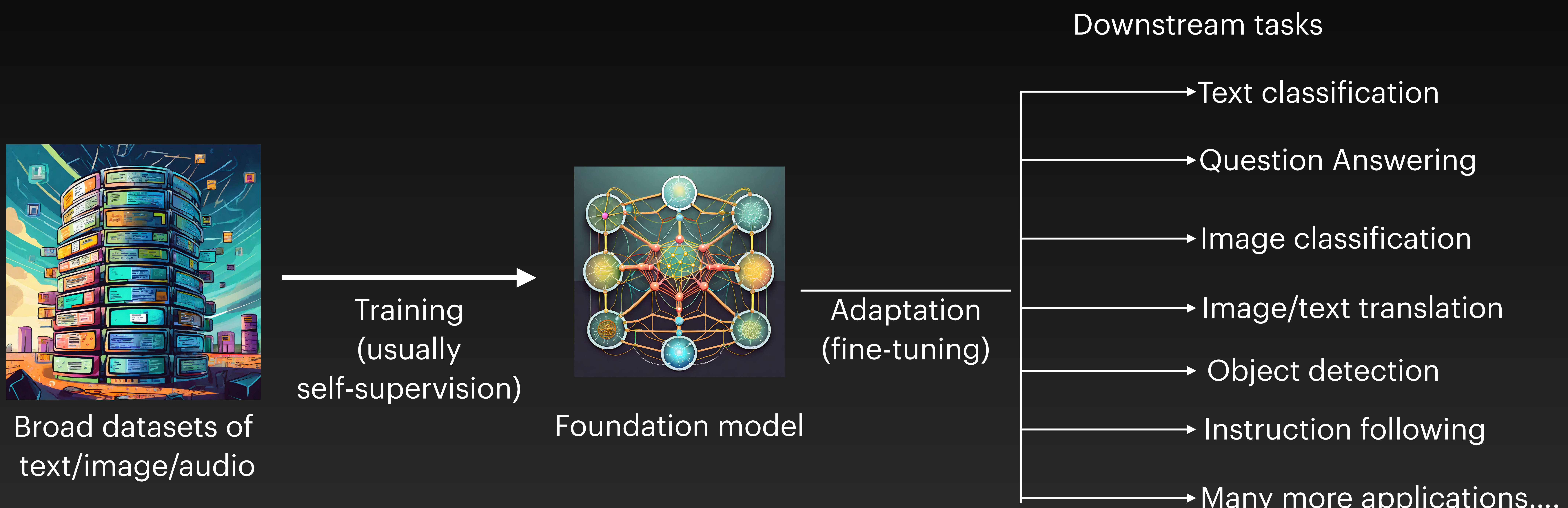
Investigating Foundation Models Through the Lens of Security

Bimal Viswanath

Dept. of Computer Science, Virginia Tech

VIRGINIA TECH™

Foundation models



Downstream tasks

Broad datasets of text/image/audio

Training (usually self-supervision)

Foundation model

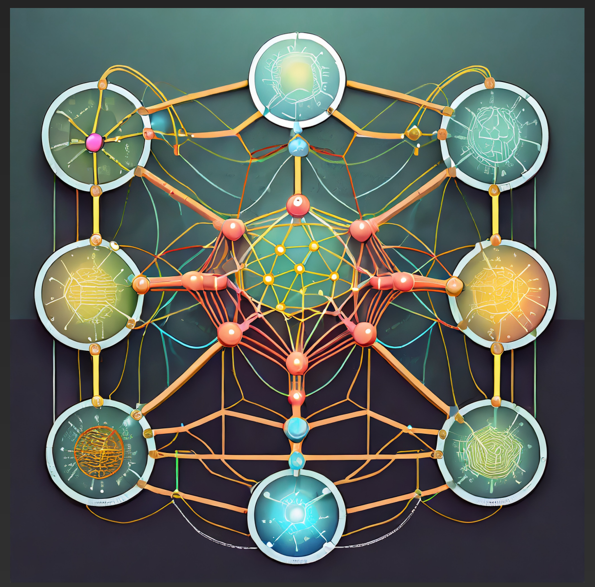
Adaptation (fine-tuning)

- Text classification
- Question Answering
- Image classification
- Image/text translation
- Object detection
- Instruction following
- Many more applications....

Example: Large Language Model (LLM)

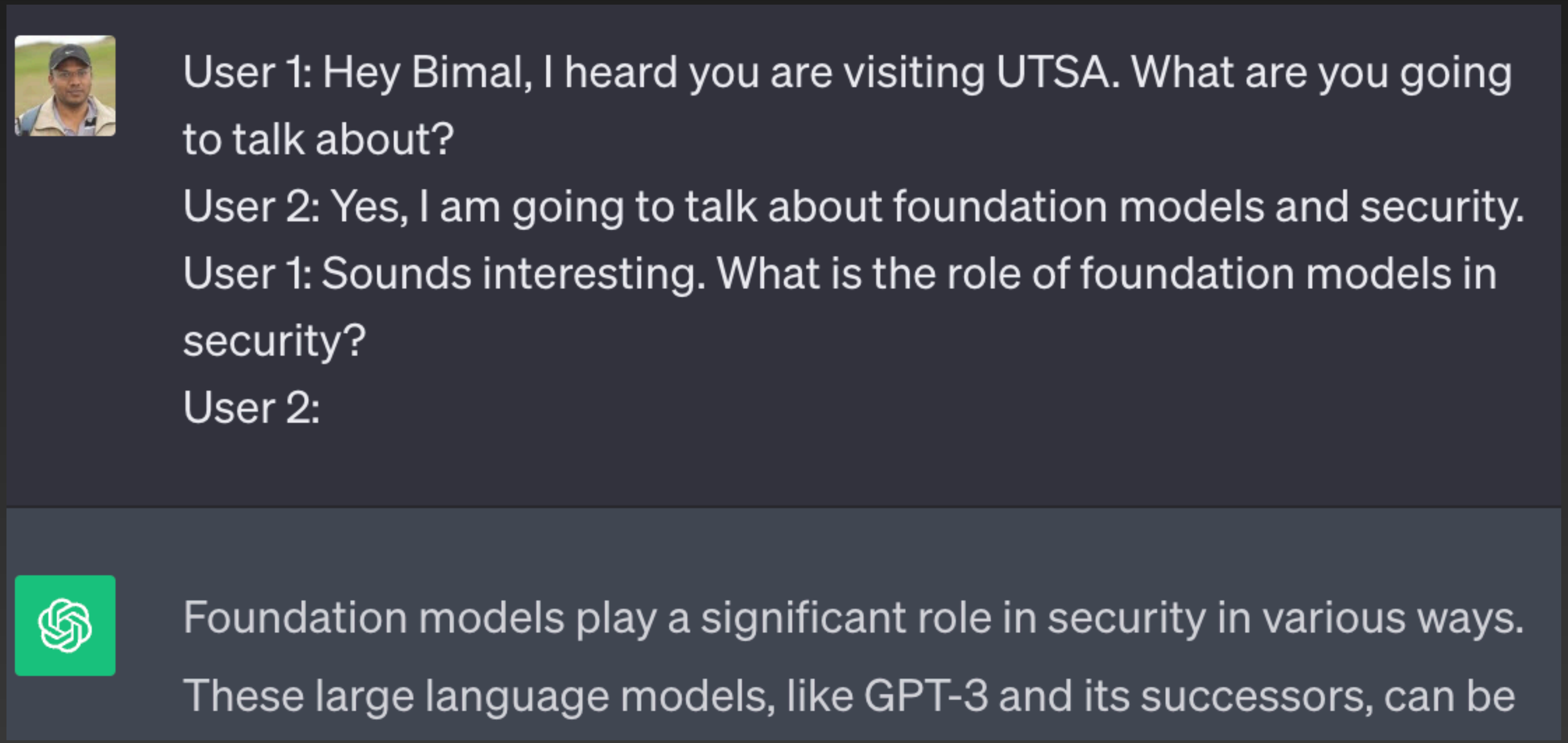
- A model that is trained to predict the next token (e.g., word) in a sequence

$$p(x_0, \ldots, x_n) = \prod_{t=0}^{n} (p(x_{t+1} \mid x_0, \ldots, x_t))$$



LLM

Adaptation
using Prompt Engineering



User 1: Hey Bimal, I heard you are visiting UTSA. What are you going to talk about?
User 2: Yes, I am going to talk about foundation models and security.
User 1: Sounds interesting. What is the role of foundation models in security?
User 2:

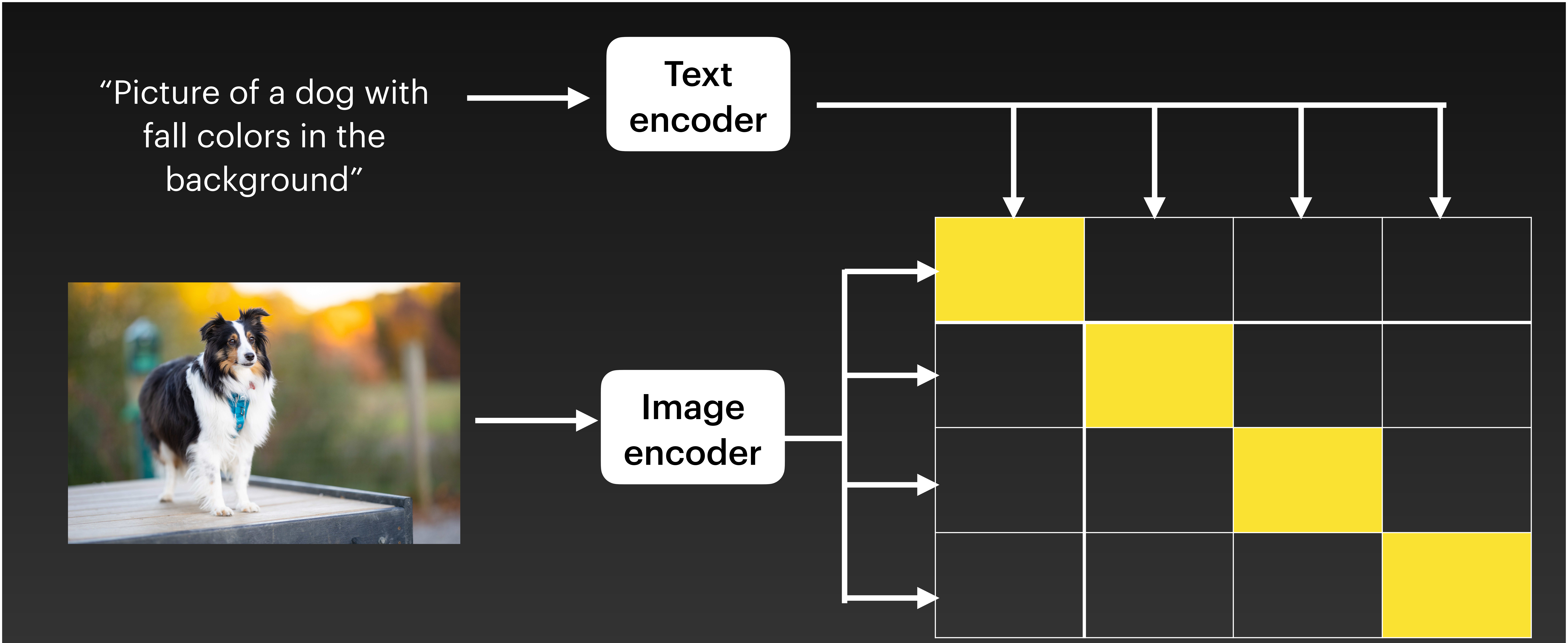
Foundation models play a significant role in security in various ways. These large language models, like GPT-3 and its successors, can be

Example: Image and text encoders

- Foundation models can be effective image and text encoders
- Examples:
  - CLIP
  - BERT
  - ViT

Example: Image and text encoders

- Foundation models can be effective image and text encoders
- Examples:
  - CLIP
  - BERT
  - ViT



“Picture of a dog with fall colors in the background”

Text encoder

Image encoder

CLIP learns a joint text-image embedding

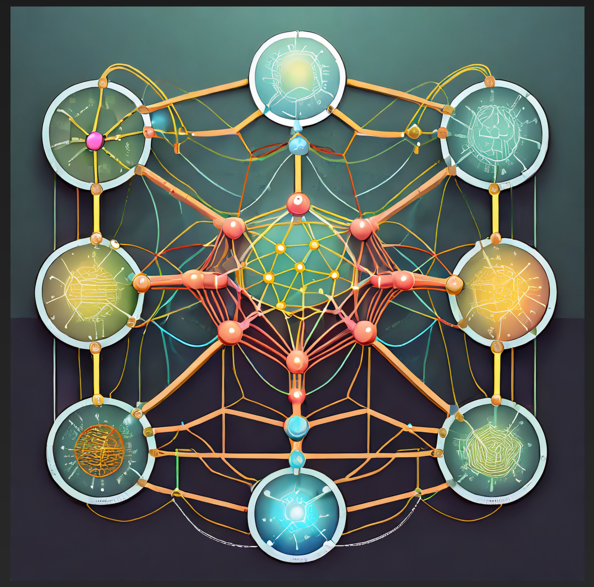
What are the implications of foundation models in security?

In the context of two problems:

1) Deepfake image detection
2) Mitigating toxicity in chatbots

Foundation models: Implications for security

Defender's perspective



Foundation model

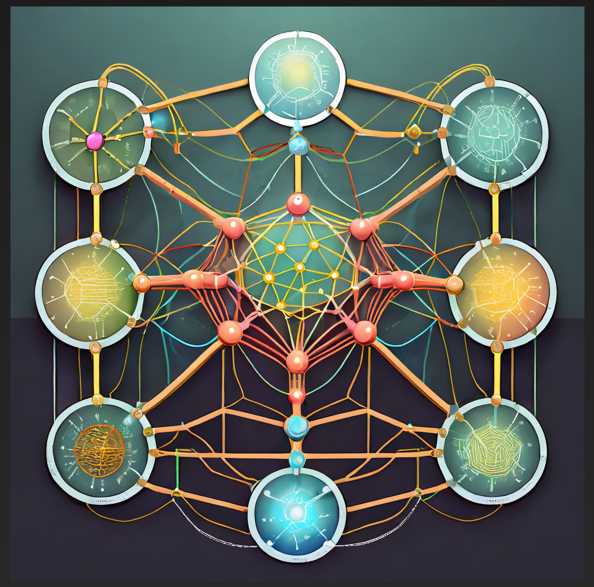
Defender

Foundation models: Implications for security

Defender's perspective

1. Simplify and improve performance of security classifiers

Focus: Deepfake image detectors

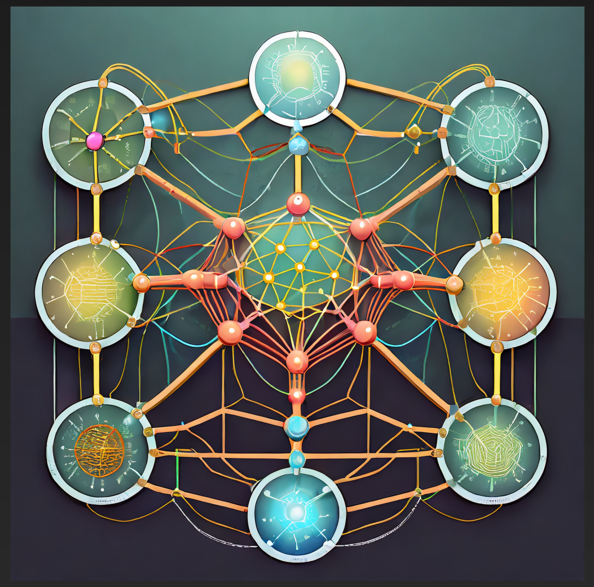


Foundation model

Defender

Foundation models: Implications for security

Defender's perspective



Foundation model

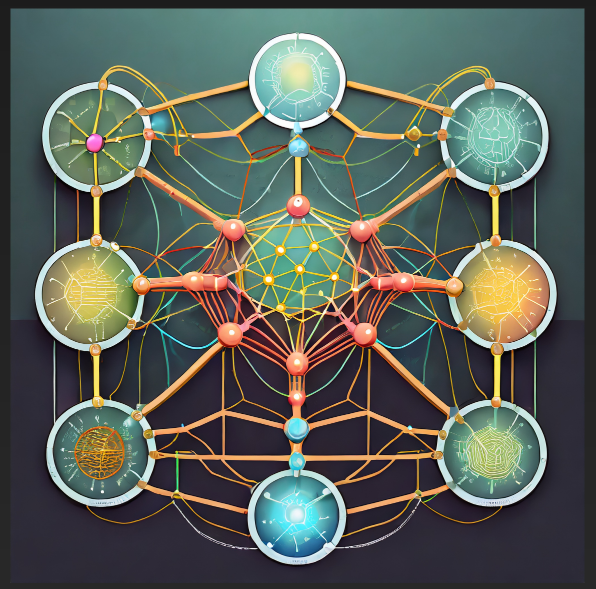
Defender

1. Simplify and improve performance of security classifiers
   Focus: Deepfake image detectors

2. Obviate the need for large labeled dataset for security classification tasks
   Focus: Mitigating toxicity in chatbots

