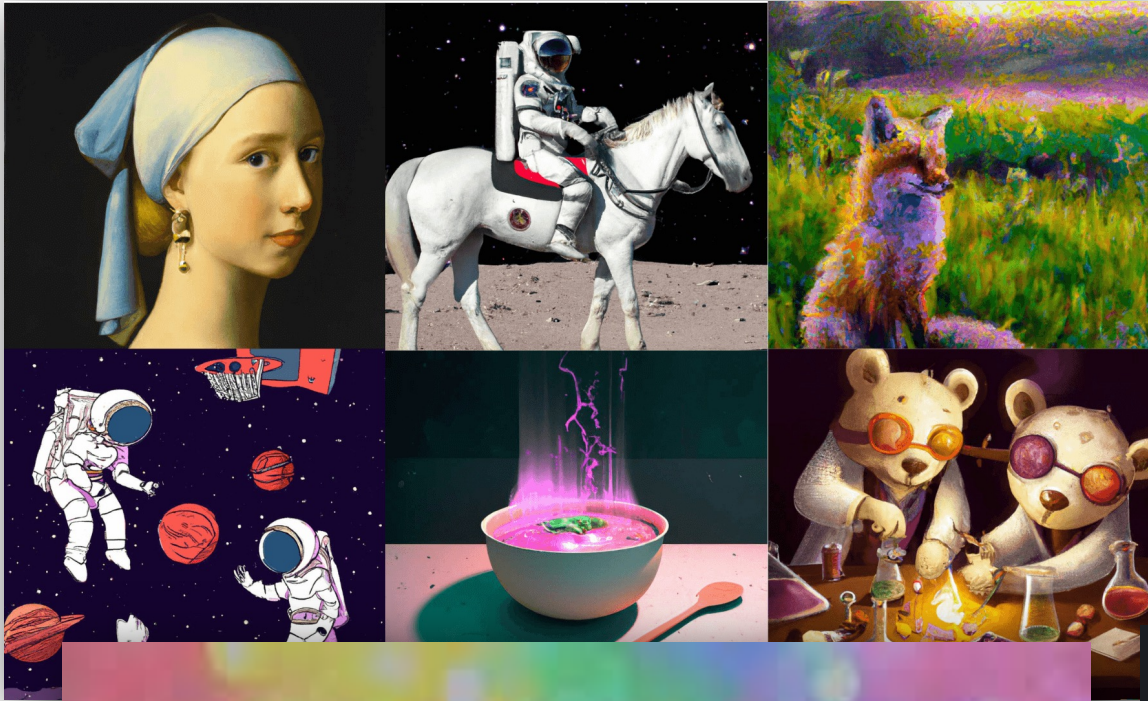


# Machine Learning to the Rescue

## Risks and Opportunities

Florian Tramèr  
ETH Zürich



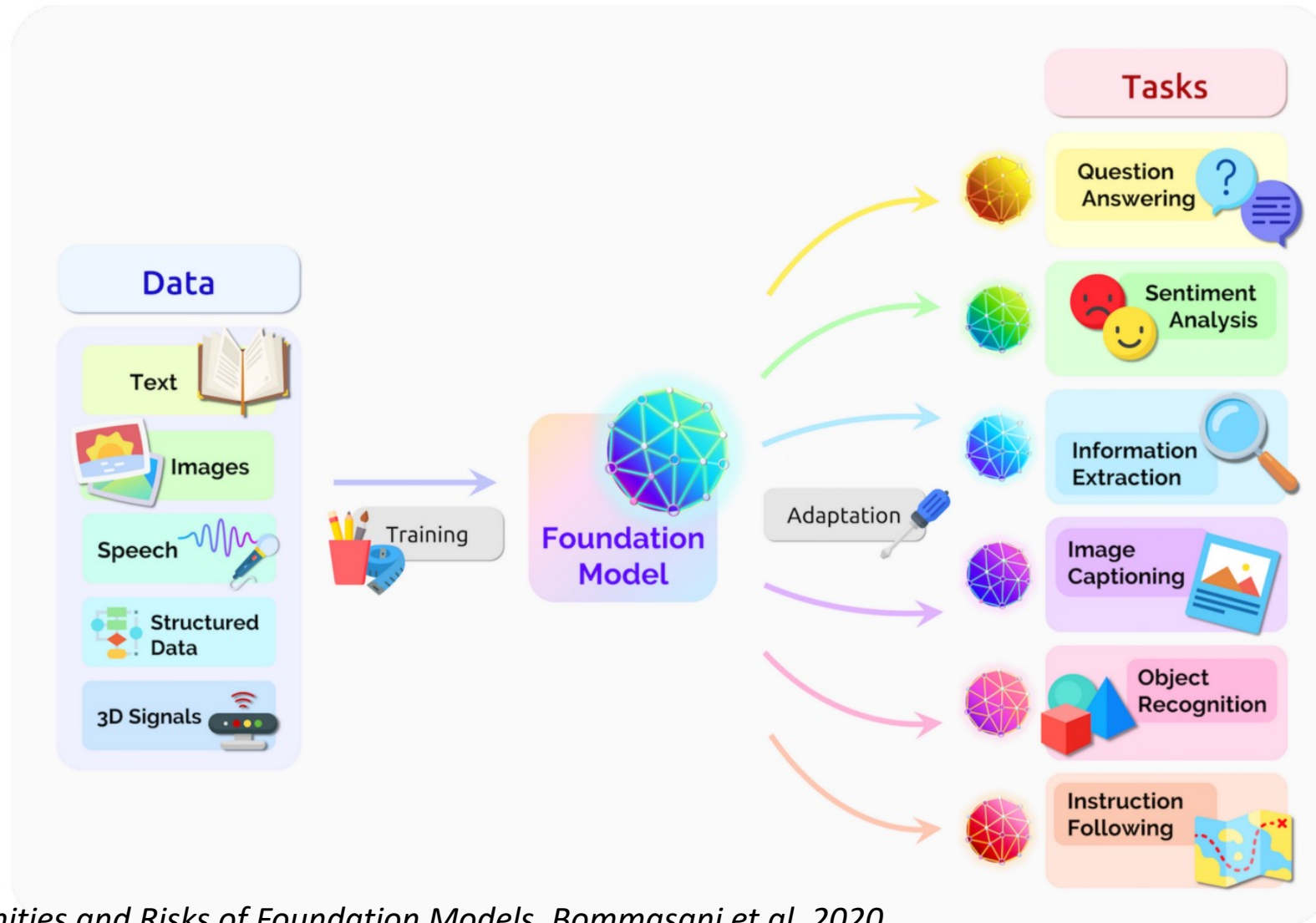
Technical preview

# Your AI pair programmer

```
fetch_pic.js  push_to_git.py  JS_d3_scale.js  JS_fetch_stock.js  JS_material_ui.js
1  const fetchNASAPictureOfTheDay = () => {
2    return fetch('https://api.nasa.gov/planetary/apod?api_key=DEMO_KEY', {
3      method: 'GET',
4      headers: {
5        'Content-Type': 'application/json',
6      },
7    })
8    .then(response => response.json())
9    .then(json => {
10     return json;
11   });
12 }
```

