon Violence, Weapons and Self Injury.
4. Visually Disturbing: content contains Emaciated Bodies, Corpses, Hanging, Air Crash and Explosions And Blasts.
5. Rude Gestures: content contains Middle Finger.
6. Drugs: content contains Drug Products, Drug Use, Pills and Drug Paraphernalia.
7. Tobacco: content contains Tobacco Products and Smoking.
8. Alcohol: content contains Drinking and Alcoholic Beverages.
9. Gambling: content contains Gambling.
10. Hate Symbols: content contains Nazi Party, White Supremacy and Extremist.

```
"""
```

You're trained, experienced, and professional content moderator who flags content that contains unsafe content.

Unsafe content includes any of the following categories:

```
{moderation_policy}.
```



You receive content (video or image description or text) summaries from users, and need to evaluate them against the provided policies with unsafe content categories. You generate a list of which of the categories the content is unsafe based on the provided description, as well as the reason behind it in a following JSON list format:

```
{  
  "violations":  
  [  
    {  
      "category": "xxx",  
      "reason": "the reason is ..."  
    },  
    {  
      "category": "yyy",  
      "reason": "the reason is ..."  
    },  
    ...  
  ]  
}
```

Categories should not be duplicated and limited to the ones explicitly stated in moderation policy. If one of the categories is violated for multiple reasons, there should be no duplicated list items for this violation, but it should be reduced to a single list item for this "category" with all the reasons combined under the "reason" field.

If the content is entirely safe, output an empty list:

```
{  
  "violations": []  
}
```

Here is the content summary:

```
##### Content evaluation #####
```

```
{content_evaluation}
```

```
##### Task #####
```

Generate an evaluation of content safety against given policies.

Memory to self-host LLM

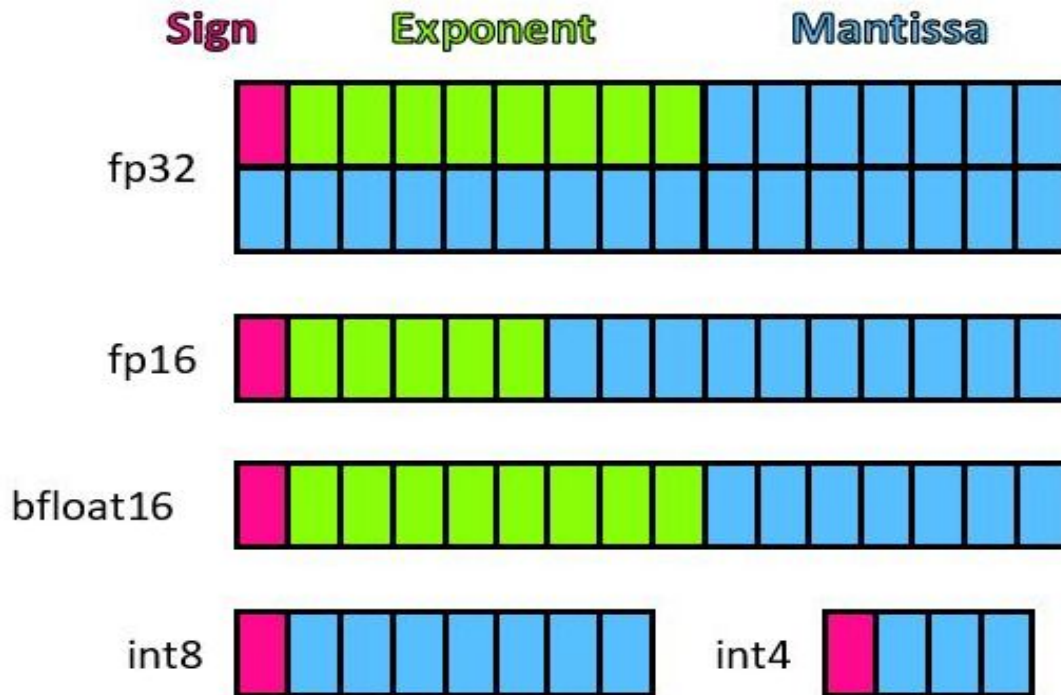
- Our assumption: need to fits 24GB VRAM for PoC (supporting text + images).
- According to formula, in the base version we wouldn't be able to fit **70B** param model (**336GB**) – but also not even **7B** (**33.6GB**).