Foundation models: Implications for security

Defender's perspective



Foundation model

Defender

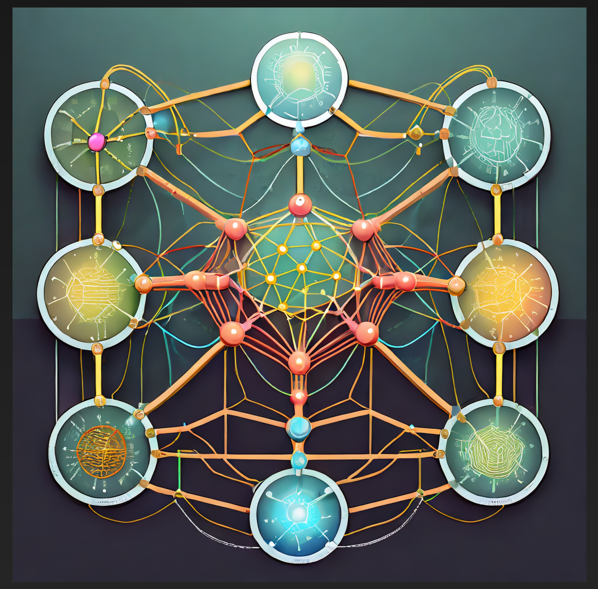
1. Simplify and improve performance of security classifiers
Focus: Deepfake image detectors

2. Obviate the need for large labeled dataset for security classification tasks
Focus: Mitigating toxicity in chatbots

3. Safely customizing foundation models
Focus: Fine-tuning foundation models to build chatbots while mitigating toxicity

Foundation models: Implications for security

Attacker’s perspective



Foundation model

Attacker

4. Create customized variants of foundation models for attacks
Focus: Evading deepfake image detectors

5. Leverage foundation models to craft adversarial samples
Focus: Evading deepfake image detectors

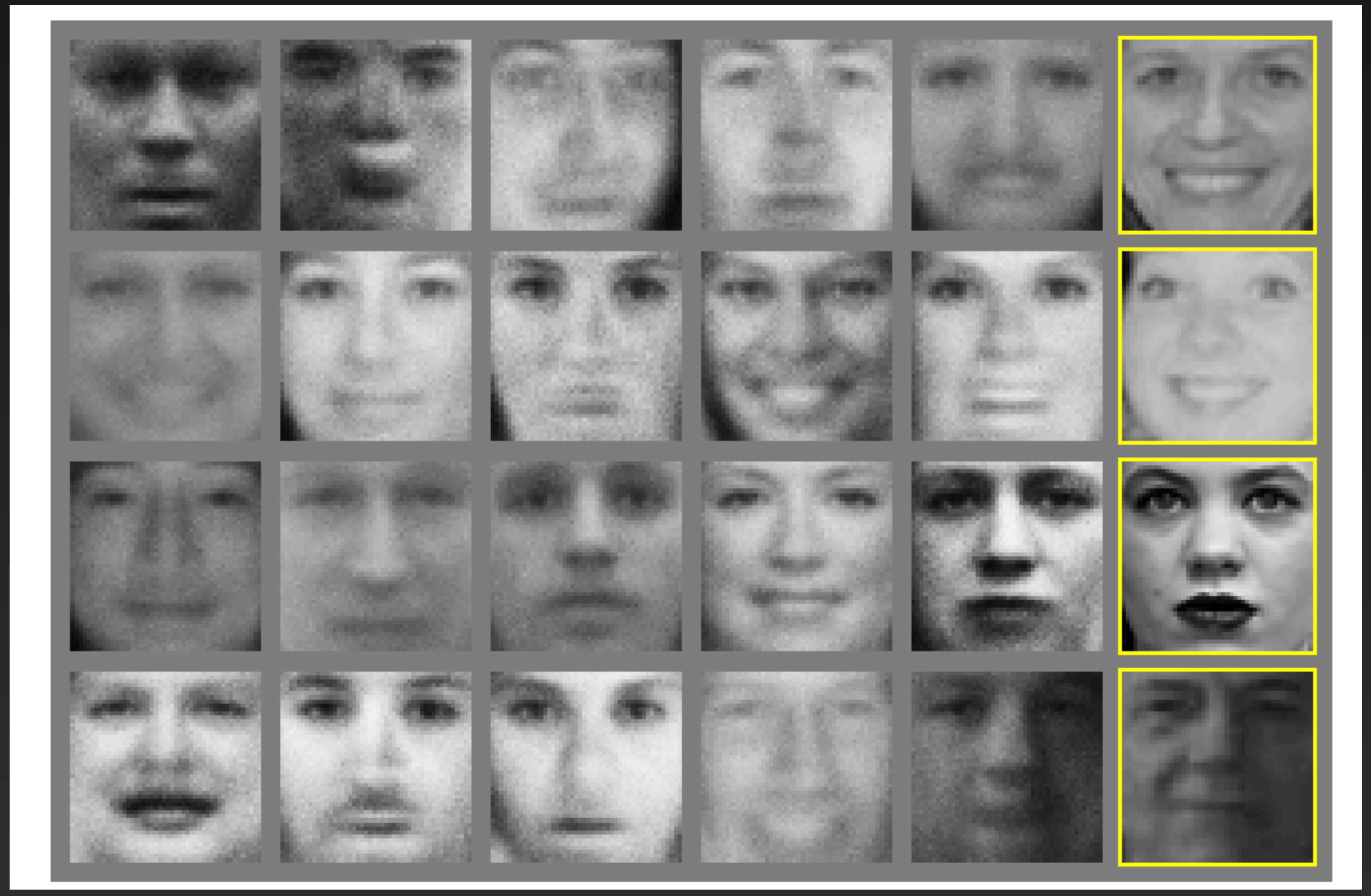
Foundation models in deepfake image detection

Defender's perspective: Simplify and improve performance of deepfake image detectors

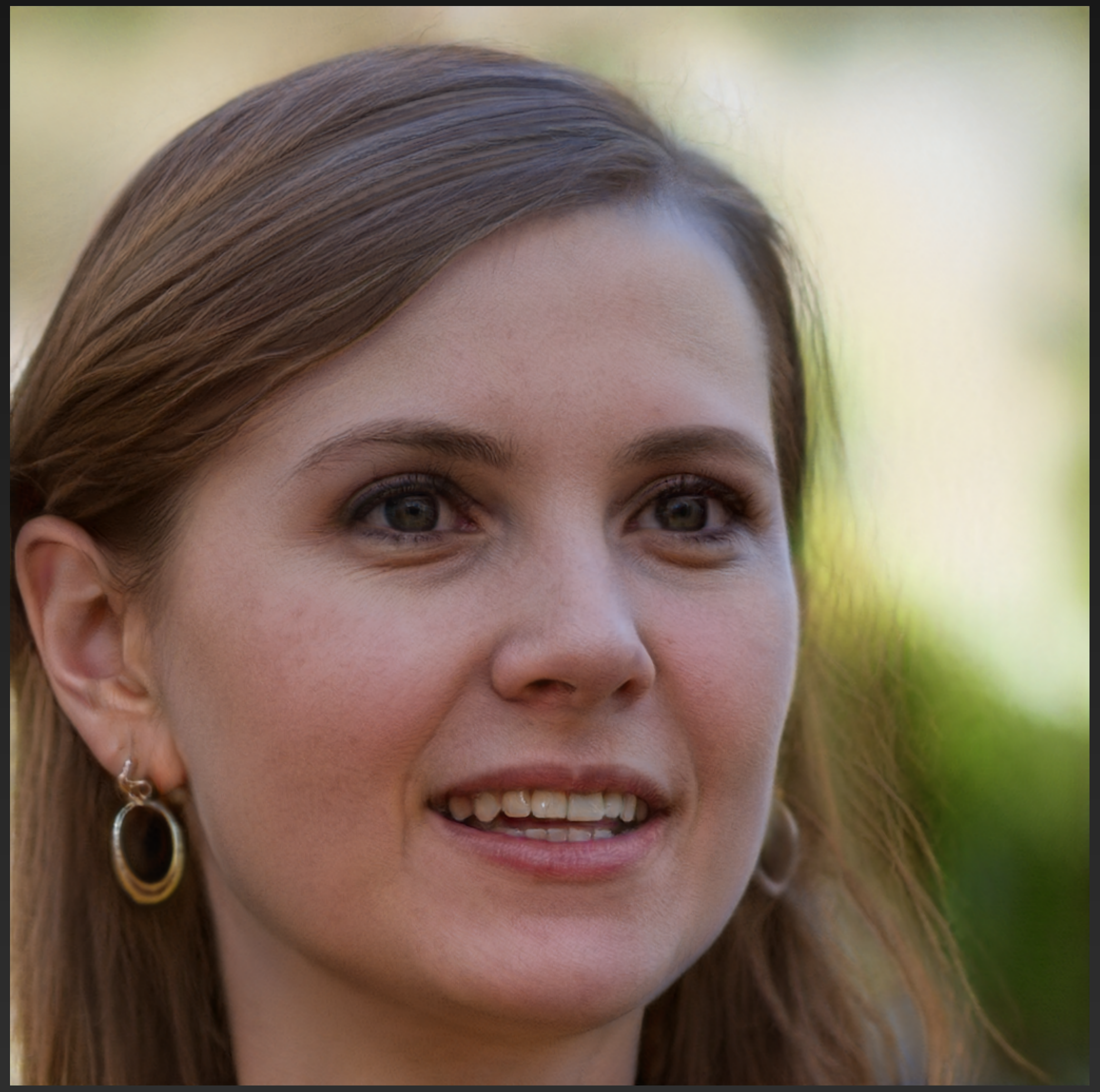
Attacker's perspective: Use foundation models to create custom deepfake generators

Deepfake images

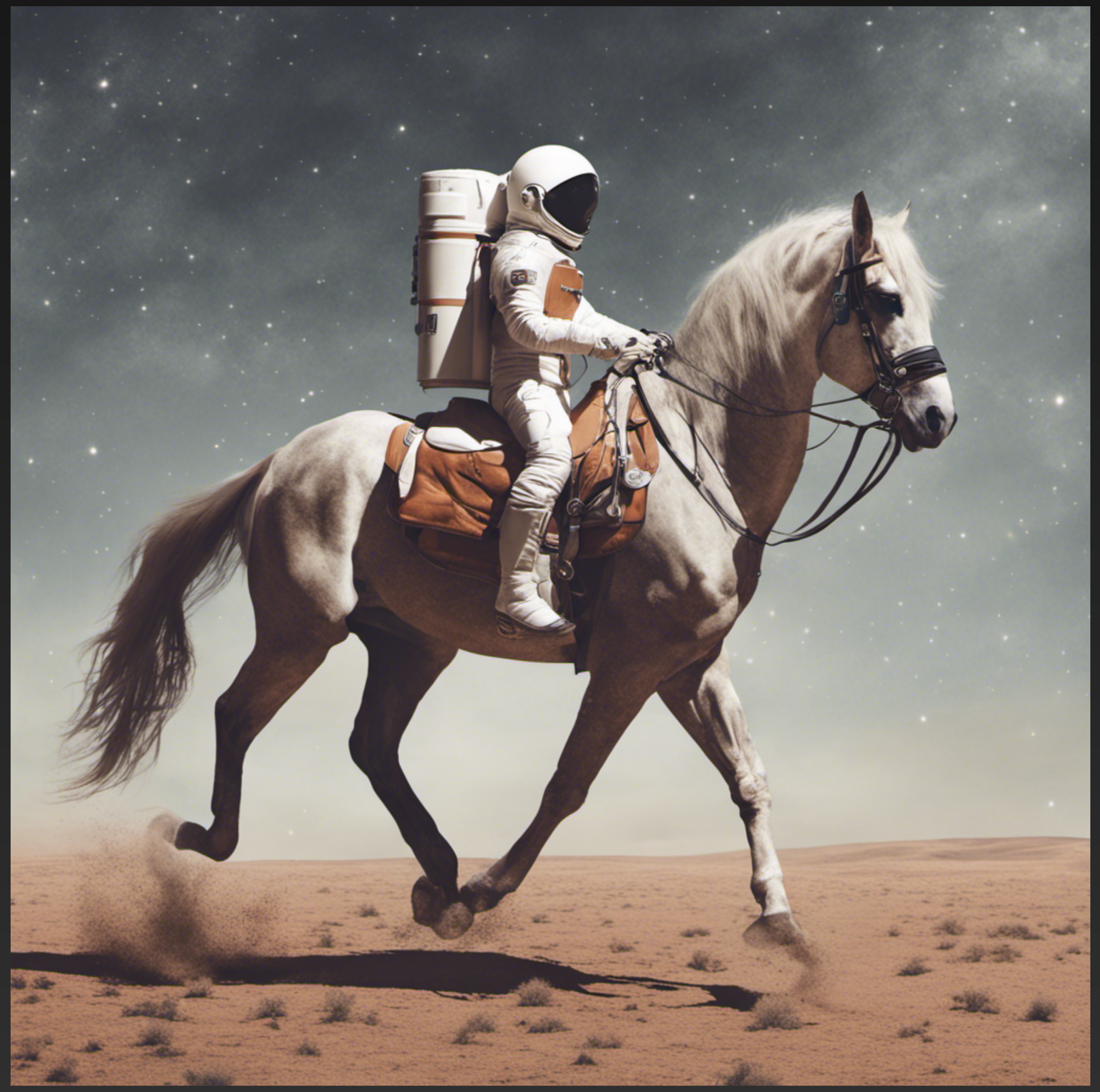
- Synthetic images generated by deep generative models



GAN (2014)



StyleGAN2 (2020)



Stable Diffusion (2023)