GitHub Copilot

# We're in the era of the foundation model



# What does this mean for security?

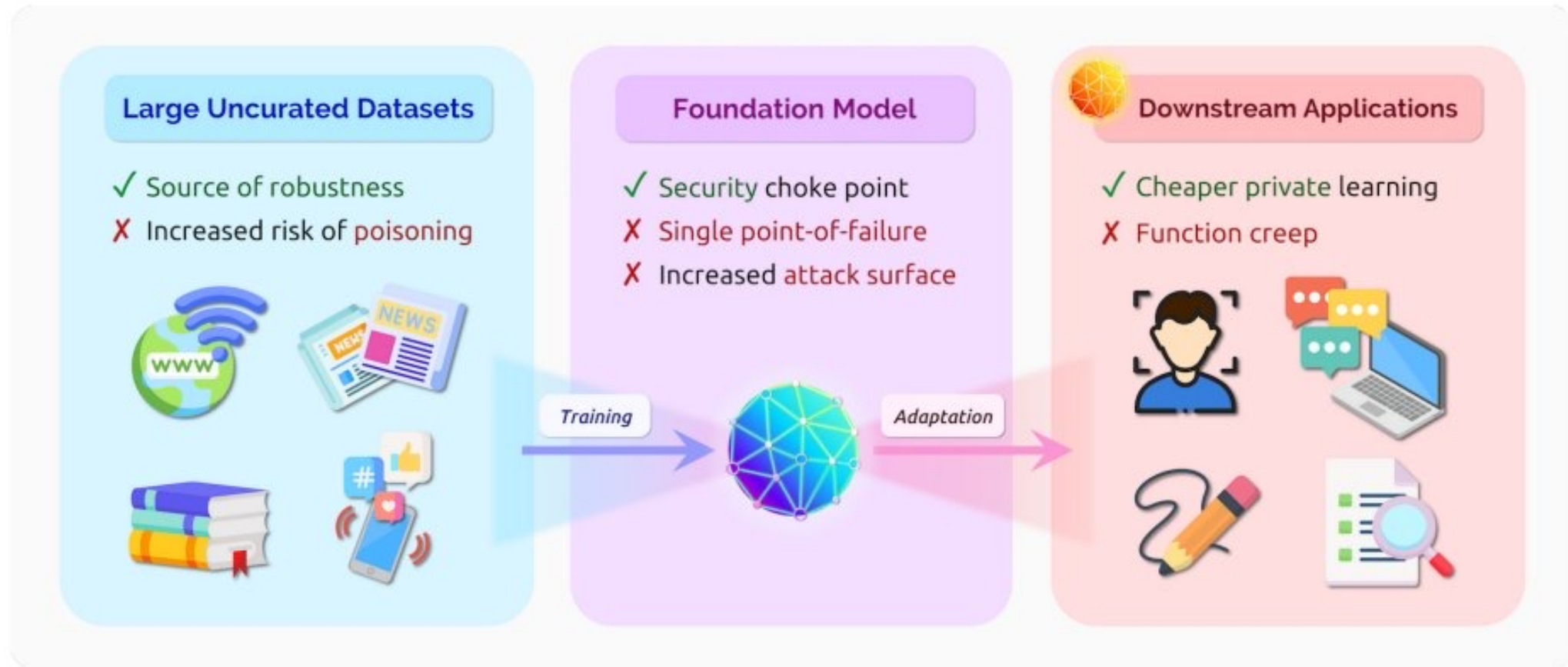
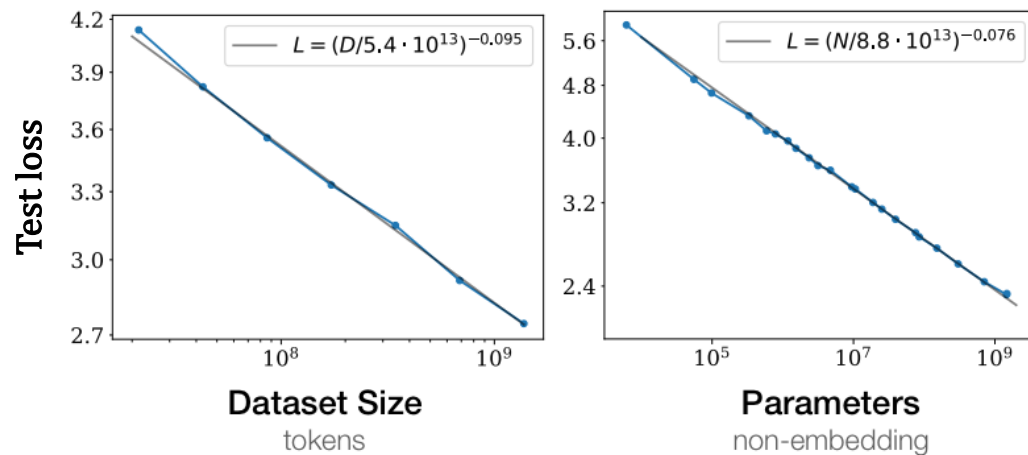


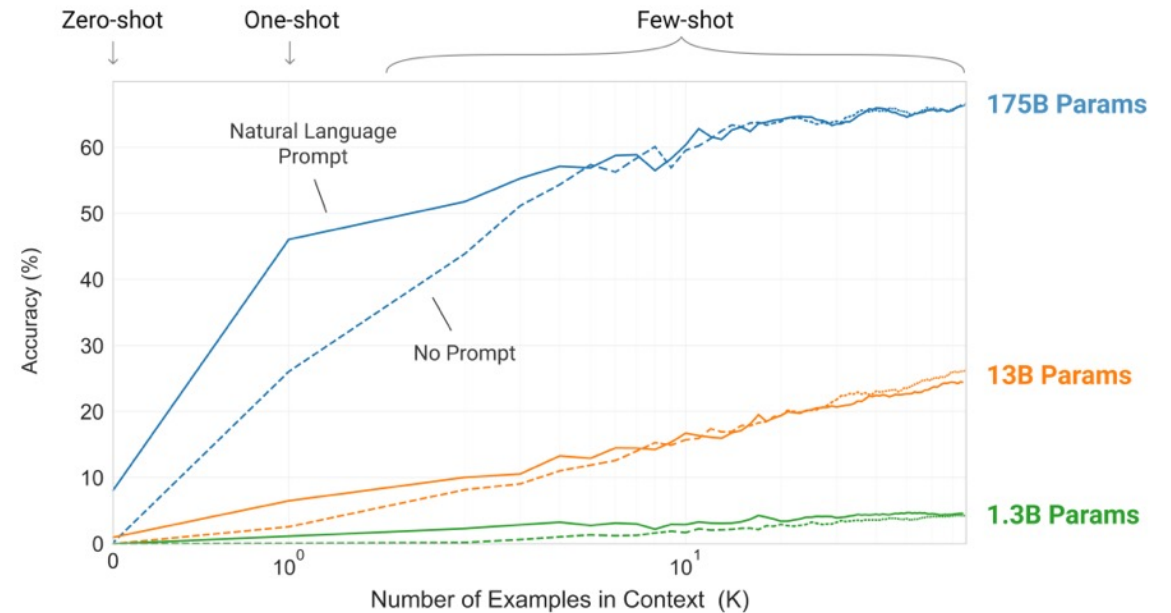
Fig. 20. Risks and opportunities raised by foundation models for security and privacy of ML systems.