$$M = \frac{(P * 4B)}{(32/Q)} * 1.2$$


Symbol	Description
M	GPU memory expressed in Gigabyte
P	The amount of parameters in the model. E.g. a 7B model has 7 billion parameters.
4B	4 bytes, expressing the bytes used for each parameter
32	There are 32 bits in 4 bytes
Q	The amount of bits that should be used for loading the model. E.g. 16 bits, 8 bits or 4 bits.
1.2	Represents a 20% overhead of loading additional things in GPU memory.

Quantization

- Neural networks weights were historically usually stored in 32-bit floating point format.
- Quantization is a set of techniques to represent the weights and activations with lower-precision data types such as float16, int8, or even smaller.





int4 

ok, so basically
im very smol

Quantization

- Results: significantly reduced memory bandwidth/footprint.
- Results: proportionally reduced space required for model storage.
- Results: noticeably improved inference speed.
- As precision is reduced: accuracy is obviously affected in the process (the question is if that tradeoff is acceptable).
- There are various methods/tools capable of performing quantization out there: usable post, and during training (the latter help mitigate potential quantization accuracy loss).
- **For 13B model with int4: 62.5GB -> ~8GB**

Going beyond text data

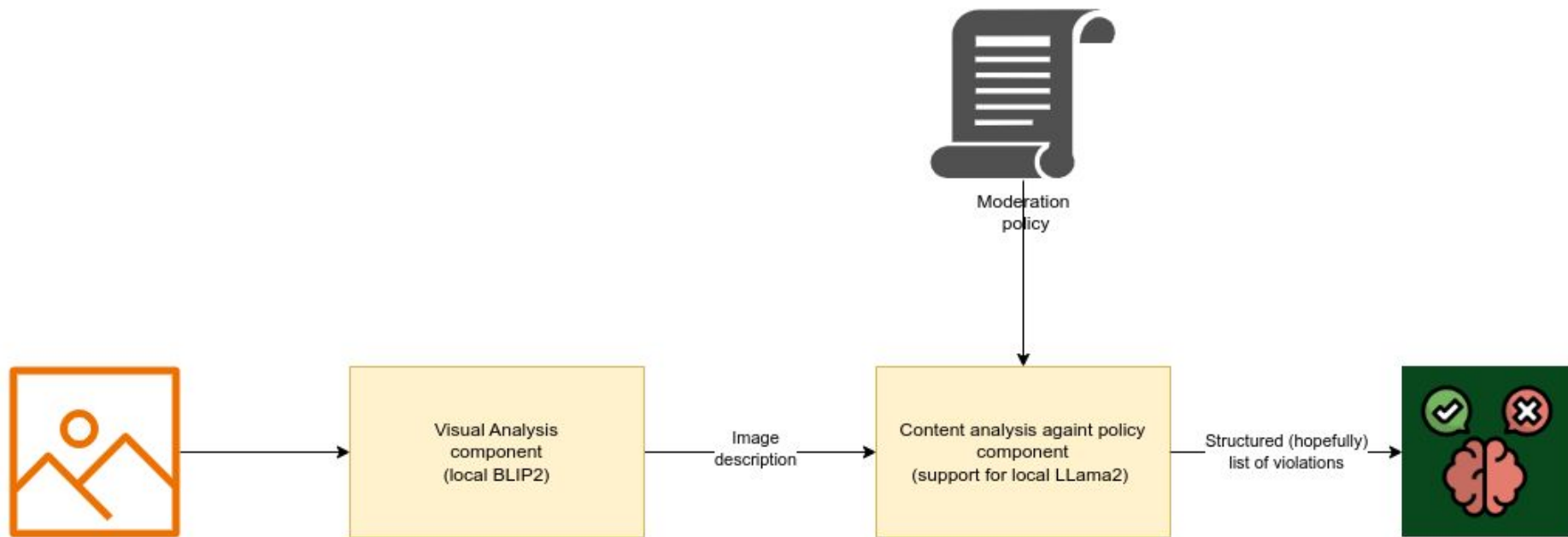
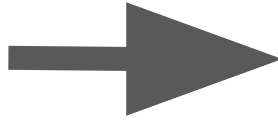