Image generators are getting better

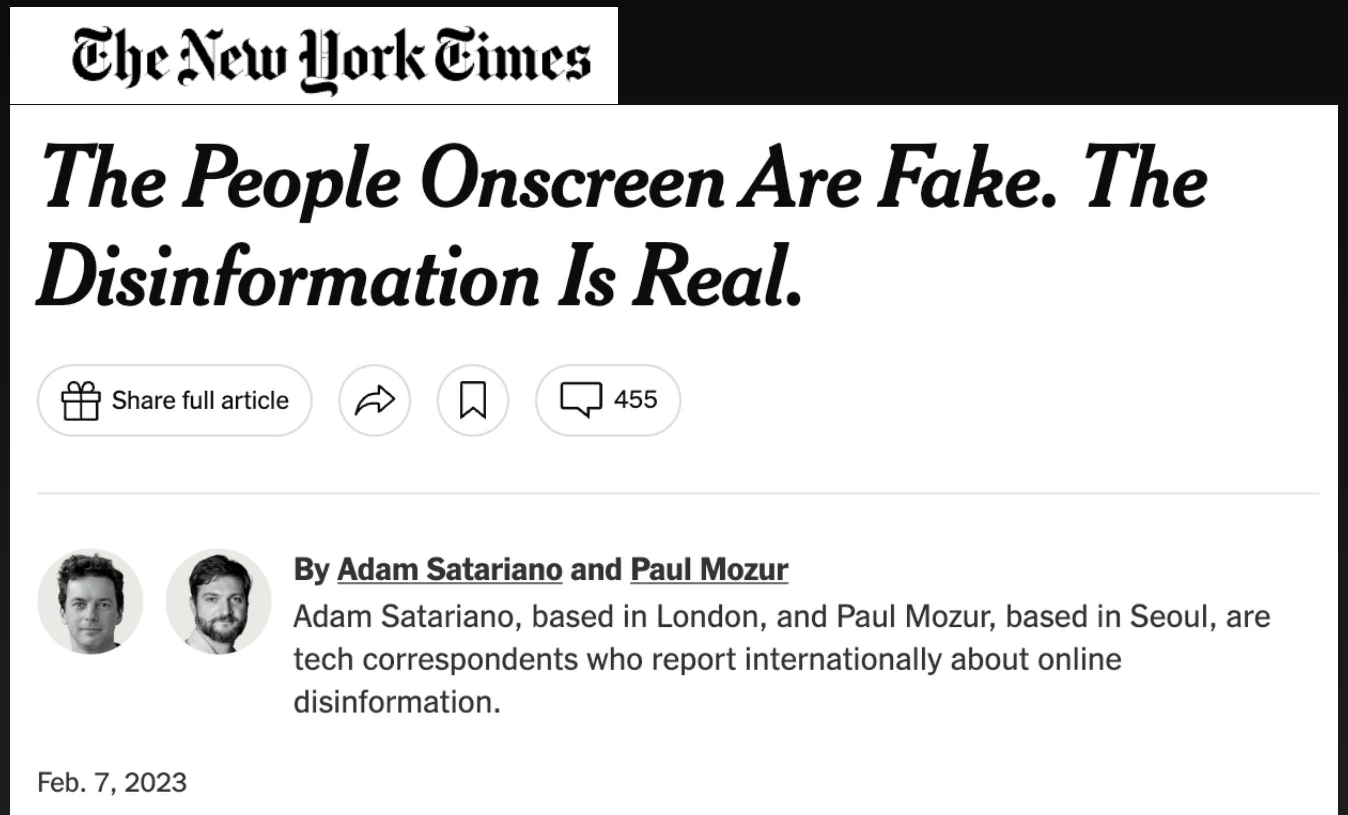
- Generating a deepfake image is as simple as typing in a text prompt

Stable Diffusion

Midjourney

DALL·E

Threats posed by deepfakes



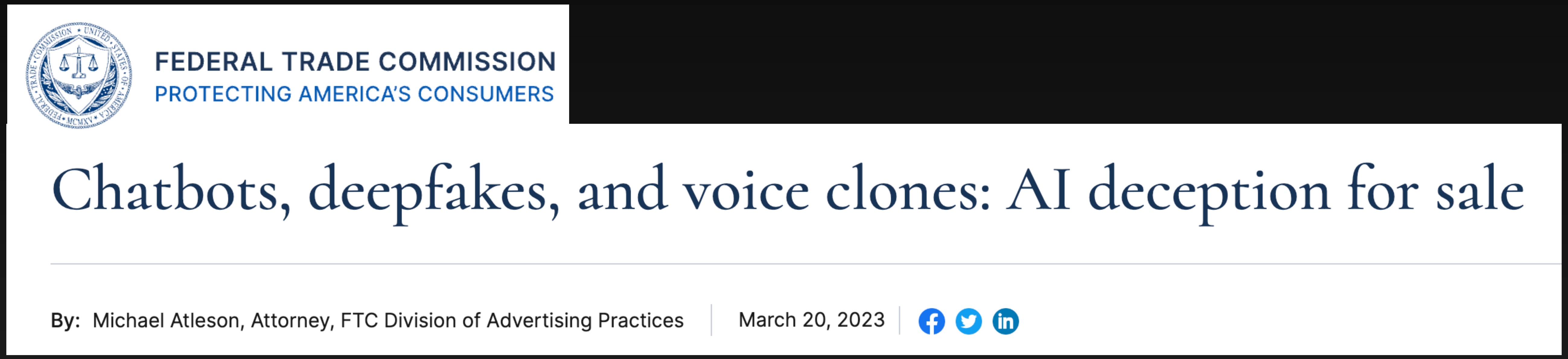
The New York Times

The People Onscreen Are Fake. The Disinformation Is Real.

Share full article 455

By Adam Satariano and Paul Mozur
Adam Satariano, based in London, and Paul Mozur, based in Seoul, are tech correspondents who report internationally about online disinformation.

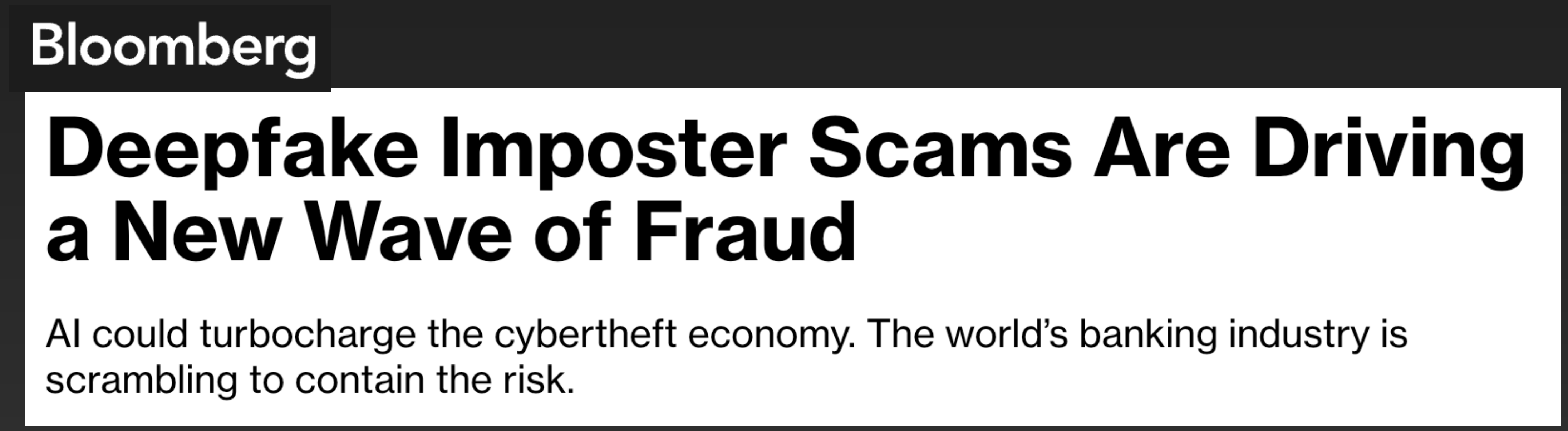
Feb. 7, 2023



FEDERAL TRADE COMMISSION
PROTECTING AMERICA'S CONSUMERS

Chatbots, deepfakes, and voice clones: AI deception for sale

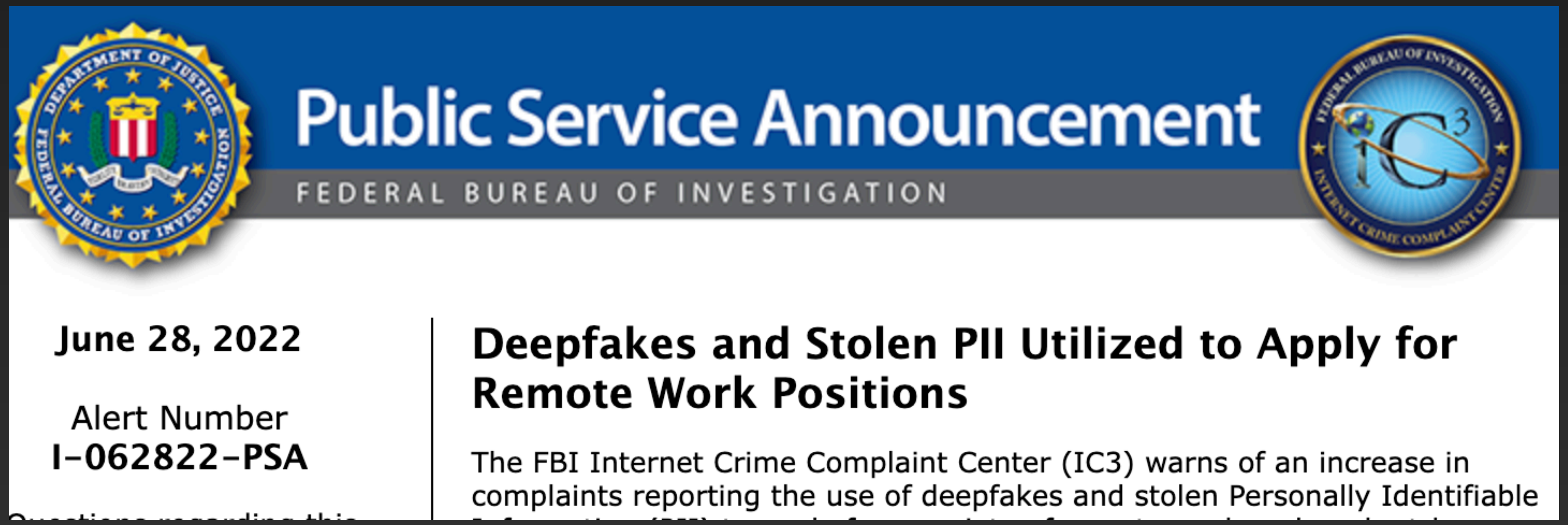
By: Michael Atleson, Attorney, FTC Division of Advertising Practices | March 20, 2023



Bloomberg

Deepfake Imposter Scams Are Driving a New Wave of Fraud

AI could turbocharge the cybertheft economy. The world's banking industry is scrambling to contain the risk.



Public Service Announcement
FEDERAL BUREAU OF INVESTIGATION

June 28, 2022

Alert Number
I-062822-PSA

Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions

The FBI Internet Crime Complaint Center (IC3) warns of an increase in complaints reporting the use of deepfakes and stolen Personally Identifiable

Can we build robust methods to detect deepfake images?

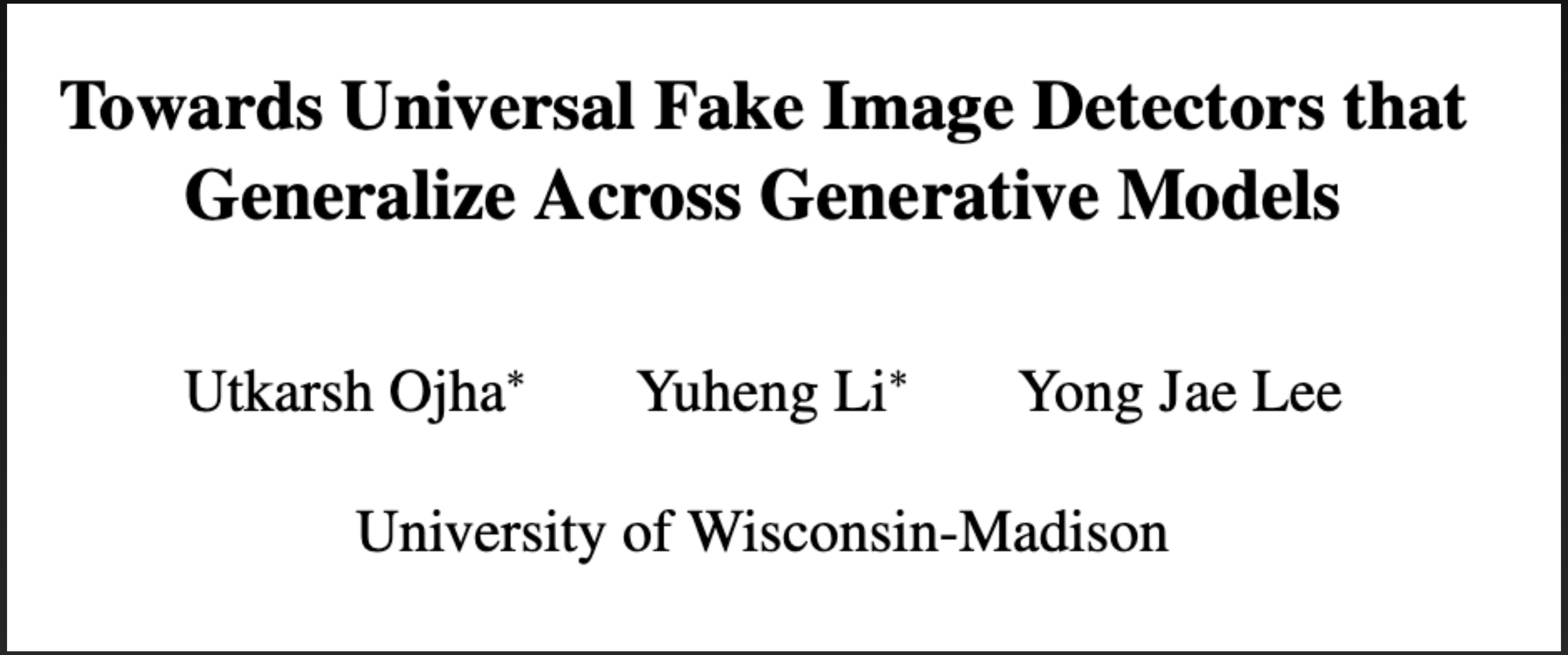
Extensive prior work on deepfake detection

Defense	Method	Performance
DCT (VISAPP 2024)	Artifacts in the frequency domain	Upto 97.7% Accuracy
UnivCLIP (CVPR 2023)	Use CLIP image-encoder features	Upto 100% Accuracy
DE-FAKE (CCS 2023)	Use CLIP text + image-encoder features	Upto 95.8% Accuracy
Resynthesis (IJCAI 2021)	Artifacts while reconstructing fake images	Upto 100% Accuracy
Patch-Forensics (ECCV 2020)	Local artifacts with small receptive fields	Upto 99.99% Average Precision
CNN-F (CVPR 2020)	CNN-based generators have detectable fingerprints	Upto 99.6% Average Precision
Gram-Net (CVPR 2020)	Artifacts in image texture statistics	Upto 99.1% Accuracy
MesoNet (IEEE WIFS 2018)	Neural networks with shallow layers	Upto 98.4% Accuracy

A grand challenge in this space is achieving good generalization performance

New notable detector: UnivCLIP (CVPR 2023)

Are vision foundation models the answer?



Towards Universal Fake Image Detectors that Generalize Across Generative Models

Utkarsh Ojha* Yuheng Li* Yong Jae Lee

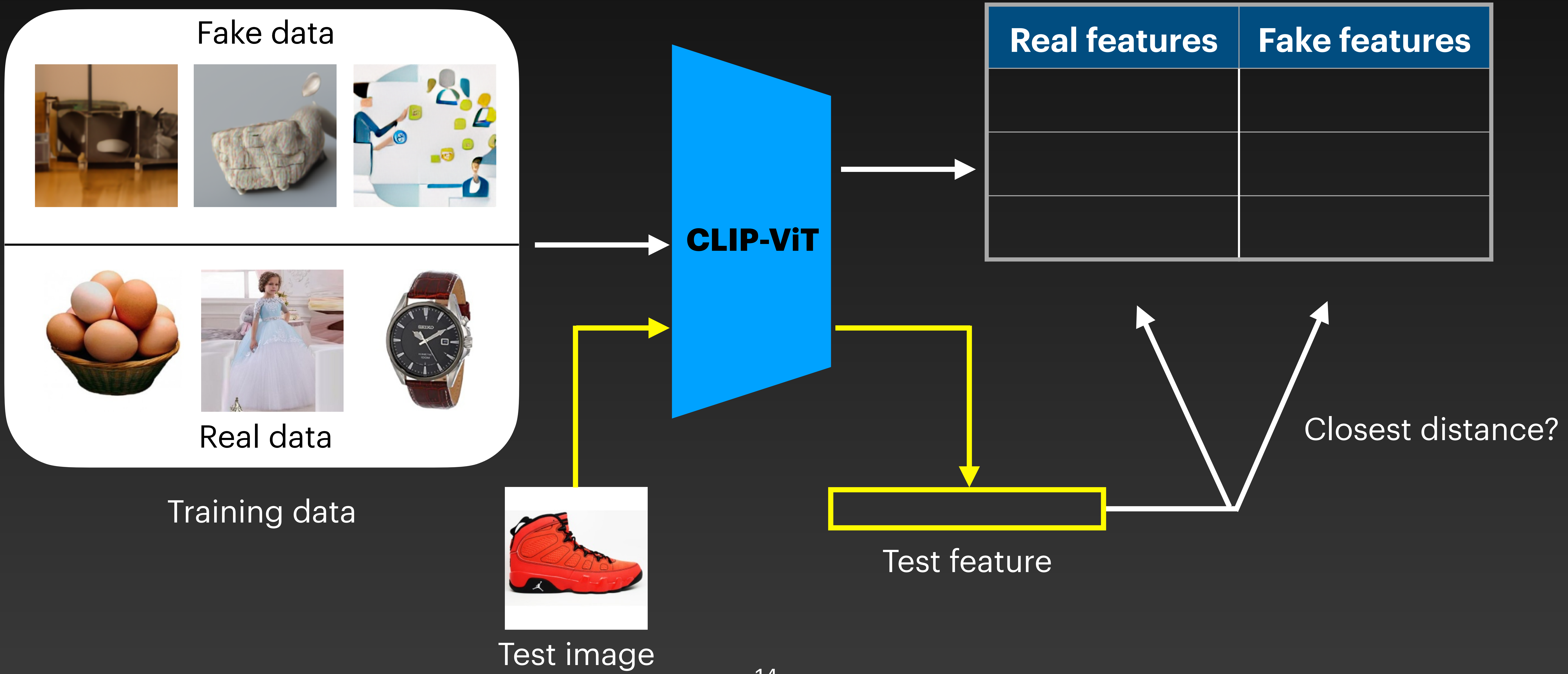
University of Wisconsin-Madison

This work claims impressive generalization performance!