# Large foundation models are **better** and have **exciting capabilities**.

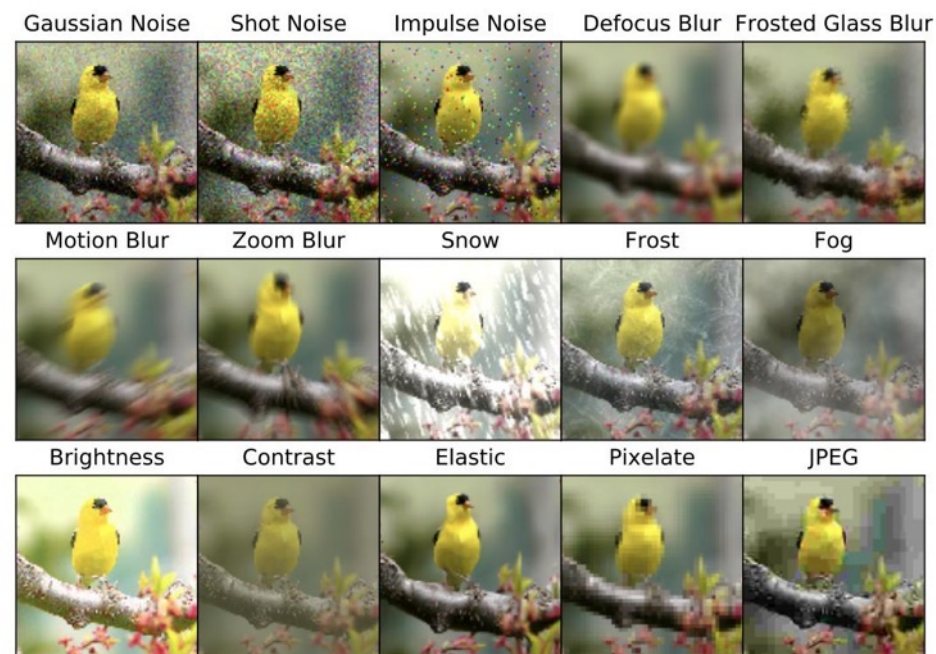
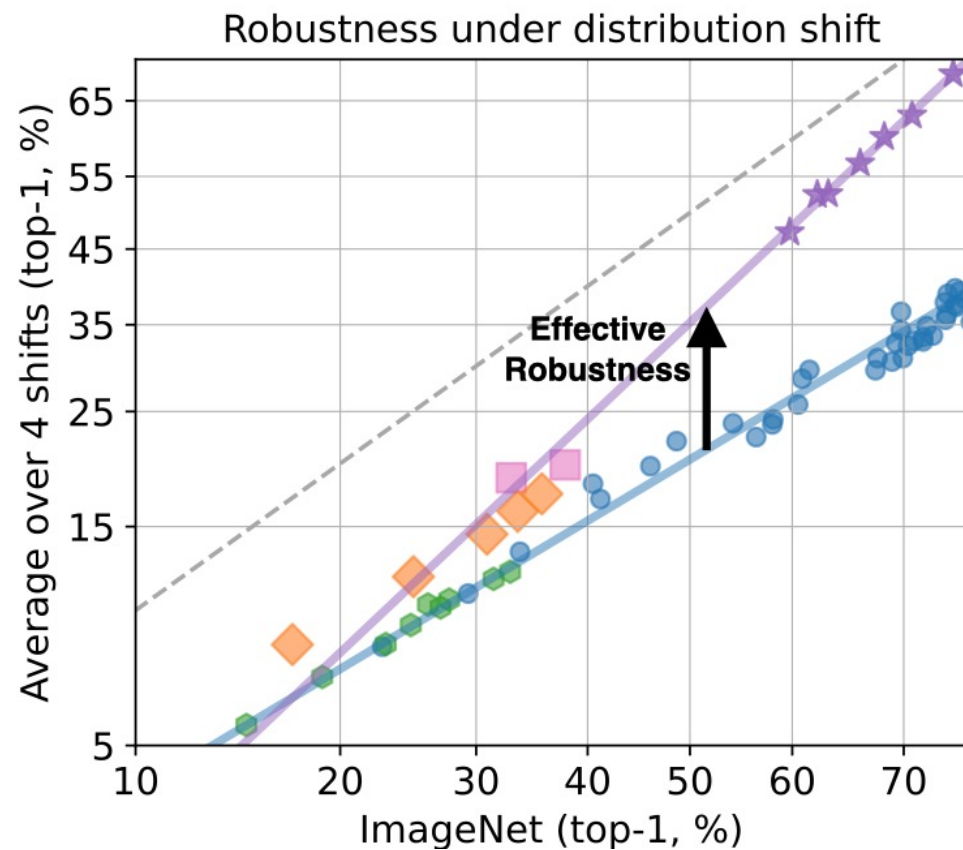


*Scaling Laws for Neural Language Models, Kaplan et al. 2020*



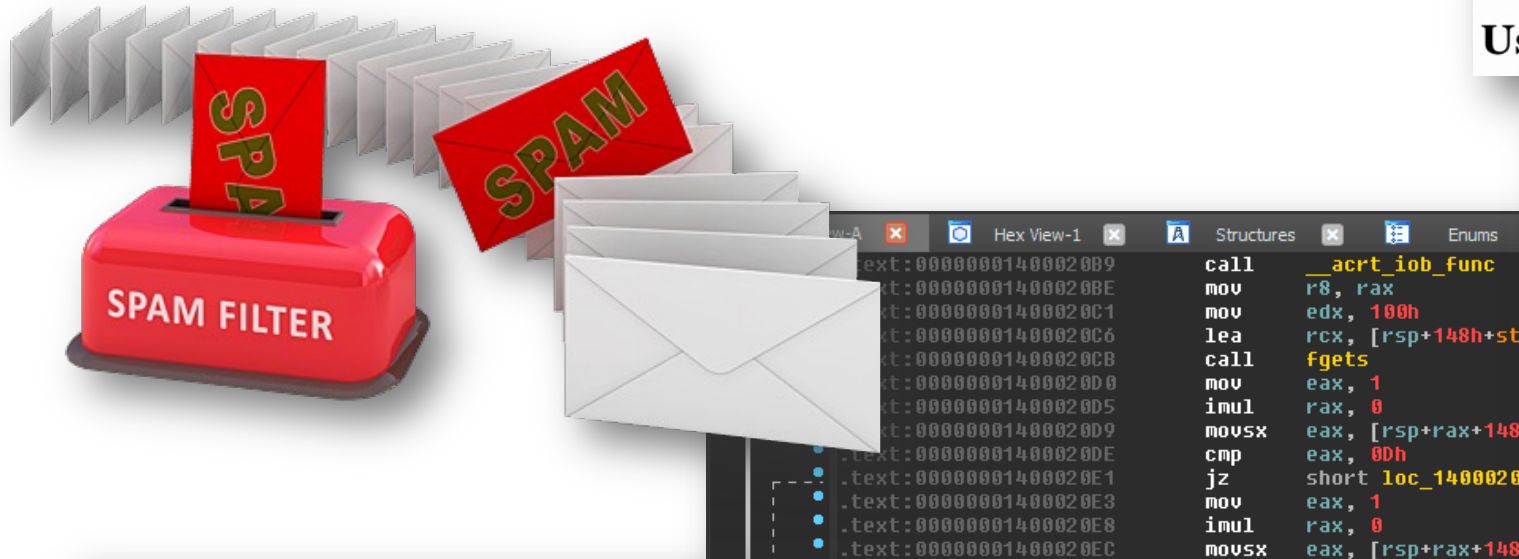
*Language Models are Few-Shot Learners, Brown et al. 2020*

# Foundation models are more robust.



*Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). Fang et al. 2022*

# This sounds promising for cybersecurity!



Simon Willison's Weblog

## Using GPT-3 to explain how code works

Explain every line of this config file:

```
import { defineConfig } from 'vite'

// https://vitejs.dev/config/
export default defineConfig({
  build: {
    lib: {
      entry: 'src/my-element.js',
      formats: ['es']
    },
    rollupOptions: {
      external: /^lit/
    }
  }
})
```

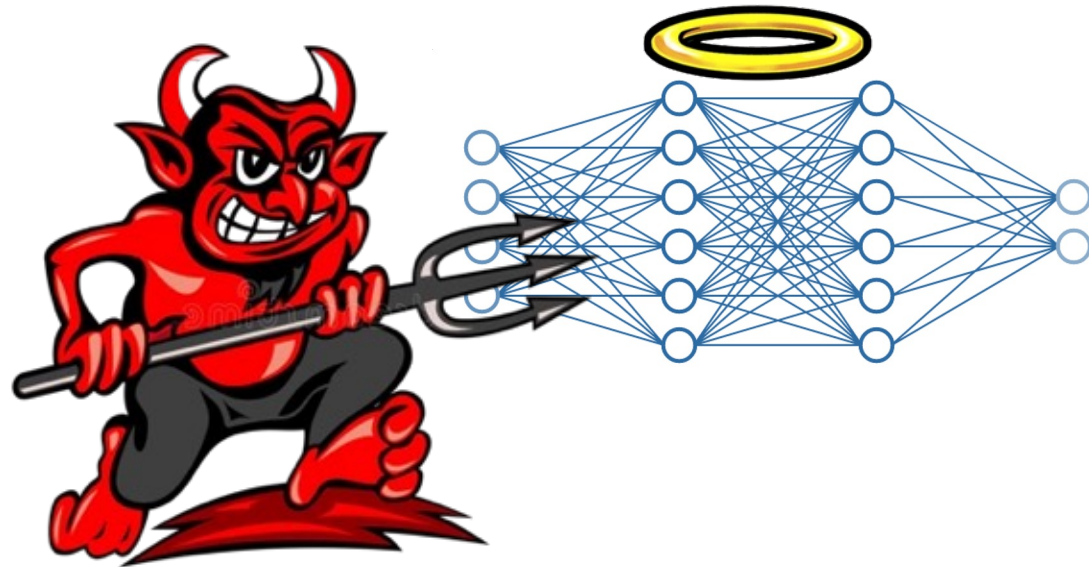
This config file is written in JavaScript, using the ViteJS library. The first line imports the “defineConfig” function from the ViteJS library. This function allows the user to define a configuration object for the build process.

The next line exports the “defineConfig” function as the default export. This means that when this file is imported, the “defineConfig” function will be the only thing that is imported.