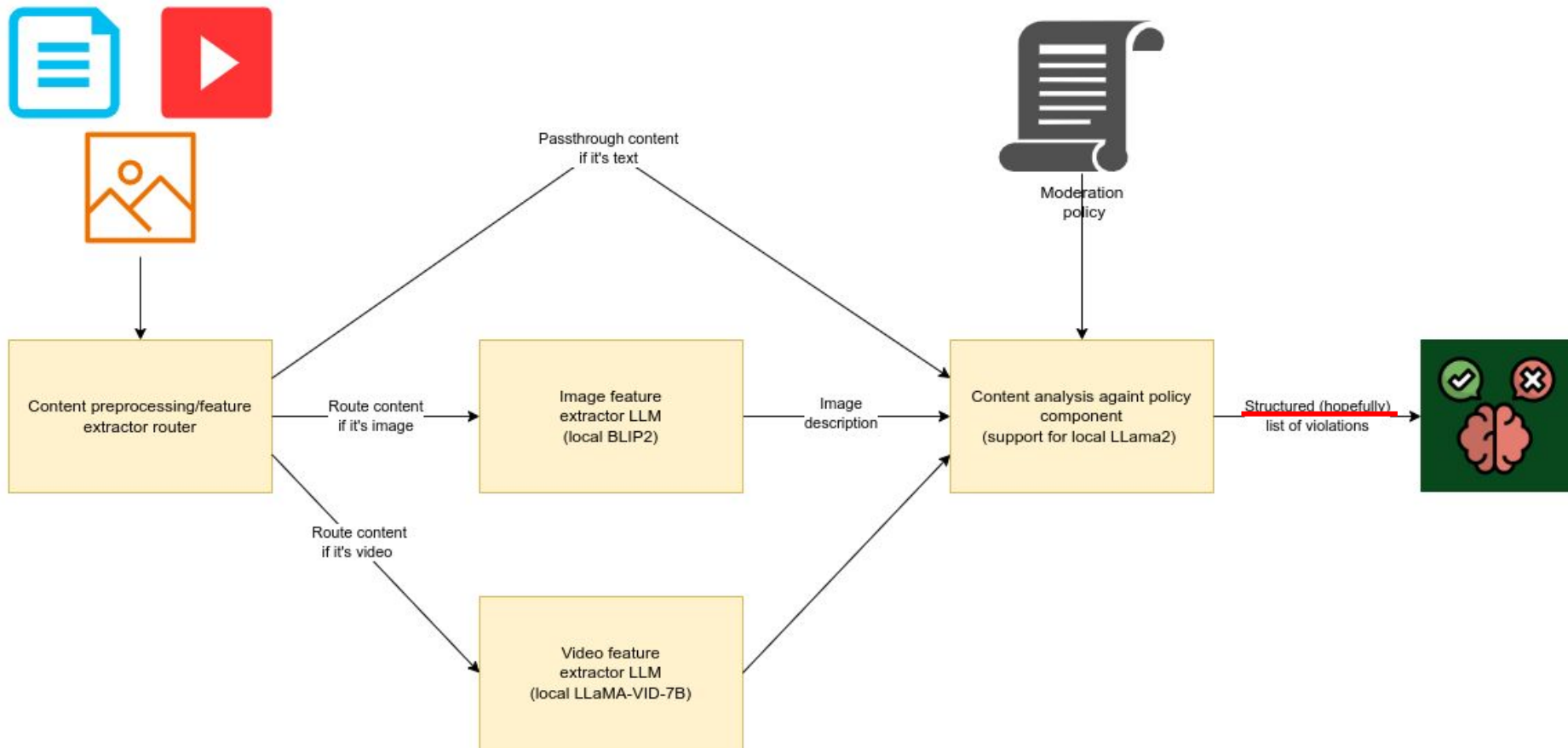
Image Descriptor

```
IMAGE_CONTENT_CHECKLIST = [  
    "Does this photo contain complete naked person?",  
    "Does this photo contain topless person?",  
    "Does this photo contain person in lingerie?",  
    "Does this photo contain sexual content?",  
    "Does this photo contain adult toys?",  
    "Does this photo contain person in swimwear?",  
    "Does this photo contain violent situation?",  
    "Does this photo contain blood?",  
    "Does this photo contain harassment words?",  
    "Does this photo contain weapons, or weapon violence?",  
    "Are there persons fighting in this photo?",  
    "Does this photo contain self-injury?",  
    "Does this photo contain corpses?",  
    "Does this photo contain air or traffic crashes?",  
    "Does this photo contain explosions or blasts?",  
    "Does this photo contain rude gestures?",  
    "Does this photo contain person showing middle finger?",  
    "Does this photo contain pills, drugs or drug paraphernalia?",  
    "Does this photo contain drug use?",  
    "Does this photo contain tobacco products?",  
    "Does this photo contain smoker or smoking?",  
    "Does this photo contain alcohol?",  
    "Does this photo contain person drinking beverages like beer?",  
    "Does this photo contain gambling?",  
    "Does this photo contain nazi party symbols?",  
    "Does this photo contain white supremacy symbols?",  
    "Does this photo contain extremist symbols?",  
    "Does this photo contain contact information?",  
]
```



```
Question: Are there persons fighting in this photo?  
Answer: no  
  
Question: Does this photo contain self-injury?  
Answer: no  
  
Question: Does this photo contain corpses?  
Answer: no  
  
Question: Does this photo contain air or traffic crashes?  
Answer: no  
  
Question: Does this photo contain explosions or blasts?  
Answer: no  
  
Question: Does this photo contain rude gestures?  
Answer: no  
  
Question: Does this photo contain person showing middle finger?  
Answer: no  
  
Question: Does this photo contain pills, drugs or drug paraphernalia?  
Answer: no  
  
Question: Does this photo contain drug use?  
Answer: no  
  
Question: Does this photo contain tobacco products?  
Answer: no  
  
Question: Does this photo contain smoker or smoking?  
Answer: no  
  
Question: Does this photo contain alcohol?  
Answer: no
```