UnivCLIP methodology

Simply extract features using a foundation model

- UnivCLIP uses the CLIP-ViT foundation model (Trained on 400M images)



Real features	Fake features

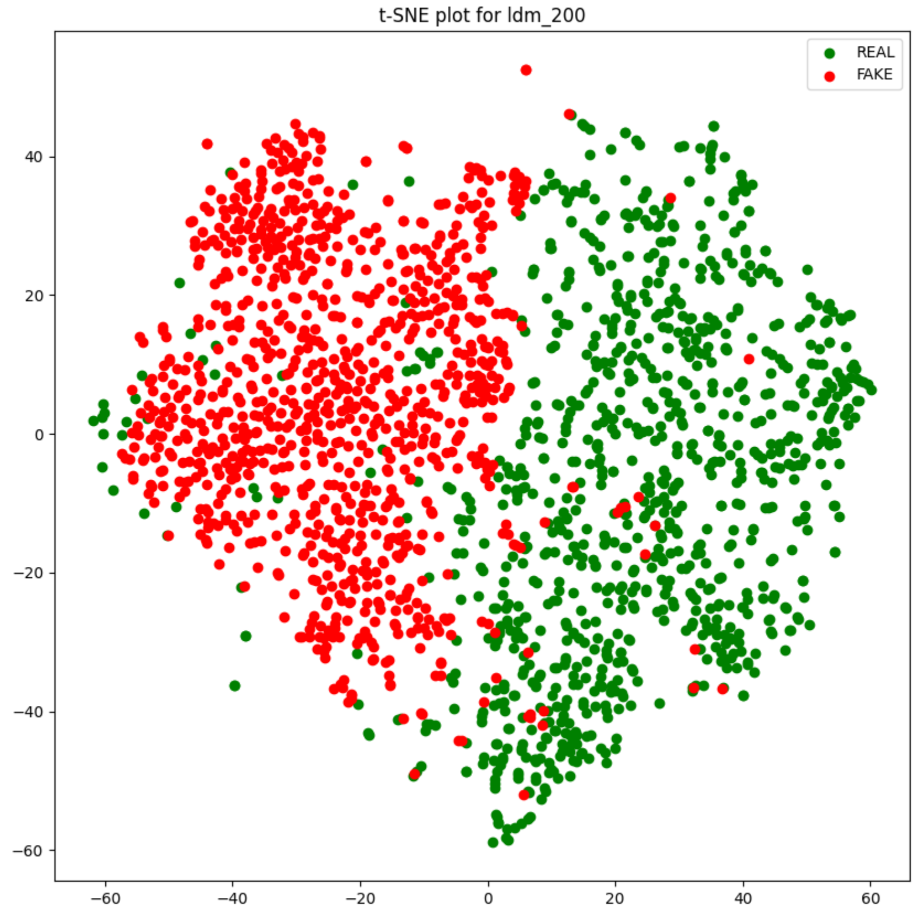
But some problems in their exp. setup

- They are not controlling for content or quality

UnivCLIP dataset

Real

Fake

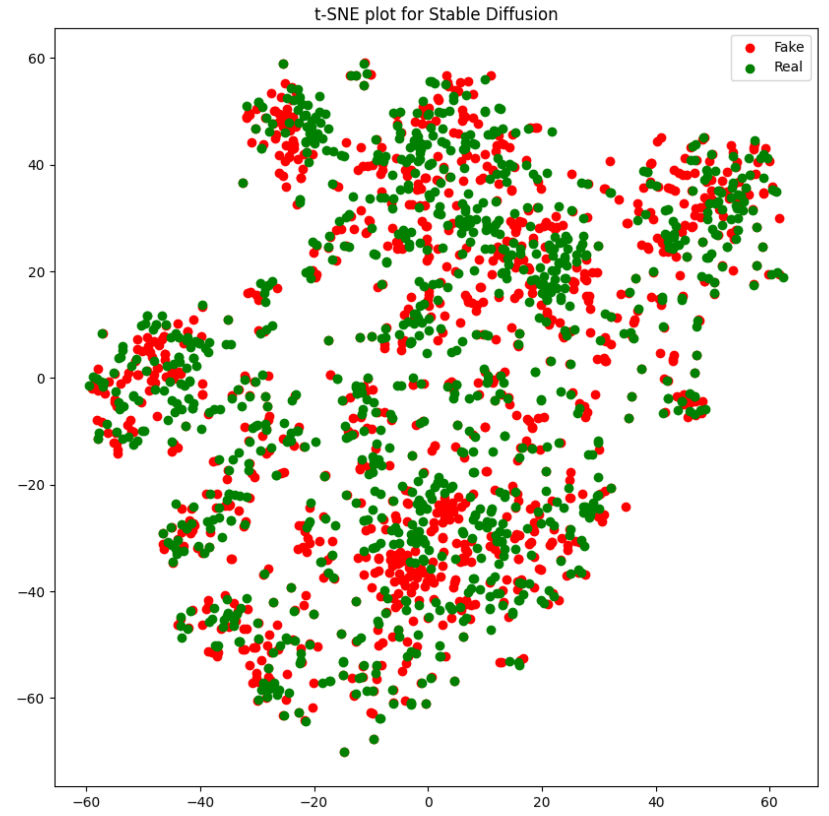


94% detection accuracy!

Our dataset (Stable Diffusion)

Real

Fake



49% detection accuracy!

Does UnivCLIP really generalize well?

- We trained UnivCLIP on our Stable Diffusion dataset (Realistic Vision)
  - Obtained an F1 score of 93%, Recall of 92% (both for fake class)

How can we test generalization in the real world?

Generalization studies in prior work

- In prior work, defenses were only evaluated with a few generative models

Deepfake defenses

GANs

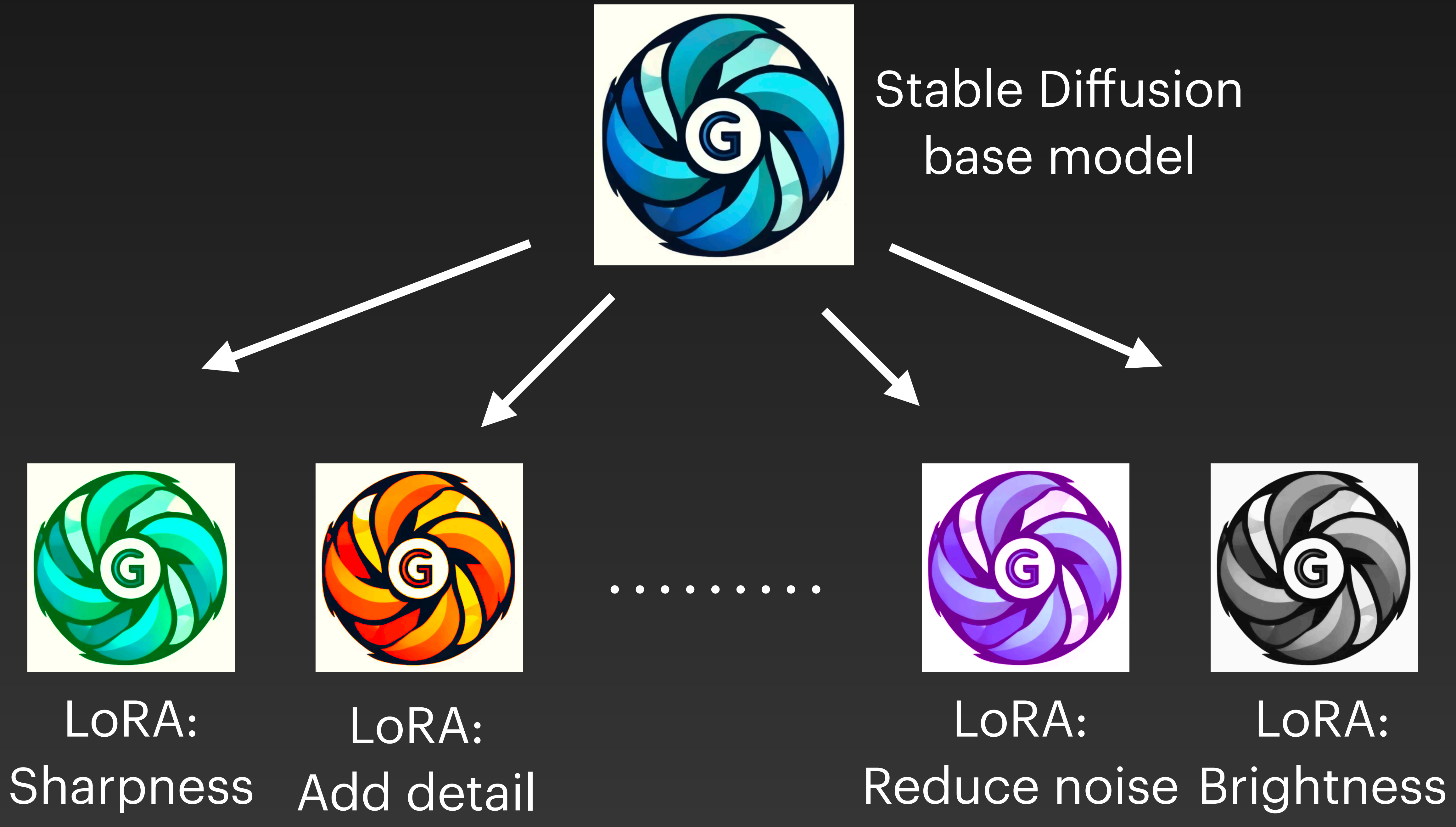
Diffusion models

Emergence of user-customized models expands the threat surface

New threat: User-customized versions of foundation models

Stable Diffusion (SD) as a case study

- Using LORA-based fine-tuning, users are creating their own versions of SD
  - Over 3,000 SD variants on CivitAI and HuggingFace



Stable Diffusion base model

LoRA: Sharpness

LoRA: Add detail

....

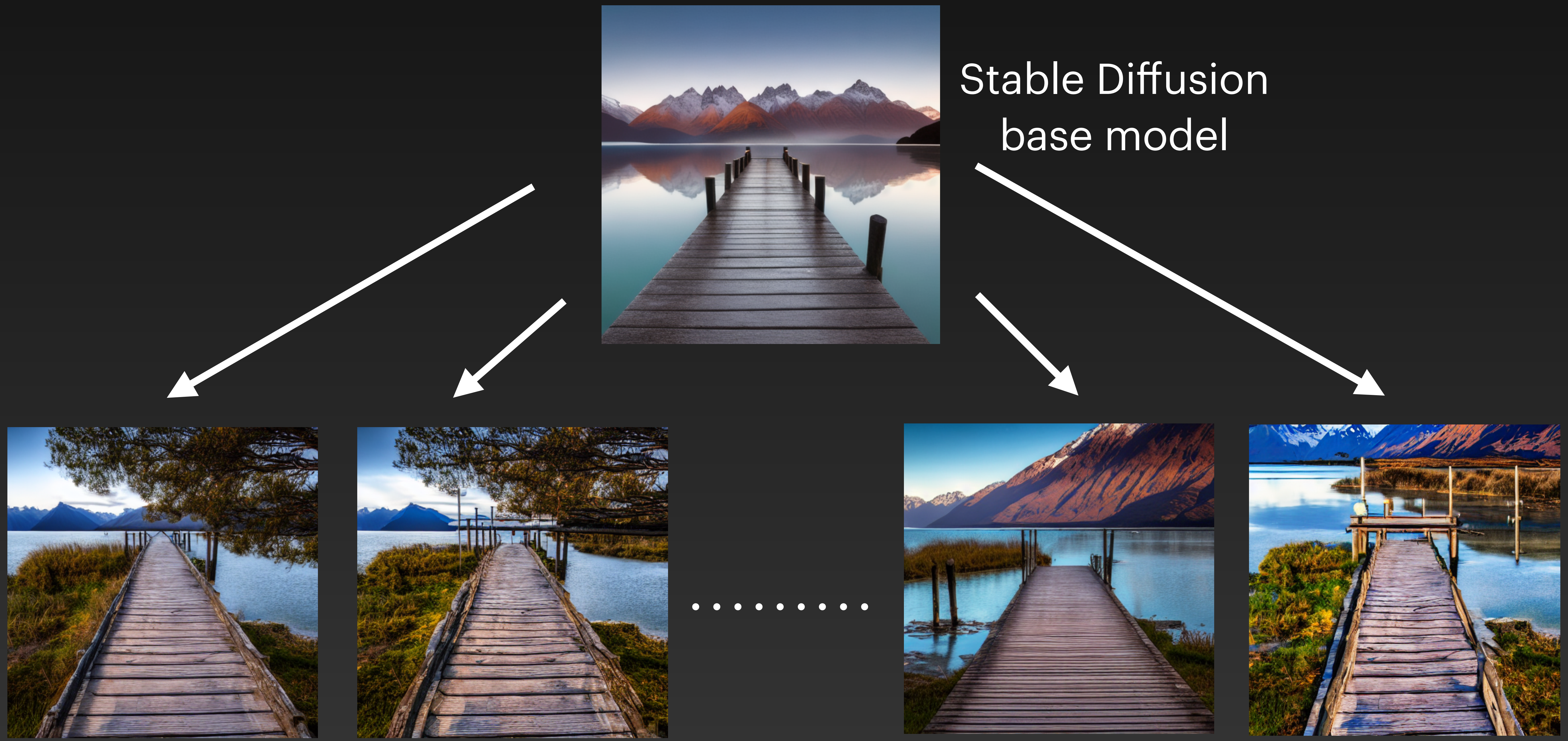
LoRA: Reduce noise

LoRA: Brightness

Generalization against user-customized models

Stable Diffusion (SD) as a case study

- Using LORA-based fine-tuning, users are creating their own versions of SD
  - Over 3,000 SD variants on CivitAI and HuggingFace