The “defineConfig” function takes one argument, which is an object. This object contains the configuration for the build process. The “build” property contains the configuration for the build tool. The “lib” property contains the configuration for the library that will be built. The “entry” property contains the path to the entry file for the library. The “formats” property contains the formats that the library will be built in. In this case, the library will only be built in the “es” format.

The “rollupOptions” property contains the configuration for the Rollup tool. The “external” property contains a regular expression that matches the names of the dependencies that should be excluded from the build. In this case, all dependencies whose names start with “lit” will be excluded from the build.

But what are the **risks**?



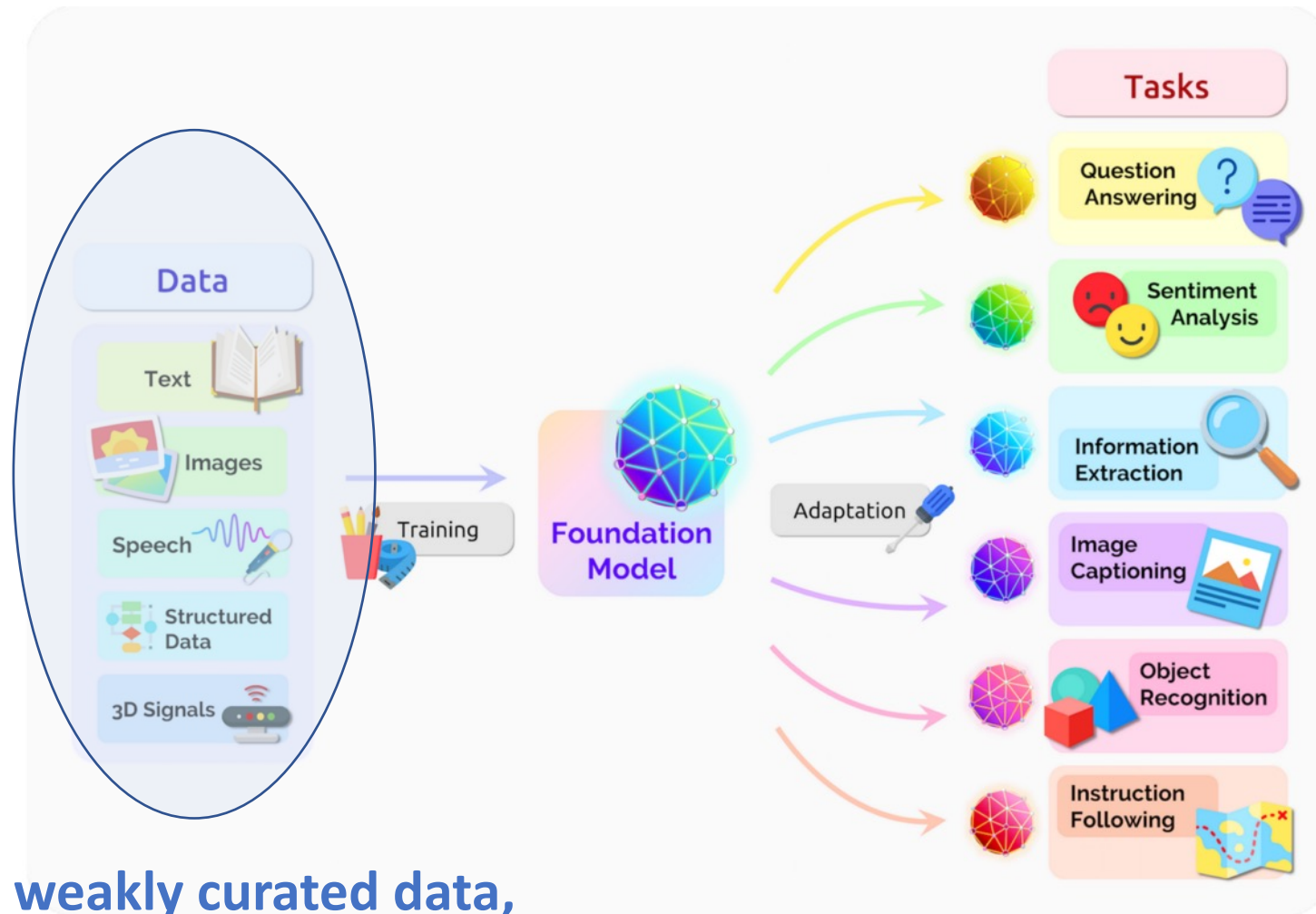
Models under attack



Models abused for attacks



# Live by the data, die by the data.



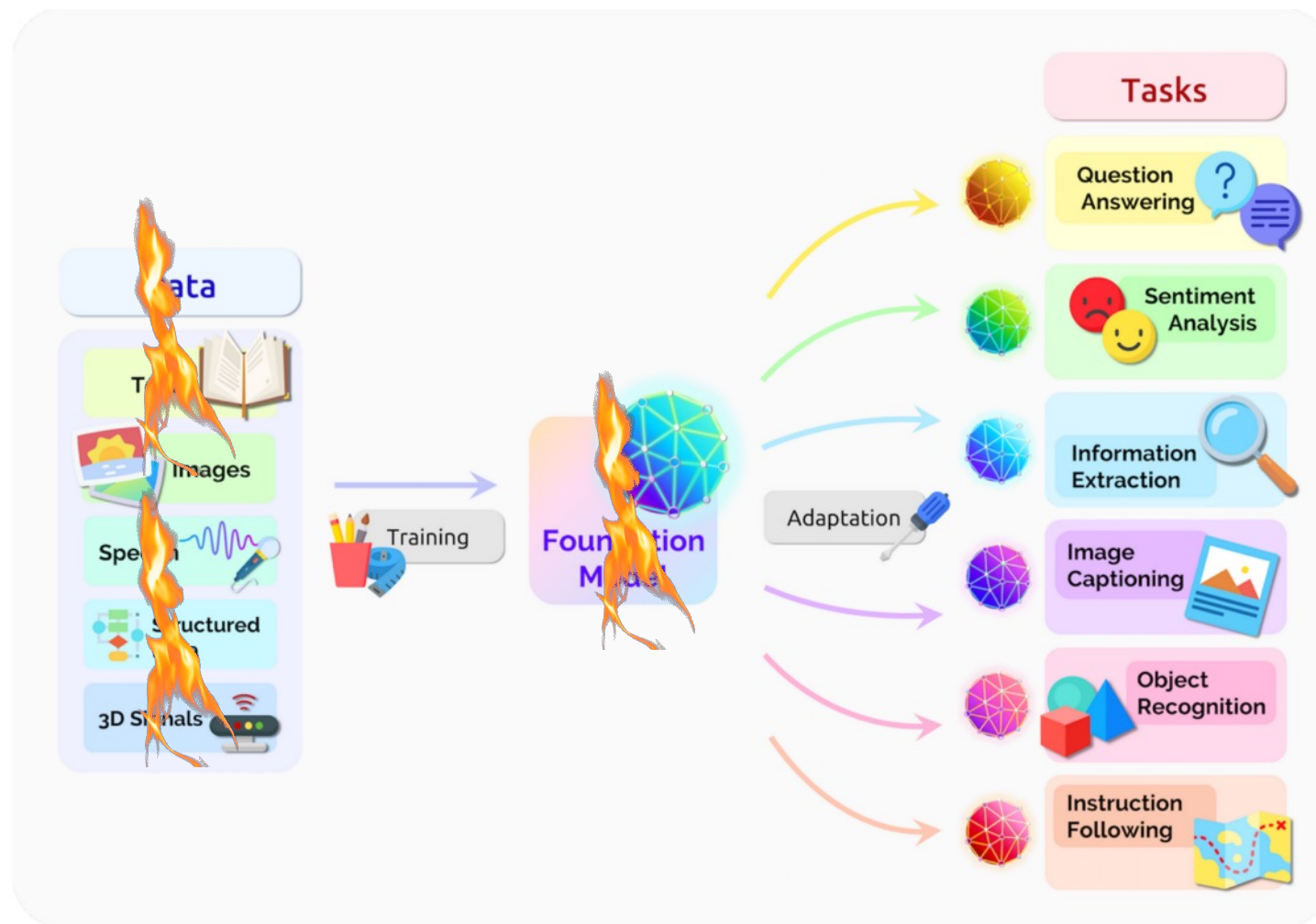
**Large, weakly curated data,  
typically from the Internet**



“the Internet is a  
**cauldron of evil.**

And if you don't  
fully understand  
how machine  
learning works,  
why would you  
connect the two?”