Going beyond text data – supporting multimodality



Desired/correctly formatted examples

```
[  
  ModerationPolicyViolation(  
    category='Alcohol',  
    reason='The photo contains a person  
           drinking alcoholic beverages.'  
  )  
]
```

```
[  
  ModerationPolicyViolation(  
    category='Explicit Nudity',  
    reason='the reason is the photo contains sexual  
content'  
  ),  
  ModerationPolicyViolation(  
    category='Suggestive',  
    reason='the reason is the photo contains a  
topless person and person in lingerie'  
  )  
]
```

Incorrectly formatted examples

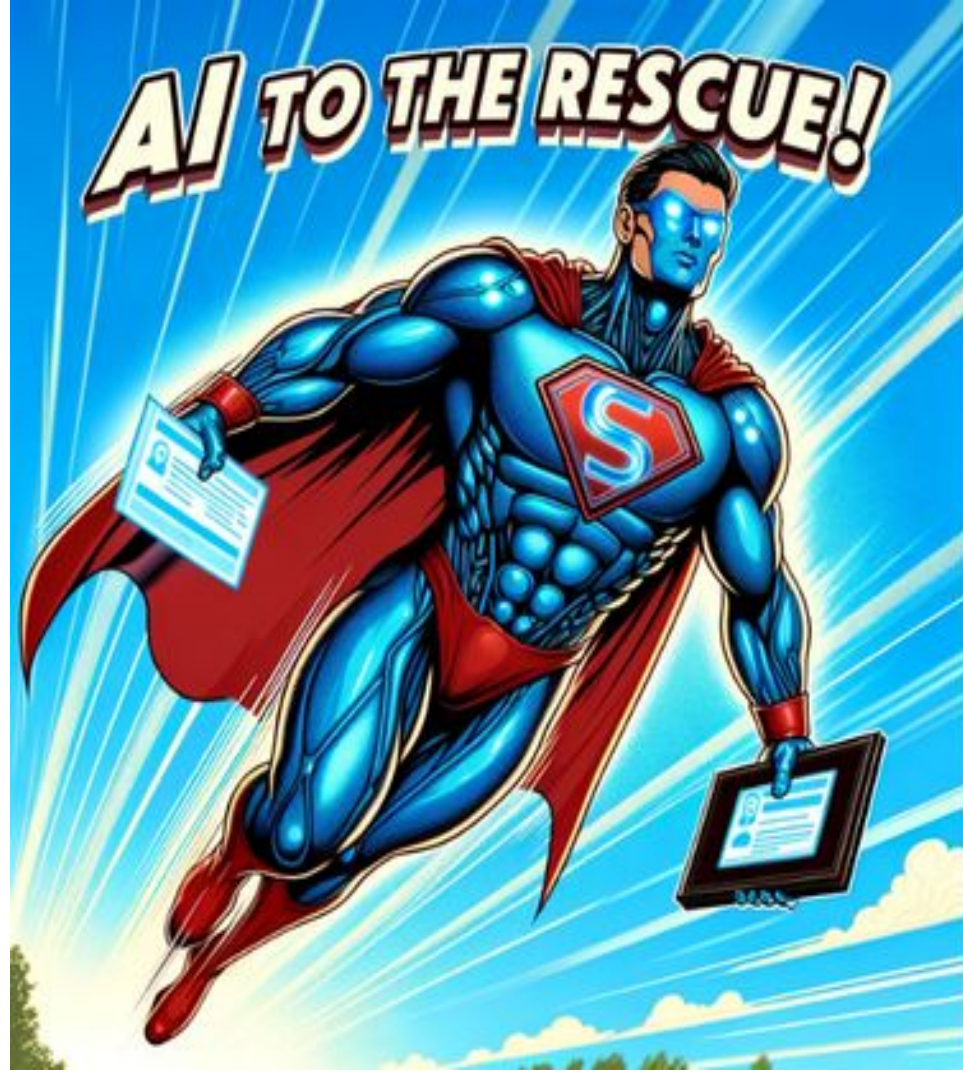
```
[  
  ModerationPolicyViolation(  
    category='none',  
    reason='No unsafe content was found  
           in the photo.'  
  )  
]
```

I've detected following violations of moderation policy:

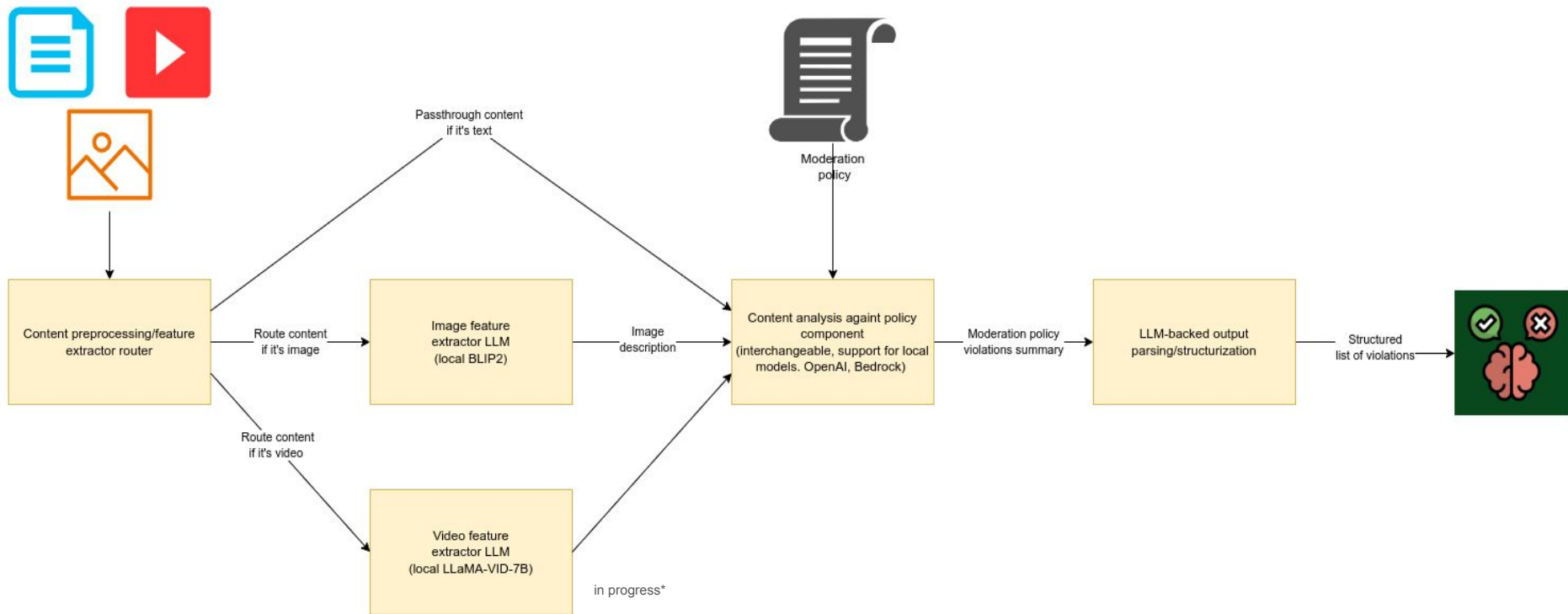
1. Explicit Nudity: the reason is the photo contains sexual content
2. Suggestive: the reason is the photo contains a topless person and person in lingerie
3. Rude Gestures: the reason is the photo contains rude gestures other than showing middle finger

What can you do when your model has a problem with outputting a correct/properly formatted response?

Simple – use yet another AI!



The final pipeline



Evaluation (simplified + image only)

Simplified evaluation (for now/in progress):