Stable Diffusion base model

LoRA: Sharpness

LoRA: Add detail

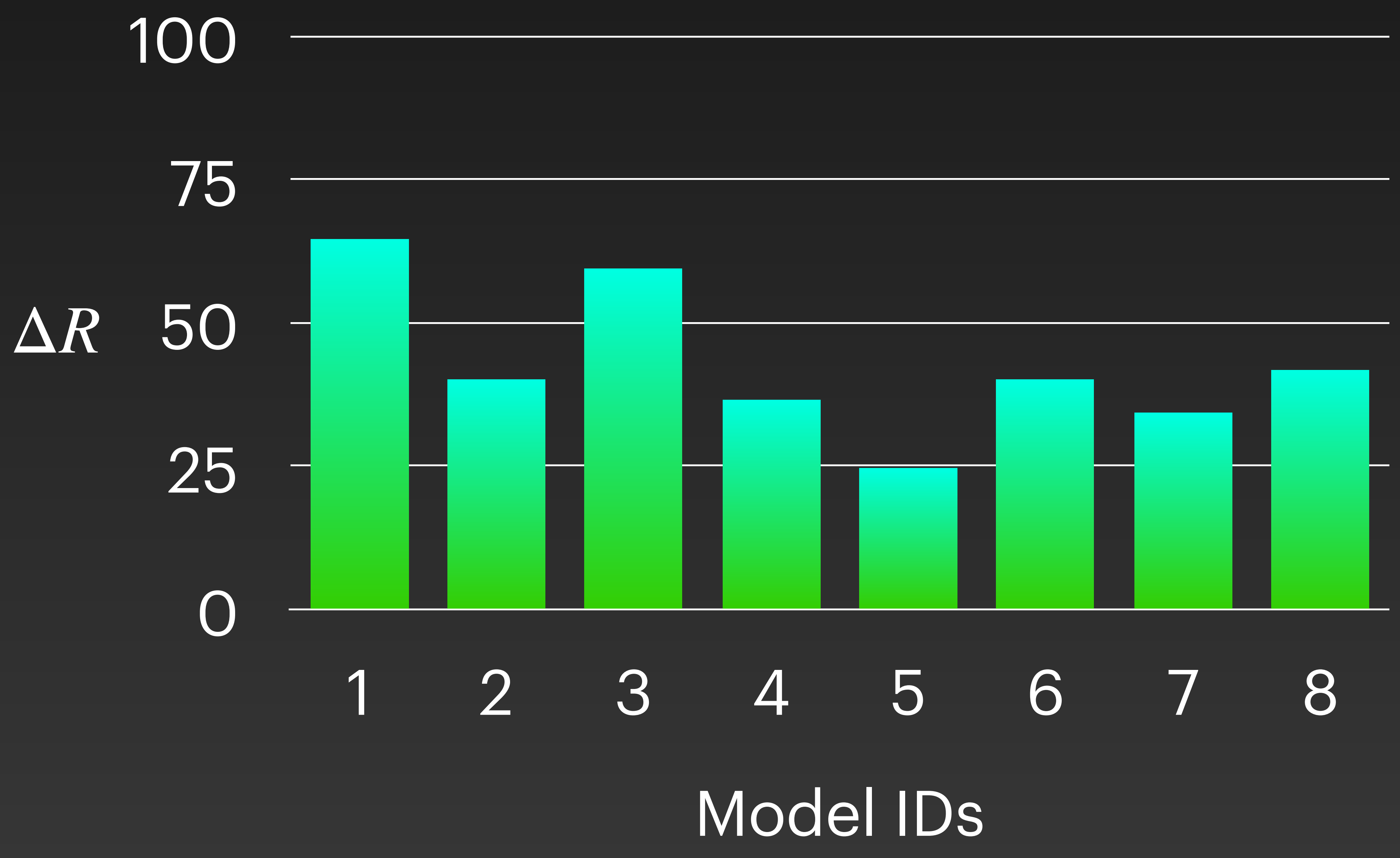
LoRA: Reduce noise

LoRA: Brightness

UnivCLIP generalizes poorly

UnivCLIP claims to be SOTA in generalization perf.

- We tested generalization on 8 user-customized SD variants (from CivitAI)
  - We measure ΔR (perc. degradation in Recall of fake images)



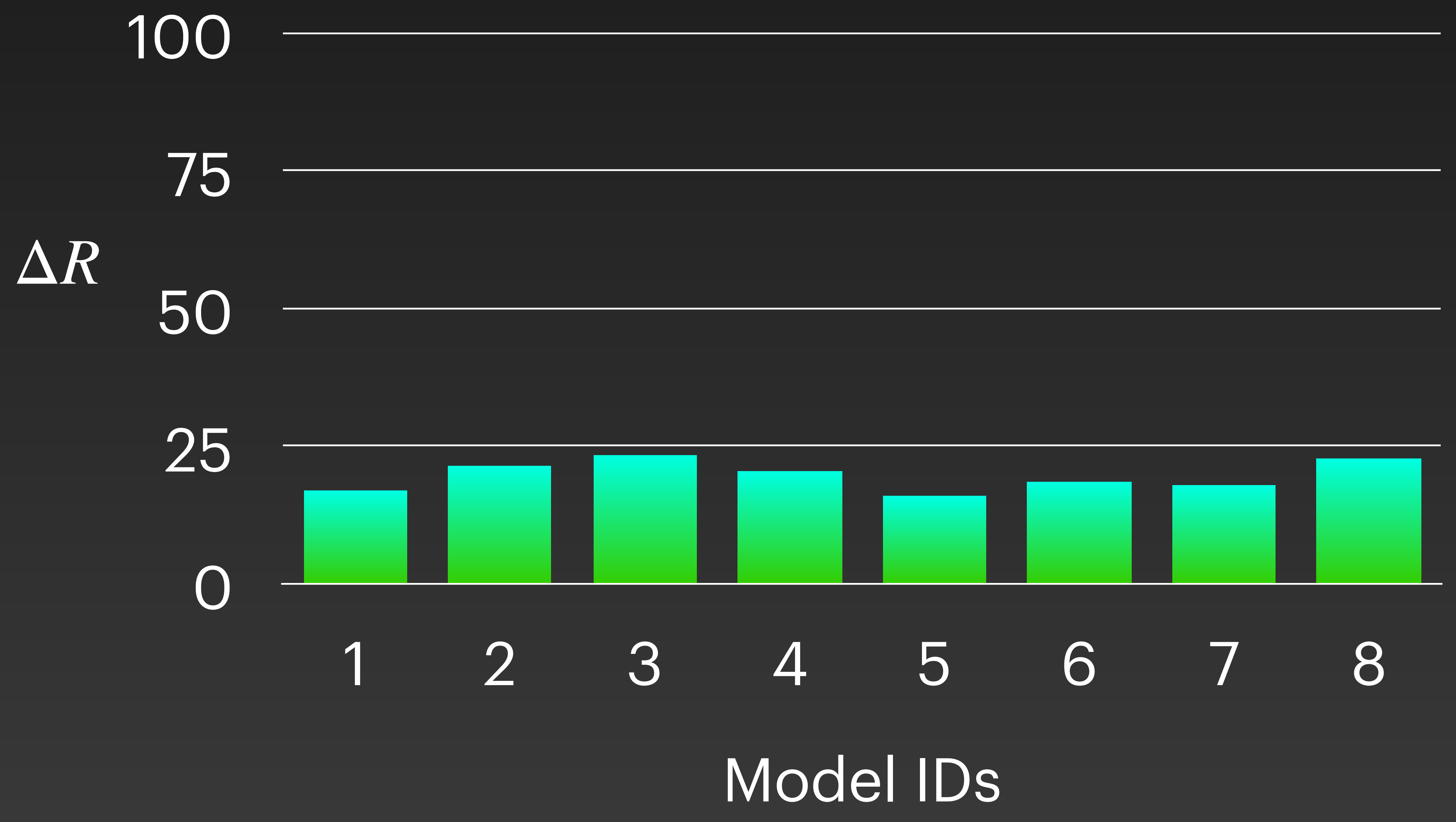
UnivCLIP shows significant degradation in Recall.

Average ΔR of 42%

Max ΔR of 64.5%

How well do other defenses generalize?

- Except once defense, all the defenses generalize poorly
  - Max ΔR from 64.5% to 90%
- Defense leveraging artifacts in frequency spectrum, shows the most promise



DCT achieves
Average ΔR of 19.6%
Max ΔR of 23.5%

Can we still improve generalization using foundation models?

- Idea: Fuse features from foundation model with domain-specific features

Avg. ΔR

Only UnivCLIP Features	UnivCLIP + DCT features
42%	8%

More effective when features from domain-agnostic foundation models are combined with domain-specific frequency features

Opportunities and challenges

Opportunities and challenges

- Challenges:
  - Easy customizability of foundation models presents new challenges
    - Open challenge: Customized generators threaten existing defenses

Opportunities and challenges

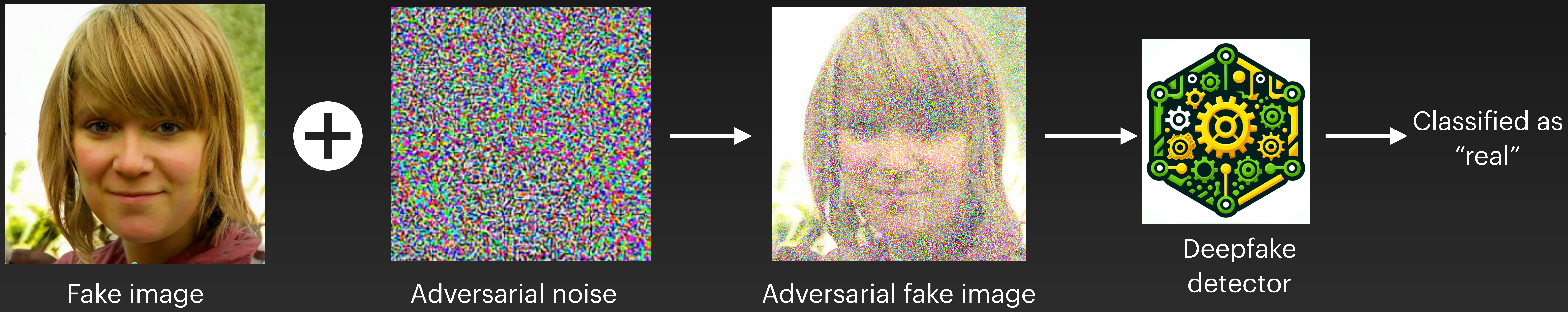
- Challenges:
  - Easy customizability of foundation models presents new challenges
    - Open challenge: Customized generators threaten existing defenses
- Opportunities:
  - Can simplify defense pipelines
  - Combined with domain-specific features can provide perf. benefit

Foundation models in deepfake image detection

Attacker’s perspective: Use foundation models to craft adversarial fake images

How can we evade deepfake detectors?

- A traditional idea is to add adversarial noise (perturbations)



Fake image

Adversarial noise

Adversarial fake image

Deepfake detector

Classified as “real”

Such adversarial perturbations can degrade image quality

Can we create adversarial images without adding noise?

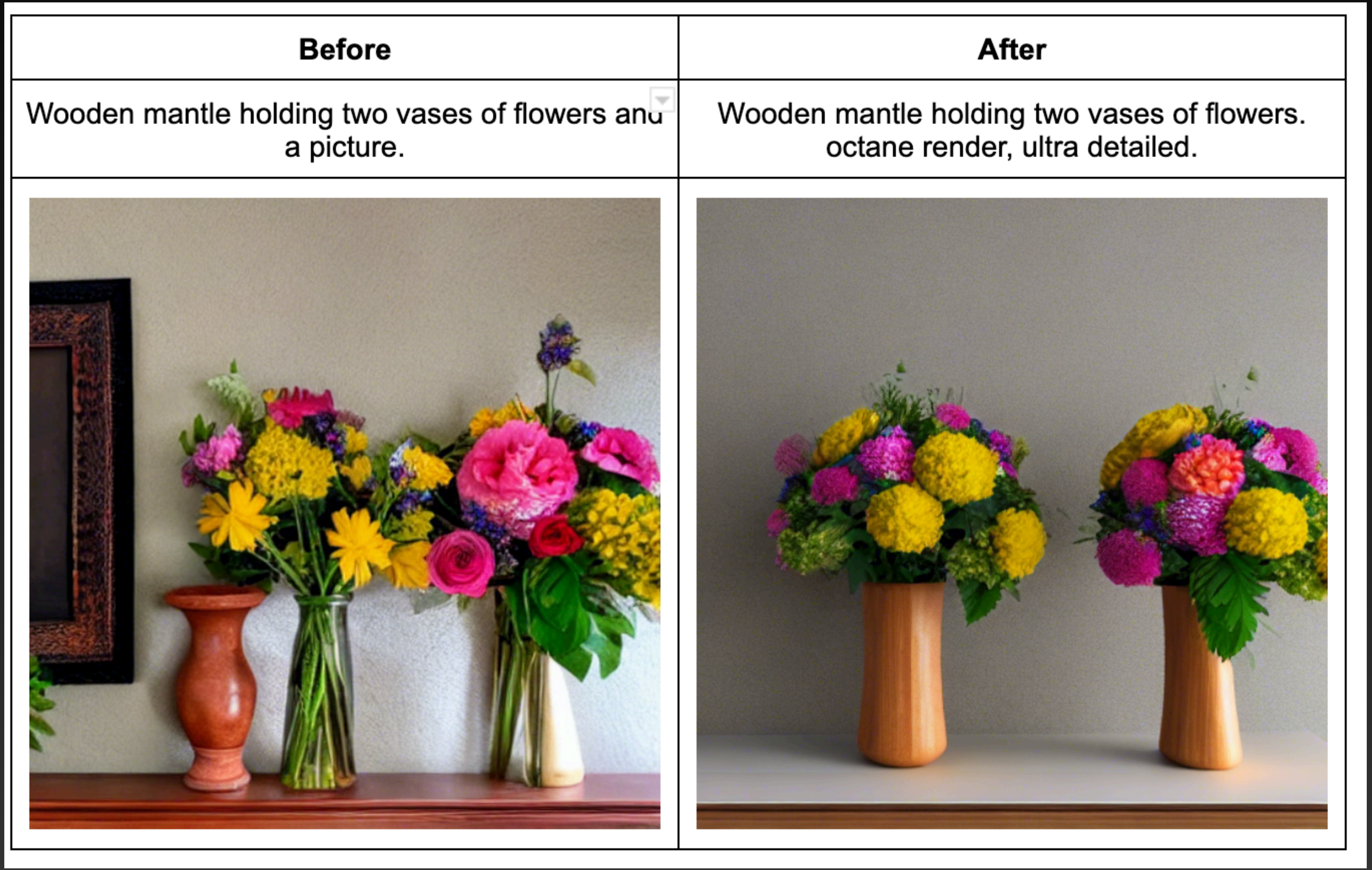
- We tried arbitrary prompt modifications with Stable Diffusion
- Tested against the CNN-fingerprint defense

(Correctly) Detected as fake

(Wrongly) Detected as real

Can we create adversarial images without adding noise?

- We tried arbitrary prompt modifications with Stable Diffusion
- Tested against the CNN-fingerprint defense



Before	After
Wooden mantle holding two vases of flowers anu a picture.	Wooden mantle holding two vases of flowers. octane render, ultra detailed.