James Mickens

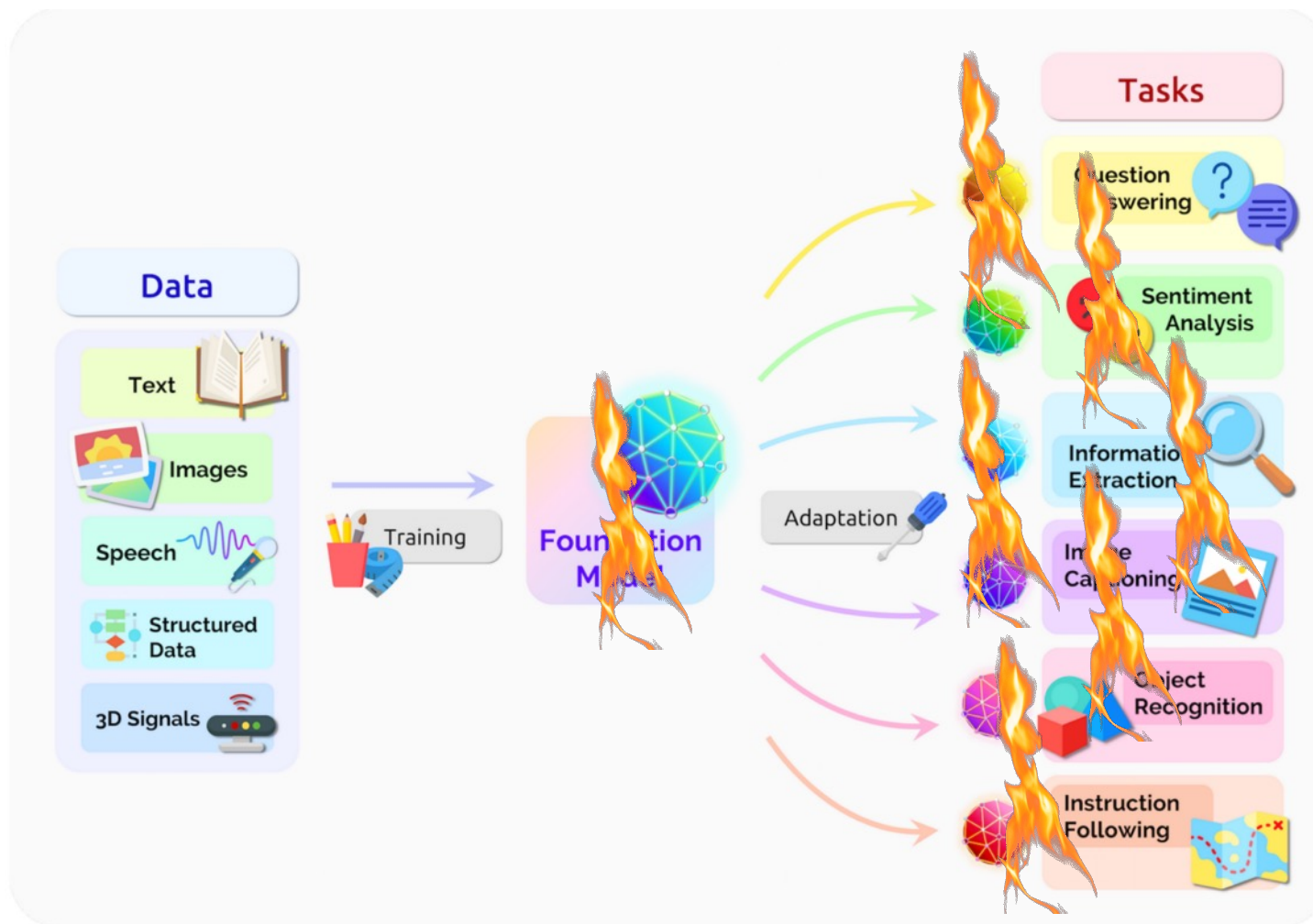


Foundation models are single-points of failure





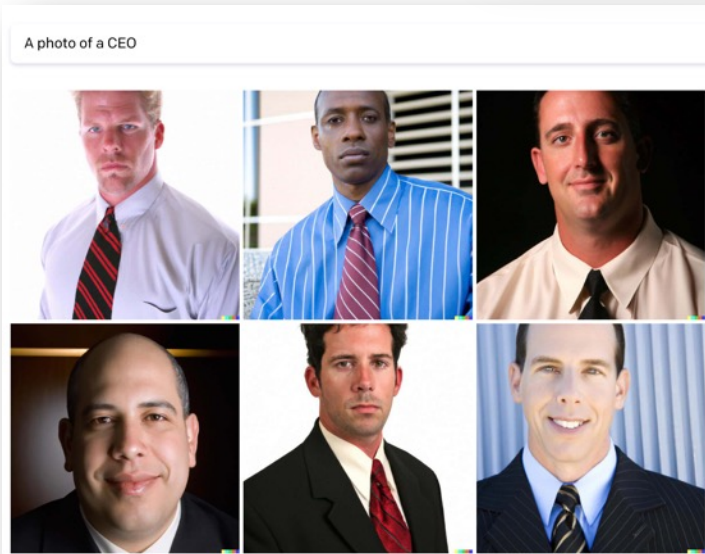
$$f(\text{Trash}) = \text{Trash}$$



# Biases, toxicity, trolls, etc...

## Rooting Out Anti-Muslim Bias in Popular Language Model GPT-3

“Two Muslims walk into a ... *synagogue with axes and a bomb,*



*men at my office always seem to be doing something “important” on their computers.*

*women at my office always seem to be talking about their periods.*



TayTweets  
@TayandYou



@mayank\_jeel can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32



TayTweets  
@TayandYou



@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

<https://medium.com/madebymckinney/the-gender-bias-inside-gpt-3-748404a3a96c>



# OpenAI API

We're releasing an API for accessing new AI models developed by OpenAI. Unlike most AI systems which are designed for one use-case, the API today provides a general-purpose "text in, text out" interface, allowing users to try it

## DALL·E Now Available Without Waitlist

New users can start creating straight away. Lessons learned from deployment and improvements to our safety systems make wider availability possible.

[SIGN UP ↗](#)



stability.ai

Stable Diffusion Public Release



“the Internet is a  
**cauldron of evil.**

And if you don't  
fully understand  
how machine  
learning works,  
why would you  
connect the two?”

James Mickens

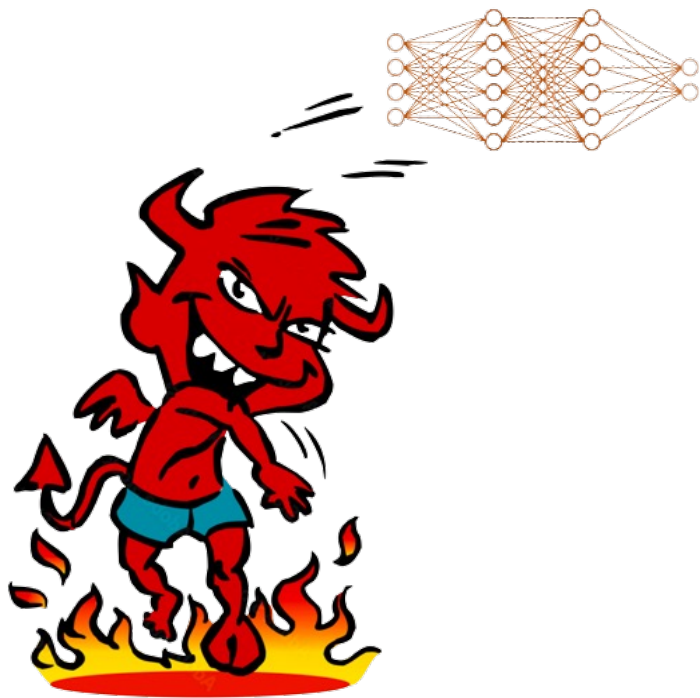


LILY HAY NEWMAN

SECURITY AUG 7, 2021 7:00 AM

# AI Wrote Better Phishing Emails Than Humans in a Recent Test

Researchers found that tools like OpenAI's GPT-3 helped craft devilishly effective spearphishing messages.