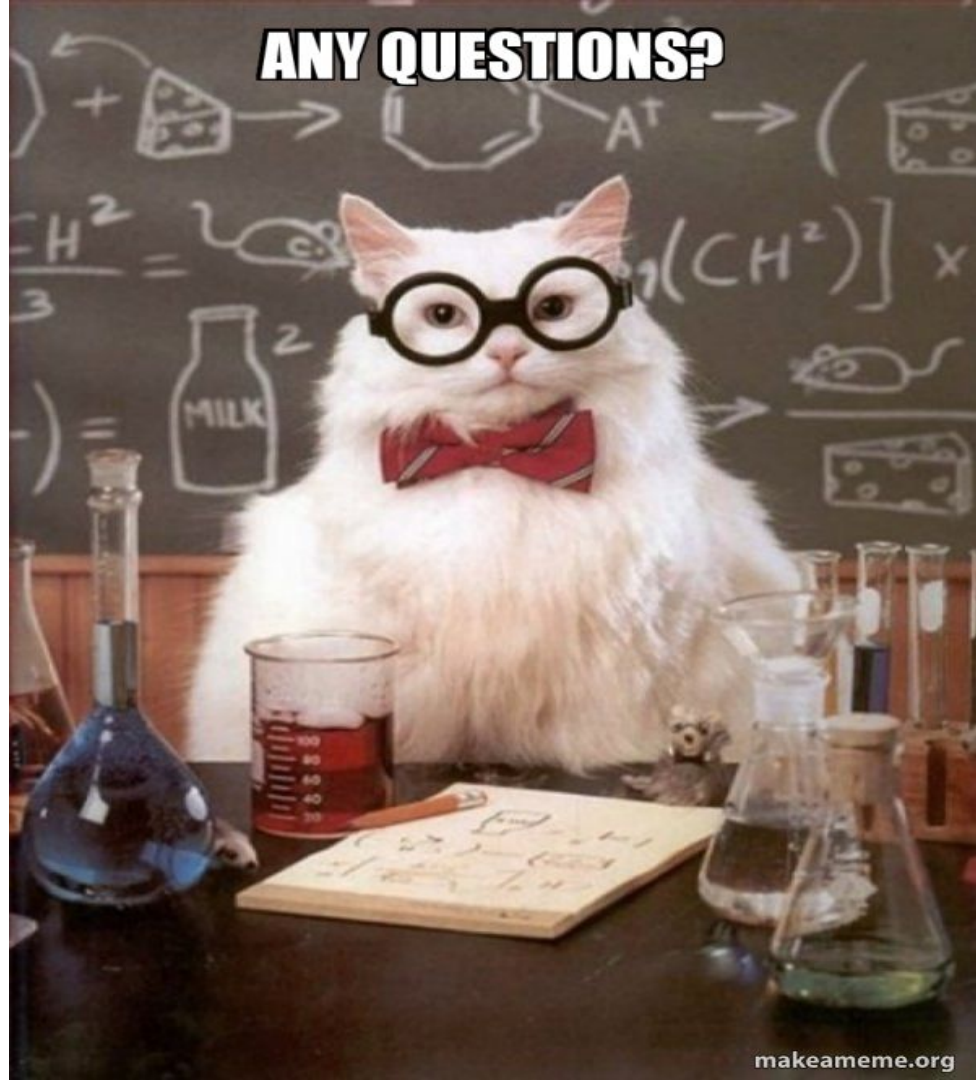
- Limited test set.
- Flagged correctly if any of the reasons/violations category from the ground truth was picked up, or if nothing it detected if the image was from the control, entirely non-harmful images group (does not account for the fact that multiple violations may exist/be detected, or highlights recall/precision tradeoff).
- No “class”/violation-level metrics.

Solution	% images correctly flagged or not – “standard” harmful content	% images correctly flagged or not – customer/domain-specific content
AWS Rekognition	~79%	Not supported
SightEngine	~74%	Not supported
Our custom pipeline	~ 89%	~ 62%

Lessons learned/future plans?

- Confirmed the feasibility of L(L)M utilization for multimodal pipelines with the current state of the tech – at least for our multimodal use case.
- Modularity, especially allowing easy prompt or model replacement is a key property of designed (pipeline/solution). Field is evolving very rapidly to say the least – since the inception of this PoC/project models like Mixtral, Qwen 1.5, Claude-3 or VideoMamba became available.
- Thanks o foundation models adaptability, complex solutions are possible even with very limited/no training data.
- But ultimately, we will need it for evaluation purposes – in our initial work it was lacking e.g. just binary checking if the content to be flagged was flagged for one of the reasons in the ground truth – – aspect to be improved.
- More thorough budget/cost evaluation will be coming. Inference optimization wise, much can be achieved utilizing dedicated tools (vLLM, TensorRT, etc.) – though we're yet to establish what models will be best to use for the best accuracy/costs tradeoff.

ANY QUESTIONS?



Adventure time! A journey for flagging dangerous multimodal content using LLMs

Michał Mikołajczak

Interested/need help in similar AI/Data/MLOps topics?

Or maybe you are a student interested in such areas?

Visit <https://www.datarabbit.ai/>
or contact us at contact@datarabbit.ai