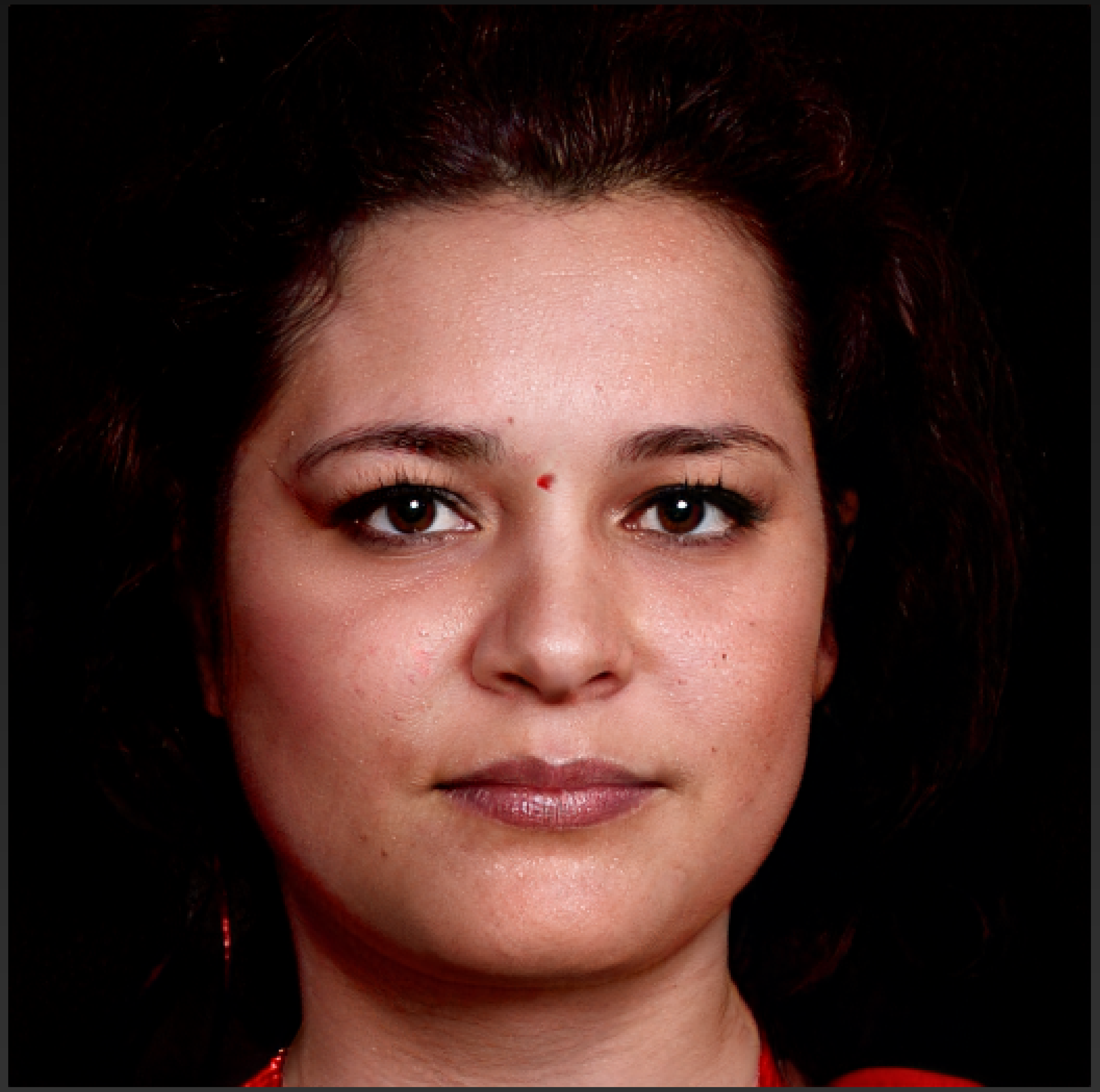
(Correctly) Detected as fake

(Wrongly) Detected as real

Can we create adversarial images without adding noise?

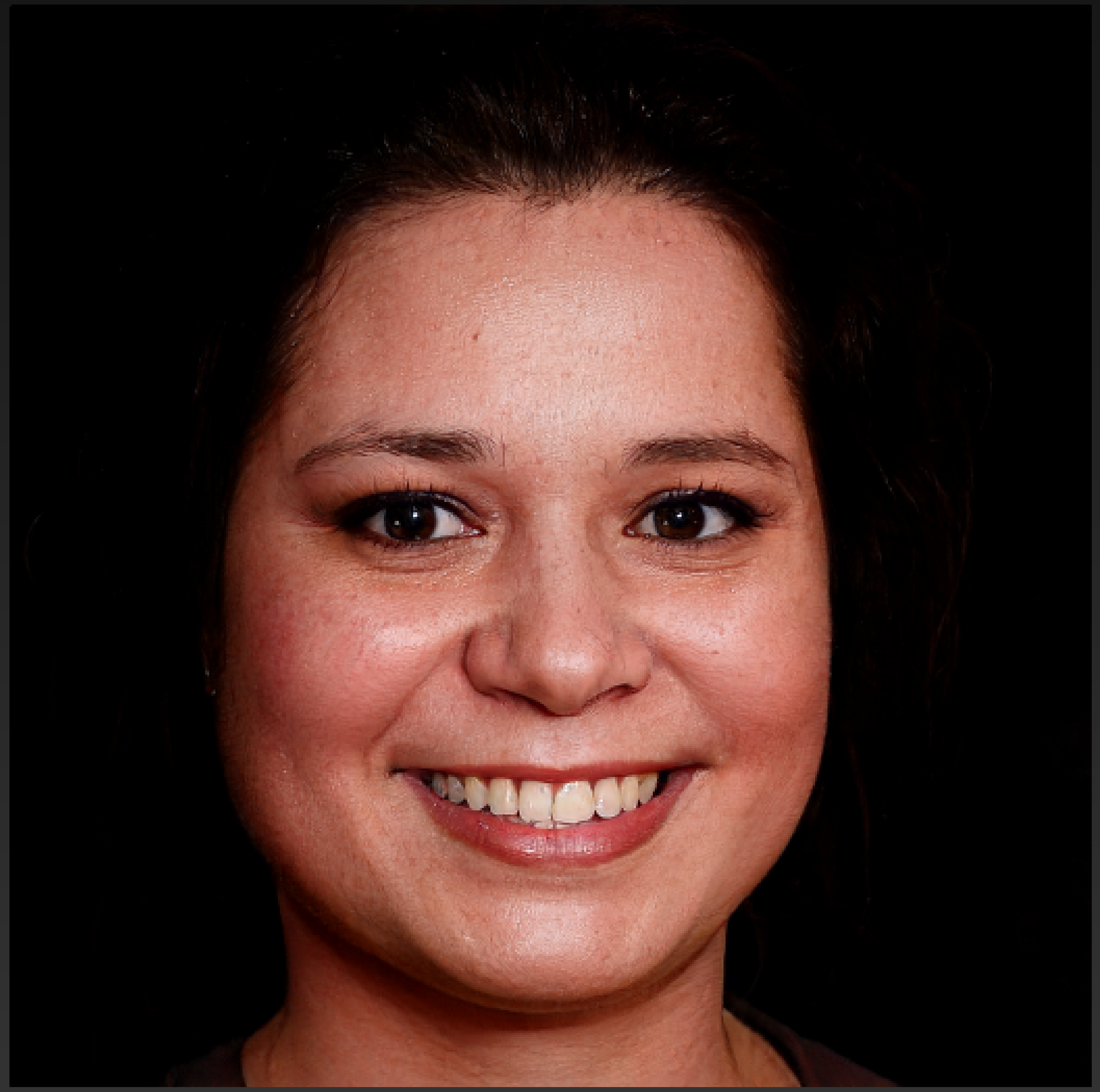
- We tried arbitrary prompt modifications with Stable Diffusion
- Tested against the CNN-fingerprint defense

A face



(Correctly) Detected as fake

A face with a smile



(Wrongly) Detected as real

Can we create adversarial images without adding noise?

- It is possible to create adversarial fake images
  - With careful modification of the content with no additional noise
  - While preserving high-level content semantics

How can we systematically create such adversarial images?

Leveraging foundation models to create adversarial images

- We assume a black-box setting, with no queries to the victim model

Leveraging foundation models to create adversarial images

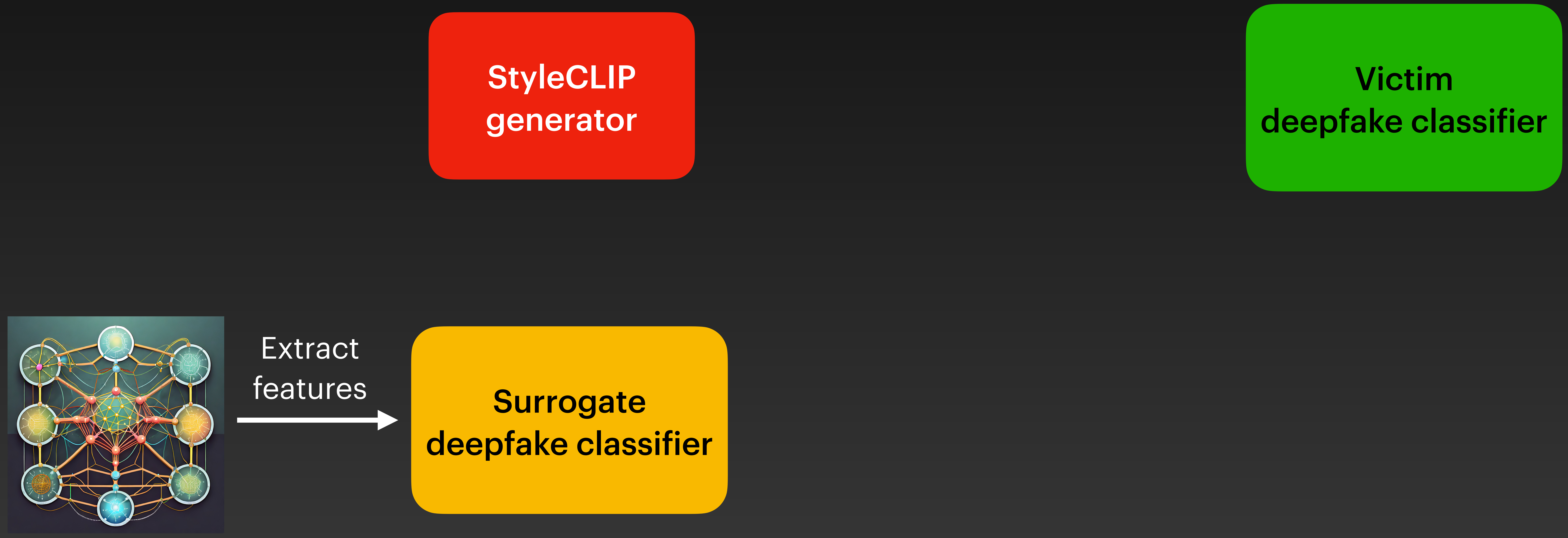
- We assume a black-box setting, with no queries to the victim model

StyleCLIP generator

Victim deepfake classifier

Leveraging foundation models to create adversarial images

- We assume a black-box setting, with no queries to the victim model



StyleCLIP generator

Victim deepfake classifier

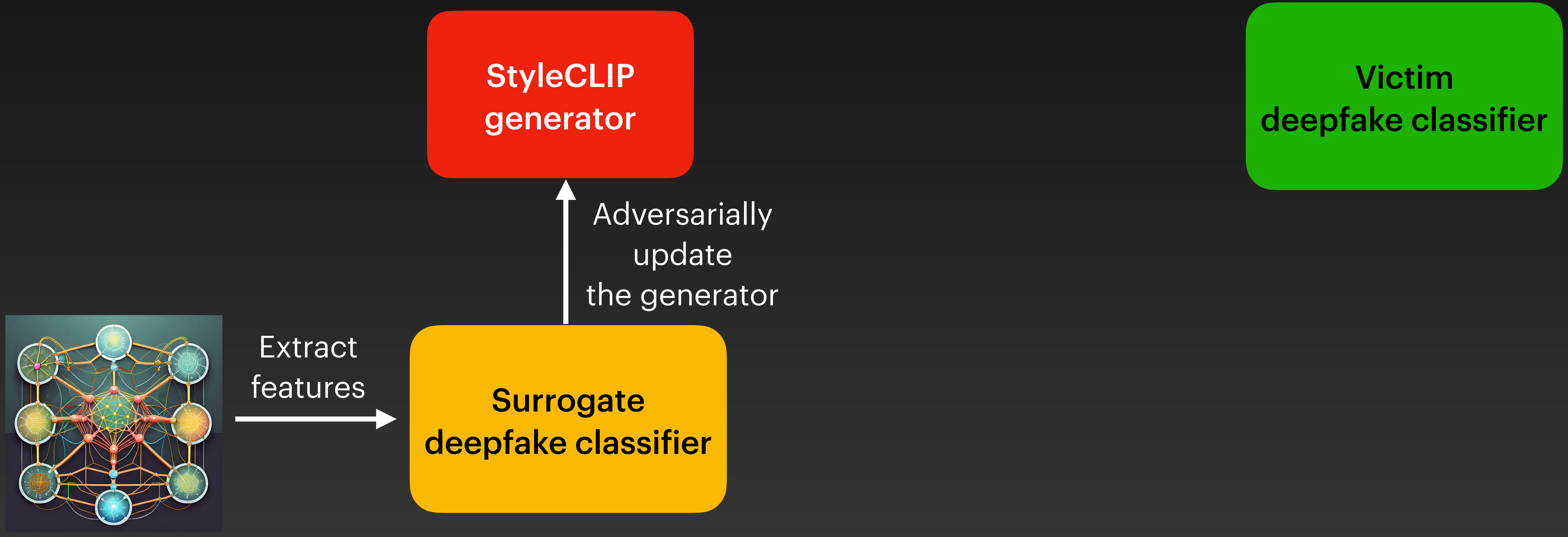
Extract features

Surrogate deepfake classifier

Foundation model

Leveraging foundation models to create adversarial images

- We assume a black-box setting, with no queries to the victim model



StyleCLIP generator

Victim deepfake classifier

Adversarially update the generator

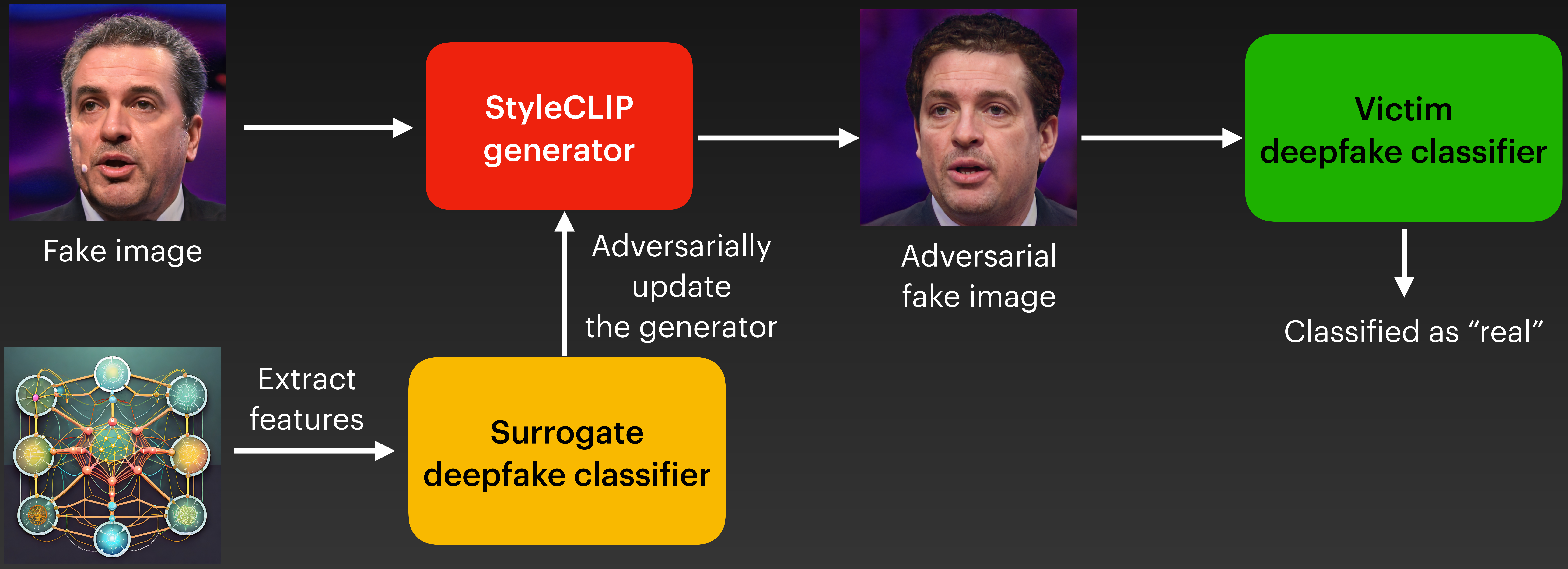
Extract features

Surrogate deepfake classifier

Foundation model

Leveraging foundation models to create adversarial images

- We assume a black-box setting, with no queries to the victim model



Fake image

StyleCLIP generator

Adversarial fake image

Victim deepfake classifier

Classified as “real”