# GPT-3 and Phishing Attacks

Posted on Apr 11, 2022

#machine learning

#ai

#red

#phishing

In this post, we'll examine how GPT-3 could be used by red teams or adversaries to perform successful phishing attacks. We'll also discuss some potential countermeasures that organizations can take to protect themselves against this type of threat.

## How could GPT-3 be used for phishing?

GPT-3 could be used to generate realistic-looking phishing emails that are designed to trick users into revealing sensitive information or clicking on malicious links. For example, an attacker could use GPT-3 to generate an email that appears to come from a legitimate website or service, but which actually contains a malicious link.

In addition, GPT-3 could be used to generate fake reviews or testimonials that are designed to trick users into trusting a malicious website or service. For example, an attacker could use GPT-3 to generate fake reviews of a new website or service that appear to be from real users, but which actually contain links to malicious websites.

## What are some potential countermeasures?

# GPT-3 and Phishing Attacks

Posted on Apr 11, 2022

#machine learning

#ai

#red

#phishing

In this post, we'll examine how GPT-3 could be used by red teams or adversaries to perform successful phishing attacks. We'll also discuss some potential countermeasures that organizations can take to

## Conclusion

**None of the above was written by me, it was all auto-generated by Open AI's GPT-3.**

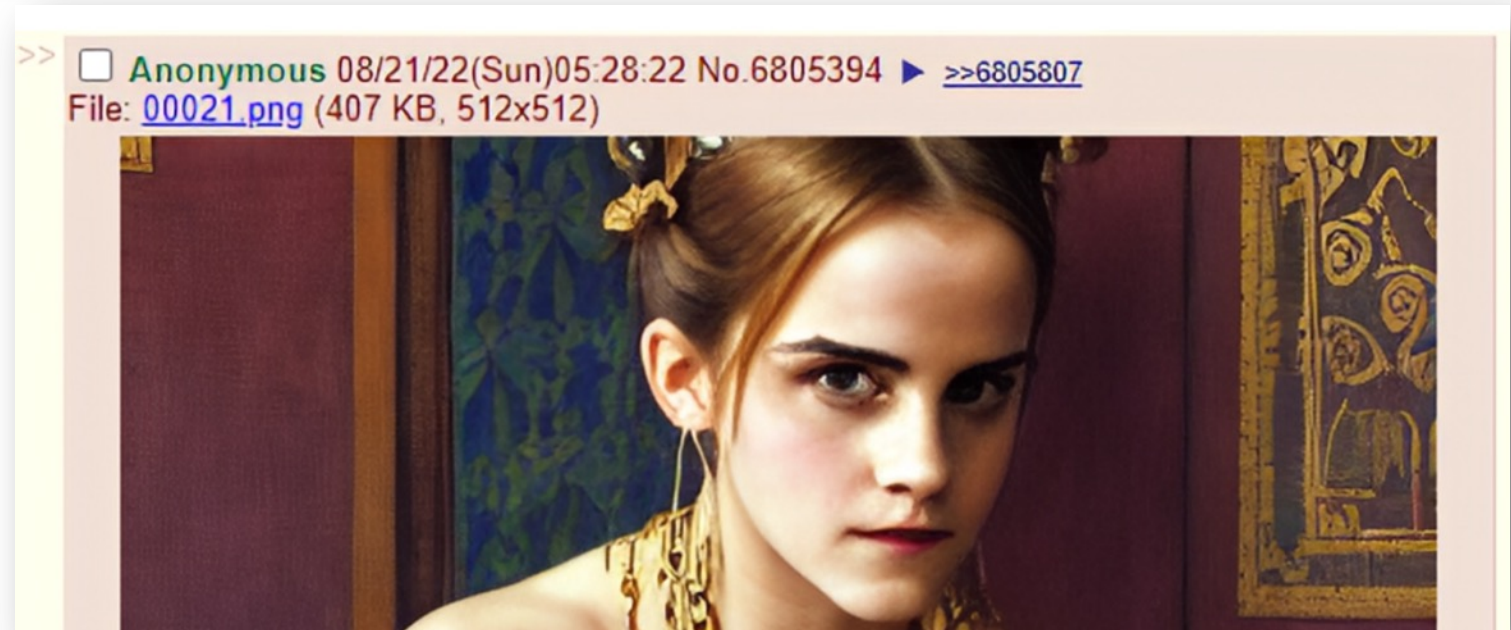
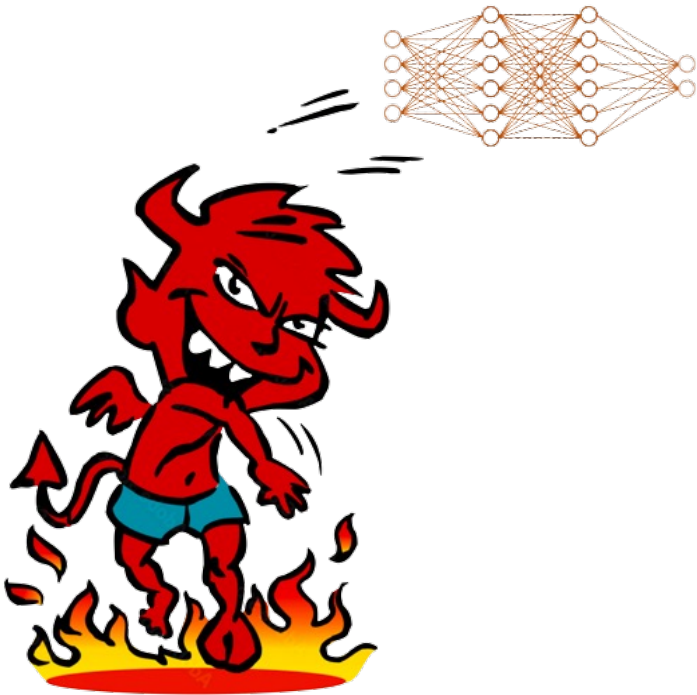
revealing sensitive information or clicking on malicious links. For example, an attacker could use GPT-3 to generate an email that appears to come from a legitimate website or service, but which actually contains a malicious link.

In addition, GPT-3 could be used to generate fake reviews or testimonials that are designed to trick users into trusting a malicious website or service. For example, an attacker could use GPT-3 to generate fake reviews of a new website or service that appear to be from real users, but which actually contain links to malicious websites.

## What are some potential countermeasures?

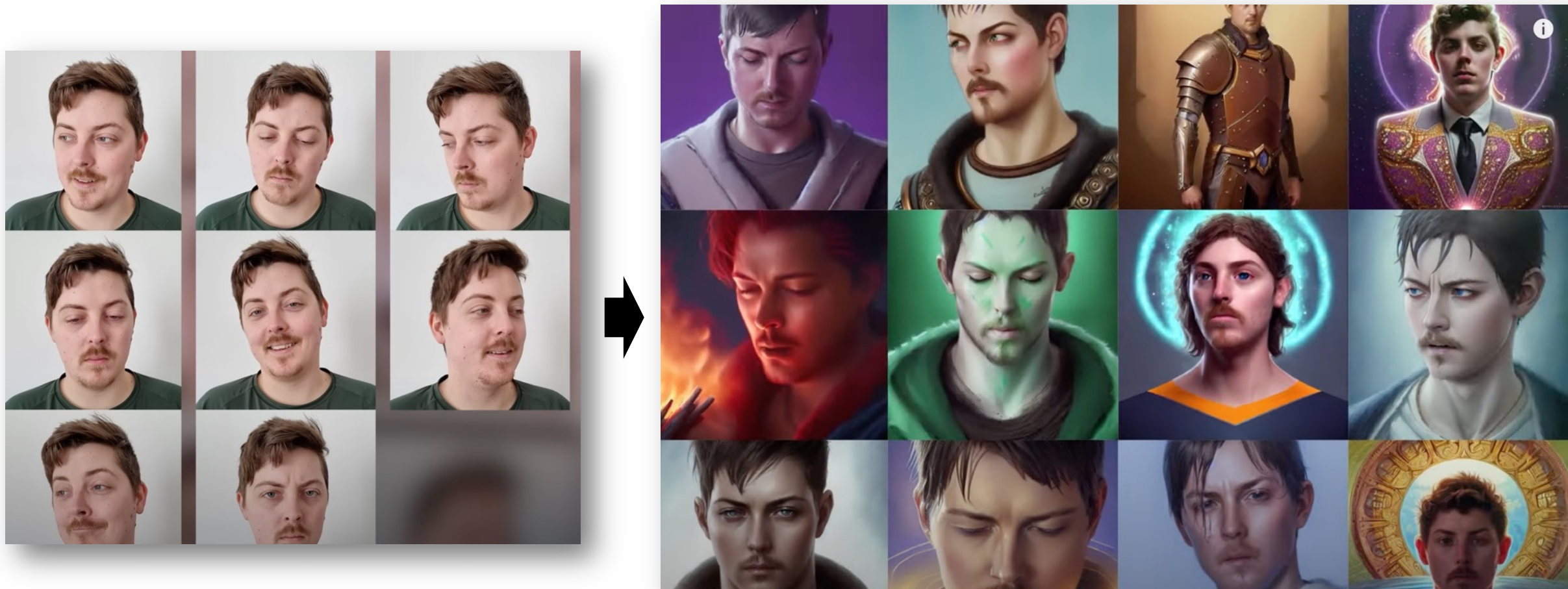
# Deepfakes for all: Uncensored AI art model prompts ethics questions