Adversarially update the generator

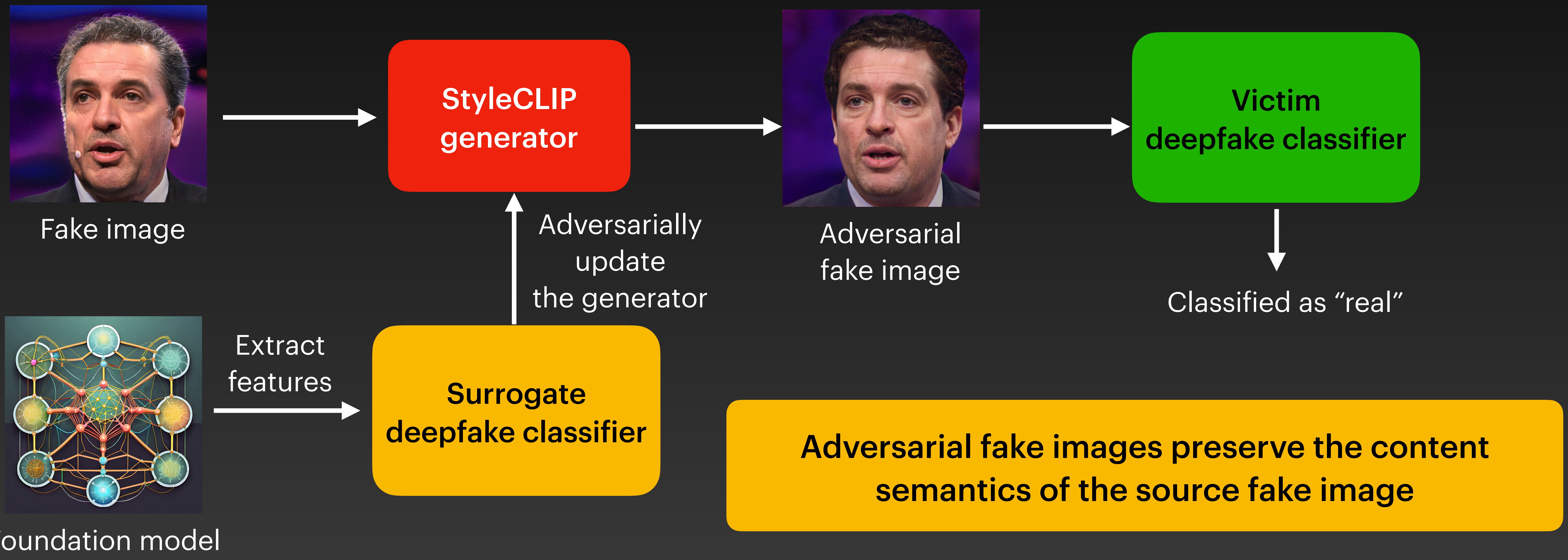
Foundation model

Extract features

Surrogate deepfake classifier

Leveraging foundation models to create adversarial images

- We assume a black-box setting, with no queries to the victim model



Fake image

StyleCLIP generator

Adversarial fake image

Victim deepfake classifier

Classified as "real"

Adversarially update the generator

Extract features

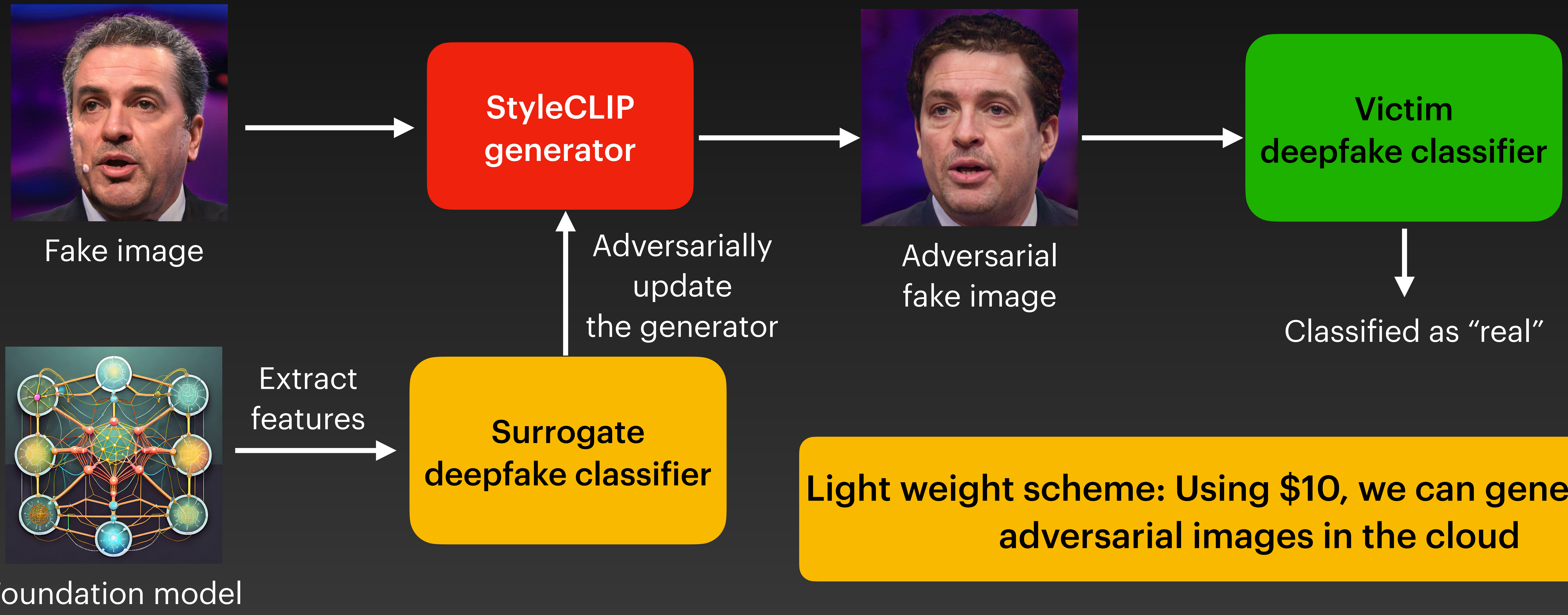
Surrogate deepfake classifier

Foundation model

Adversarial fake images preserve the content semantics of the source fake image

Leveraging foundation models to create adversarial images

- We assume a black-box setting, with no queries to the victim model



Fake image

StyleCLIP generator

Adversarial fake image

Victim deepfake classifier

Classified as “real”

Adversarially update the generator

Extract features

Surrogate deepfake classifier

Foundation model

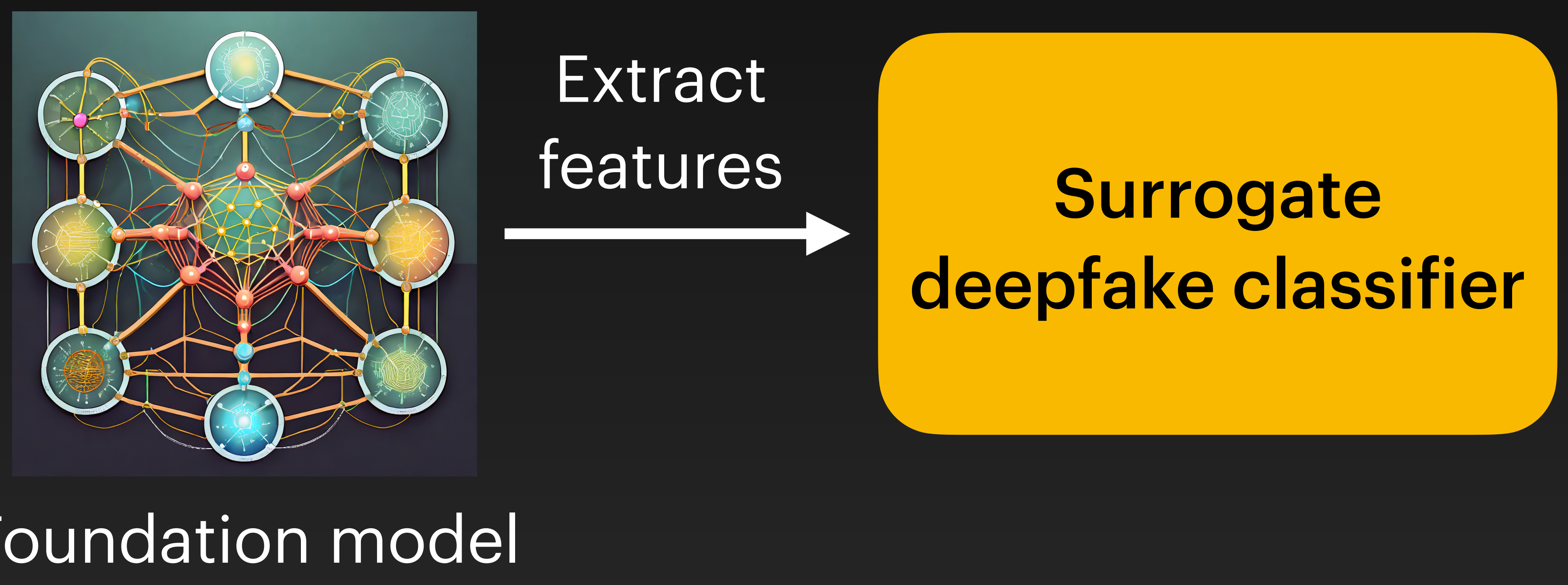
Light weight scheme: Using $10, we can generate 840 adversarial images in the cloud

Our attack is low cost

With only $10, using an NVIDIA A100 cloud GPU, we can generate 840 adversarial fake images

Our attack is powered by surrogate deepfake classifiers

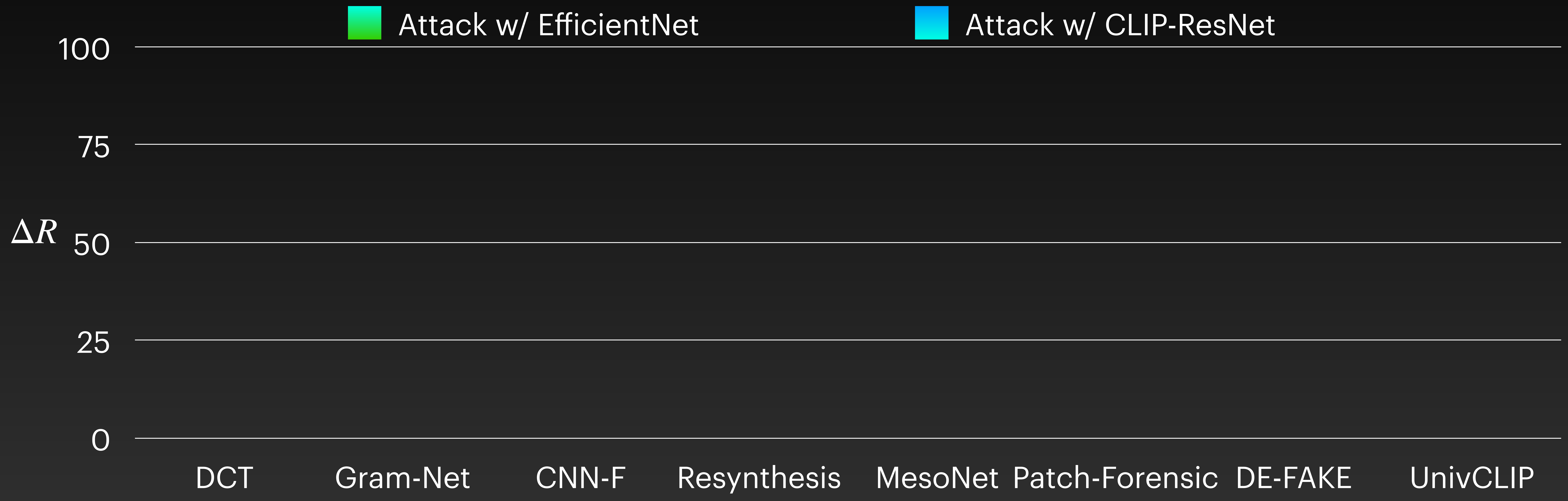
Using foundation models



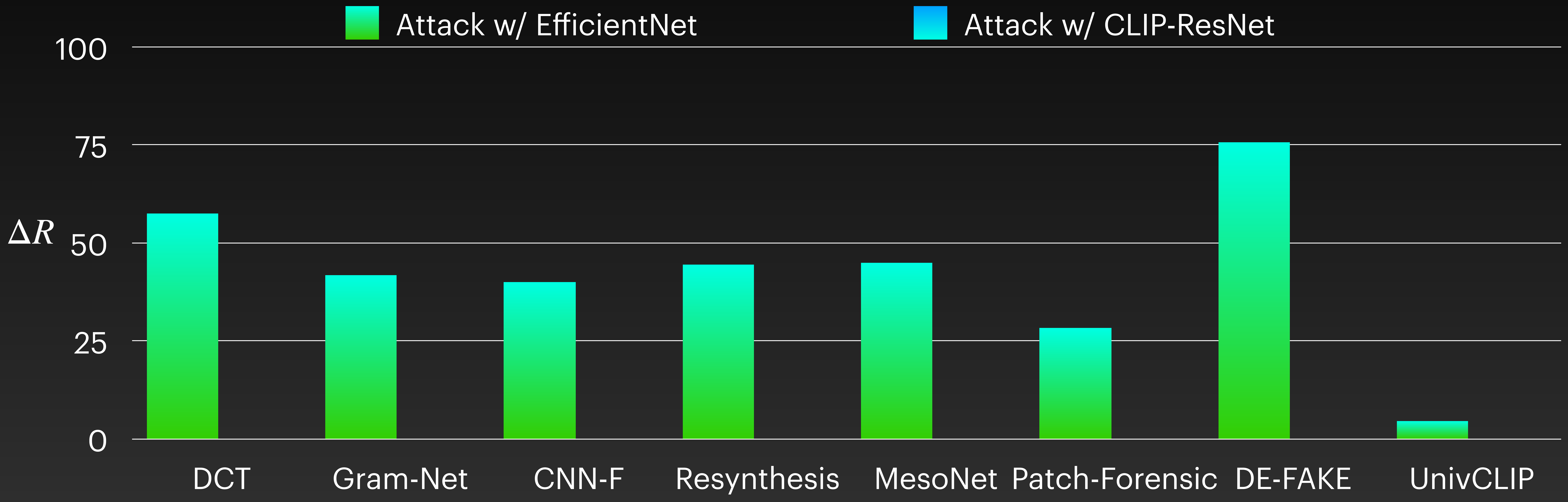
EfficientNet: Trained on 14M images
CLIP-ResNet: Trained on 400M images

How effective are these adversarial images?

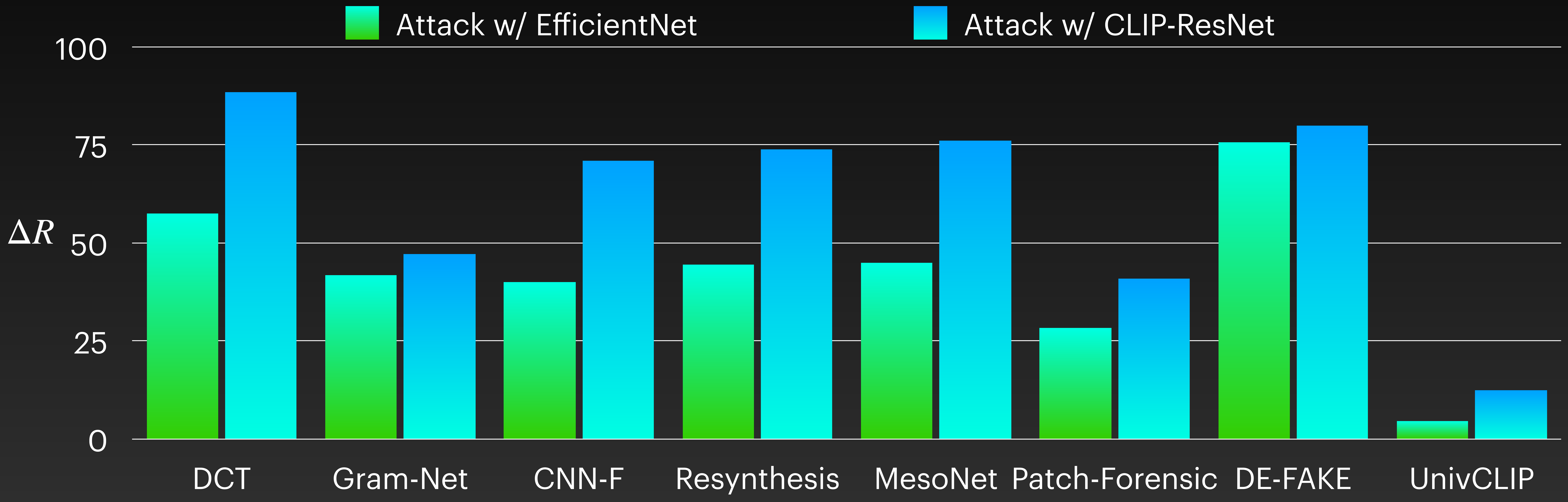
How effective are these adversarial images?