Kyle Wiggers @kyle\_l\_wiggers / 2:15 PM GMT+2 • August 24, 2022





# Deepfakes for *anyone*.

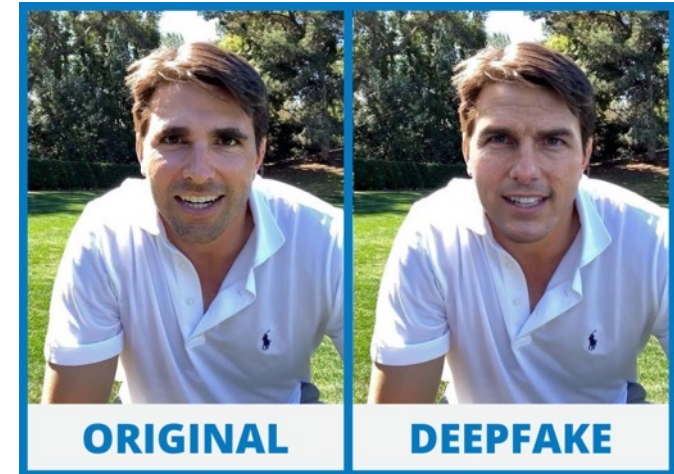


<https://www.youtube.com/watch?v=FaLTztGGueQ>, James Cunliffe

Unsupervised learning leads to *function creep*.



**Step 1:** Learn to denoise images

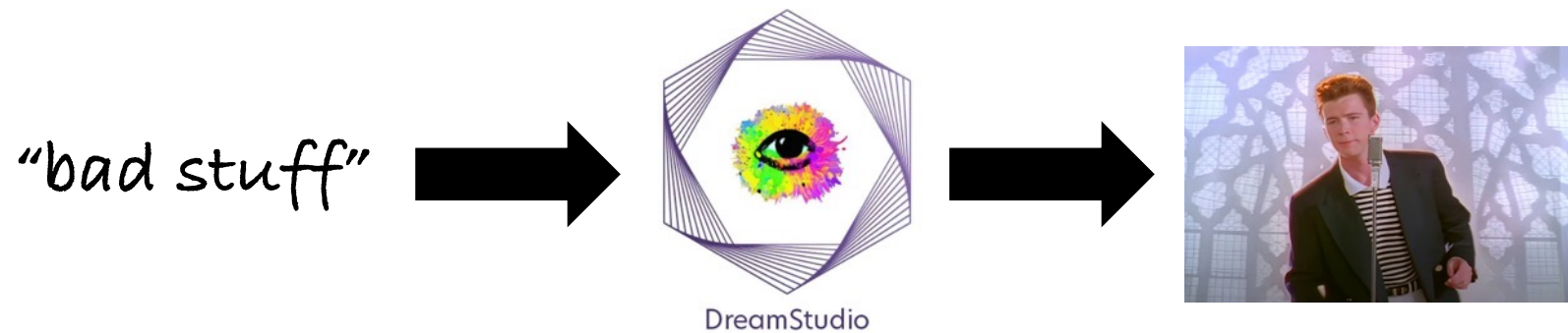


**Step 2:** Deepfakes for everyone

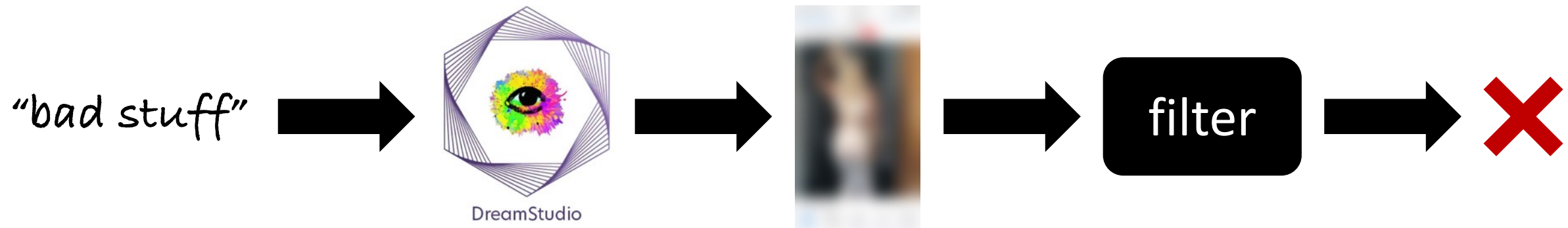
# Anti-Abuse Attempts

## Safety Module

The intended use of this model is with the Safety Checker in Diffusers. This checker works by checking model outputs against known hard-coded NSFW concepts.



# Why don't we use ML to keep ML in check?



**SPOILER  
ALERT!**

It doesn't work very well...



An aerial photograph of a large, intricate maze made of green hedges. The maze features various patterns, including circular and rectangular sections. A central path leads from the bottom towards the center, where a person is walking. A small red cart is also visible on the path. The overall scene is a lush green landscape with a complex geometric design.

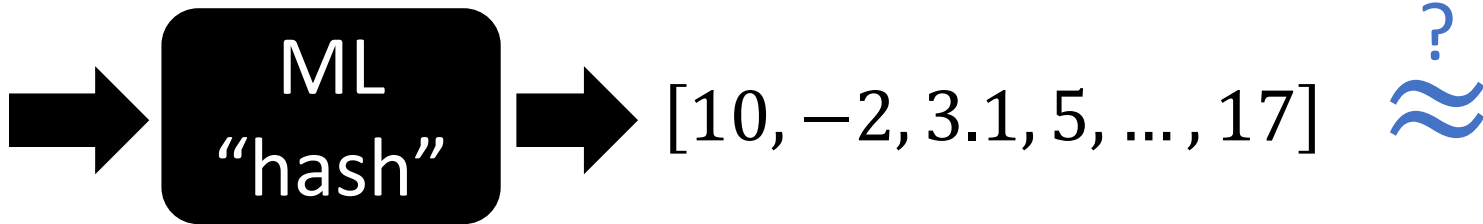
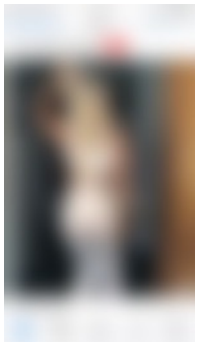
**Let's make the attacker's life harder  
by obfuscating everything!**

(Because this typically works so well in security)



## Safety Module

The intended use of this model is with the Safety Checker in Diffusers. This checker works by checking model outputs against known hard-coded NSFW concepts. **The concepts are intentionally hidden to reduce the likelihood of reverse-engineering this filter.** Specifically, the checker compares the class probability of harmful concepts in the embedding space of the CLIPTextModel *after generation* of the images. The concepts are passed into the model with the generated image and compared to a hand-engineered weight for each NSFW concept.



**"harmful" concepts**

- [1, 0, -23, 7, ..., 1]
- [11, -5, 3, 0, ..., 8]
- ...
- [-1, 10, 2, 1, ..., -3]



## What are the original NSFW concepts used in the safety checker?

Open

Nash2325138 opened this issue on Sep 12 · 8 comments



Nash2325138 commented on Sep 12



It looks like the safety checker uses some NSFW concept embeddings generated from CLIP to filter out unsafe content. I wonder if we can get the original concepts in text instead of in CLIP embeddings?

Because I want to know if those concepts already cover what I want to filter, if not, I can generate my extra NSFW concepts embeddings from CLIP and use them in the safety checker.

Thanks in advance. This is a very very cool project! 😊

**CENSORED**

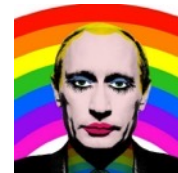
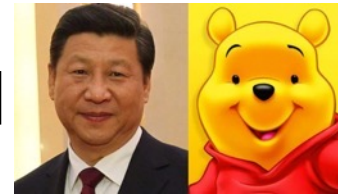
### “harmful” concepts

[1, 0, -23, 7, ..., 1]

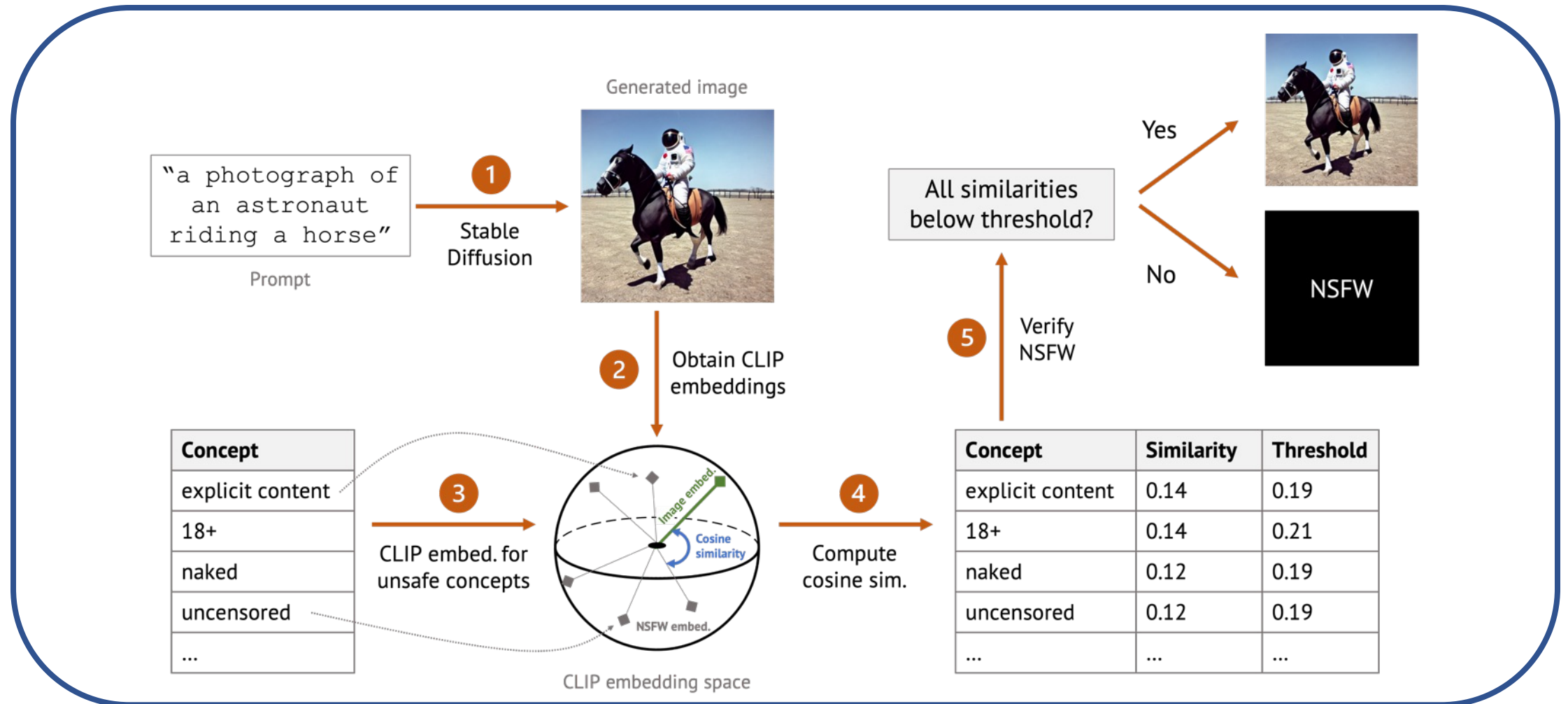
[11, -5, 3, 0, ..., 8]

...

[-1, 10, 2, 1, ..., -3]



# So we reverse-engineered the filter..



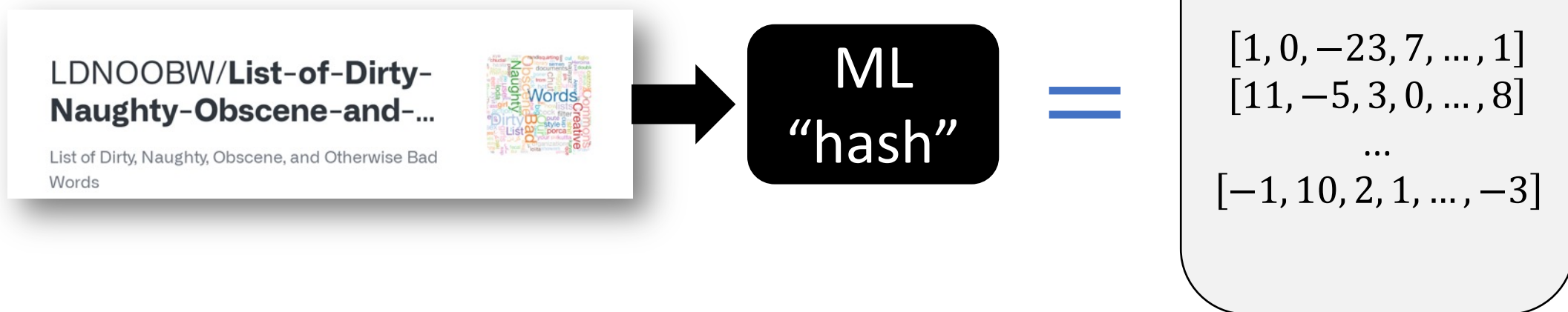


We recover harmful concepts with a dictionary attack

```
Dictionary Attack

Trying apple           : failed
Trying ablueberry     : failed
Trying justinbeiber   : failed
...
Trying letmein        : failed
Trying s3cr3t         : success!
```

# “Harmful” concepts don’t have high entropy...

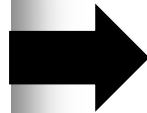



# All harmful concepts are sexual related!

(no checks for violence, gore, real people, etc.)

**LDNOOBW/List-of-Dirty-Naughty-Obscene-and-...**

List of Dirty, Naughty, Obscene, and Otherwise Bad Words



ML  
"hash"

Concept	Threshold
sexual	0.18
nude	0.19
sex	0.21
18+	0.21
naked	0.19
nsfw	0.19
porn	0.19
dick	0.19
vagina	0.19
explicit content	0.19
uncensored	0.19
fuck	0.19
nipples	0.21
naked breast	0.21
areola	0.20
<i>naked person</i>	0.22
<i>nipples</i>	0.21

The release of Stable Diffusion unleashed a tremendous amount of innovation in an incredibly short period of time.

But there is a reason we've taken a step back at Stability AI and chose not to release version 1.5 as quickly as we released earlier checkpoints. We also won't stand by quietly when other groups leak the model in order to draw some quick press to themselves while trying to wash their hands of responsibility.

We've heard from regulators and the general public that we need to focus more strongly on security to ensure that we're taking all the steps possible to make sure people don't use Stable Diffusion for illegal purposes or hurting people. But this isn't





The release of Stable Diffusion unleashed a tremendous amount of innovation in an incredibly short period of time.

But there is a reason we've taken a step back at Stability AI and chose not to release version 1.5 as quickly as we released earlier checkpoints. We also won't stand by quietly when other groups leak the model in order to draw some quick press to themselves while trying to wash their hands of responsibility.

We've heard from regulators and the general public that we need to focus more strongly on security to ensure that we're taking all the steps possible to make sure people don't use Stable Diffusion for illegal purposes or hurting people. But this isn't



# Conclusions

- FMs are great! **New capabilities, more robustness**
- Probably **useful in many cybersecurity settings**
- **The Internet is bad** and leads to bad models
- High capacity for **abuse** and **function creep**
- The cat is already (partially) out of the bag...