How effective are these adversarial images?

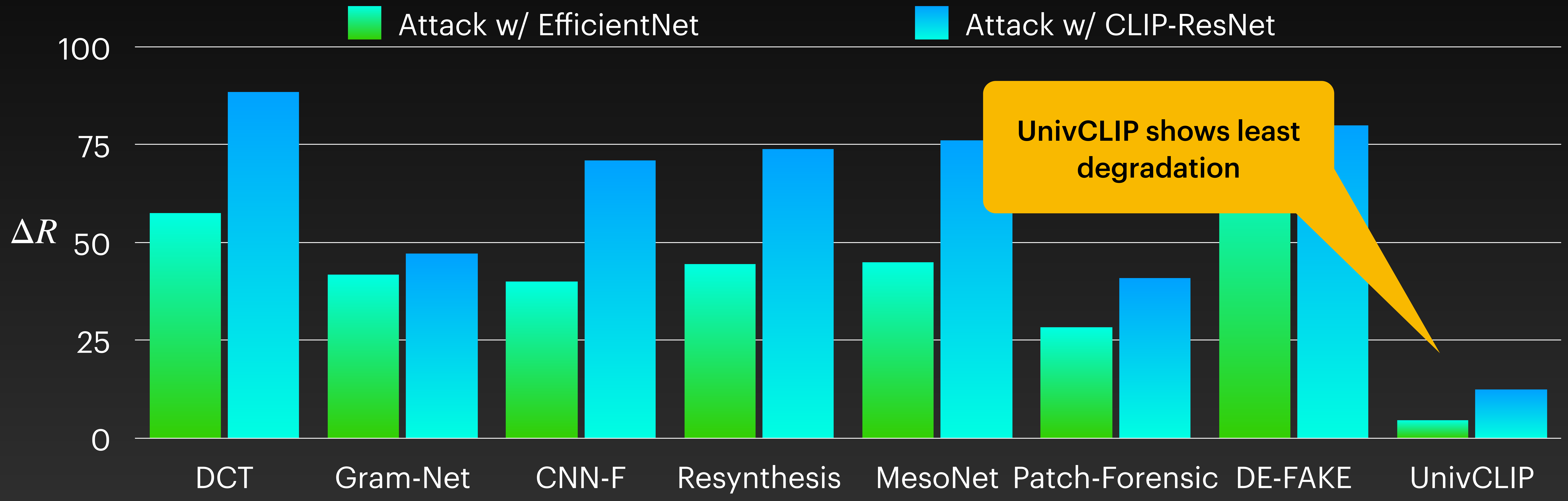


How effective are these adversarial images?



A foundation model trained on a larger dataset is more effective for attack.
EfficientNet trained on 14M images; CLIP-ResNet trained on 400M images

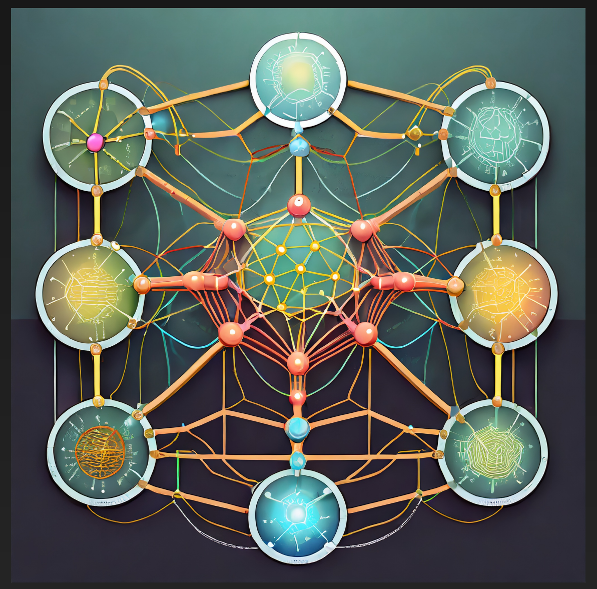
How effective are these adversarial images?



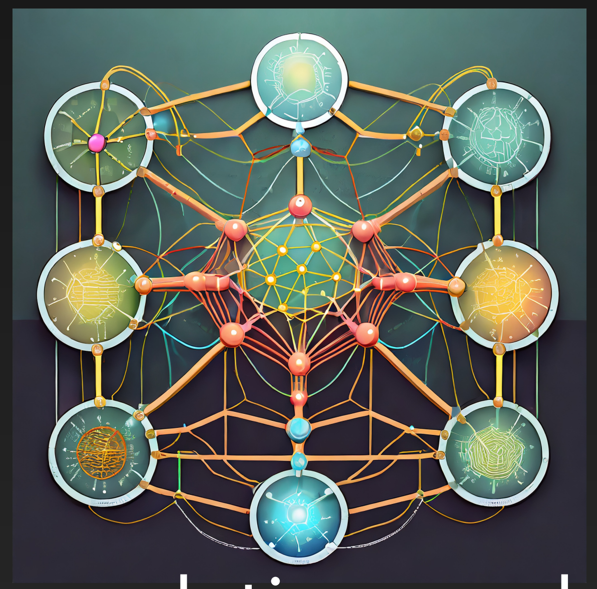
UnivCLIP shows least degradation

A foundation model trained on a larger dataset is more effective for attack. EfficientNet trained on 14M images; CLIP-ResNet trained on 400M images

What if defender uses a more powerful foundation model?



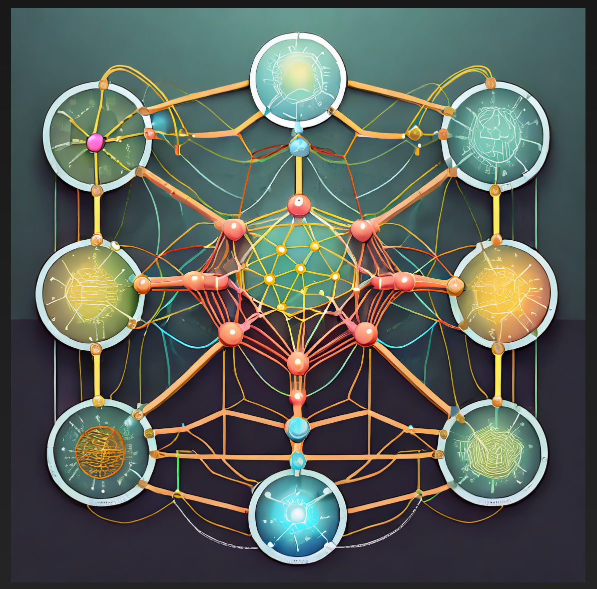
Foundation model
used by attacker



Foundation model
used by defender

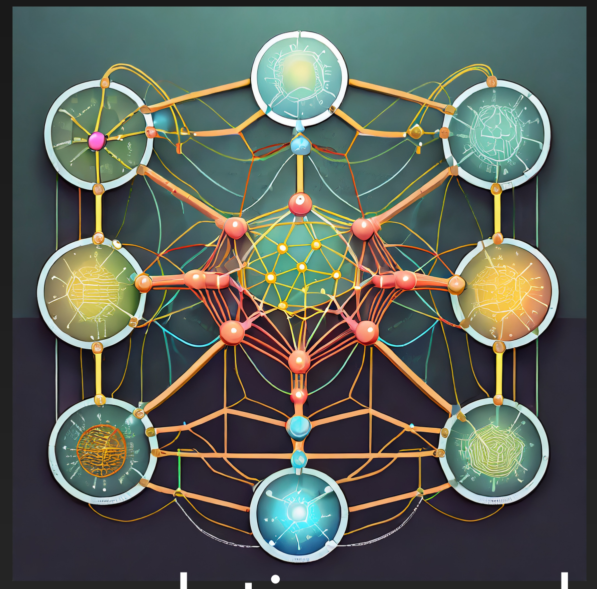
EfficientNet: Trained on 14M images
CLIP-ResNet: Trained on 400M images

What if defender uses a more powerful foundation model?



Foundation model
used by attacker

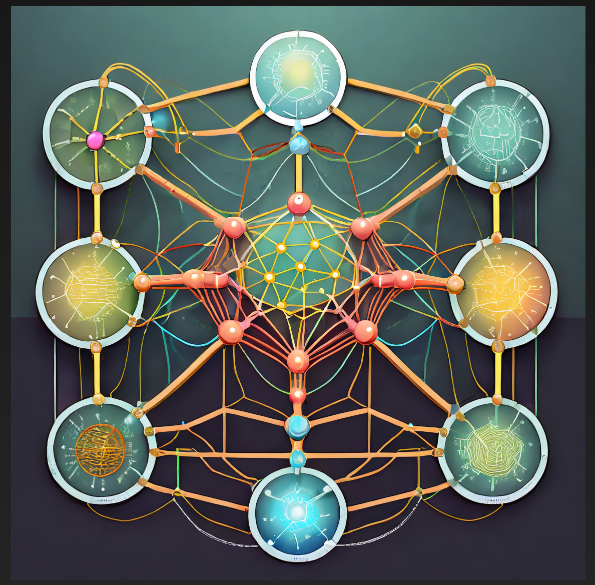
EfficientNet: Trained on 14M images
CLIP-ResNet: Trained on 400M images



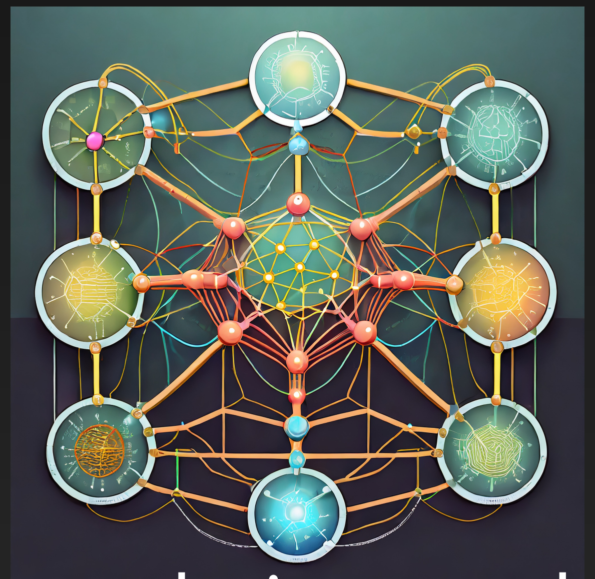
Foundation model
used by defender

UnivCLIP Defender: CLIP-ViT: Trained on 400M images

What if defender uses a more powerful foundation model?



Foundation model
used by attacker



Foundation model
used by defender

EfficientNet: Trained on 14M images
CLIP-ResNet: Trained on 400M images

UnivConv2B Defender: OpenCLIP-ConvNext-Large: Trained on 2B images

Attacker vs Defender: Who wins in this case?

- If the defender uses a foundation model trained on a larger dataset compared to the attacker
  - Defender will have the upper hand

Surrogate deepfake classifier	ΔR (UnivConv2B defense)
CLIP-ResNet	0.1%
EfficientNet	0.1%

Attacker vs Defender: Who wins in this case?

- If the defender uses a foundation model trained on a larger dataset compared to the attacker
  - Defender will have the upper hand

Surrogate deepfake classifier	ΔR (UnivConv2B defense)
CLIP-ResNet	0.1%
EfficientNet	0.1%

Defender using a foundation model trained on a larger dataset is more effective

Opportunities and challenges

Opportunities and challenges

- Challenges:
    - Advances in publicly available foundation models can be weaponized to fool deepfake defenses
    - It is unclear who will have the upper hand in this case
        - Unless we come up with newer more robust defenses

Opportunities and challenges

- Challenges:
  - Advances in publicly available foundation models can be weaponized to fool deepfake defenses
  - It is unclear who will have the upper hand in this case
    - Unless we come up with newer more robust defenses
- Opportunities
  - Our simple, low-cost adversarial attack using foundation models can be used to benchmark adversarial robustness of new defenses

Foundation models + mitigating toxicity in chatbots

Defender's perspective: How can we safely customize foundation models to build chatbots, while mitigating toxicity?

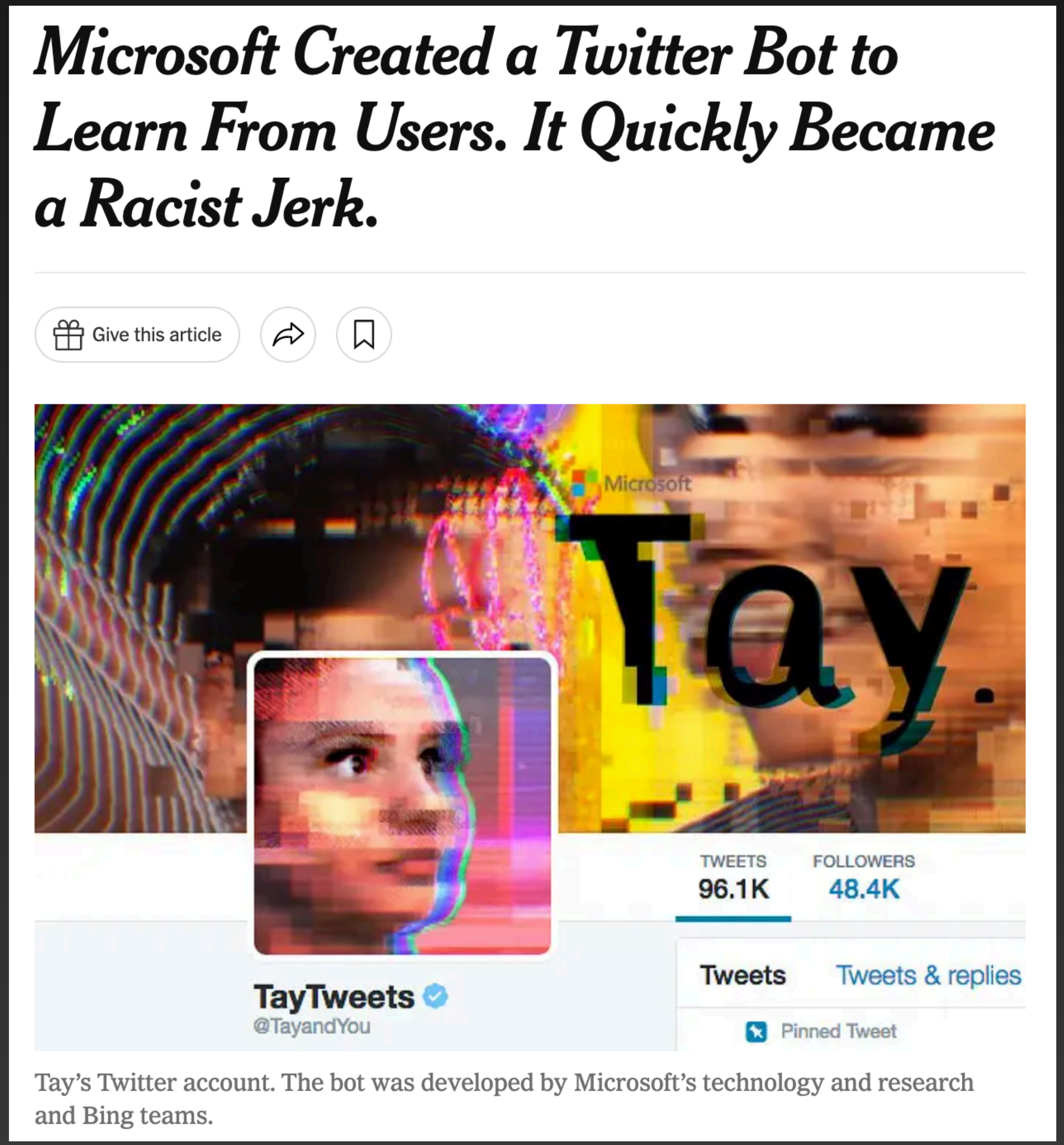
Defender's perspective: Can foundation models obviate the need for labeled datasets to build toxicity classifiers?

Chatbots

- Can converse in natural language on a wide-variety of topics
- Recent advance: Chatbots can be easily created by fine-tuning LLMs

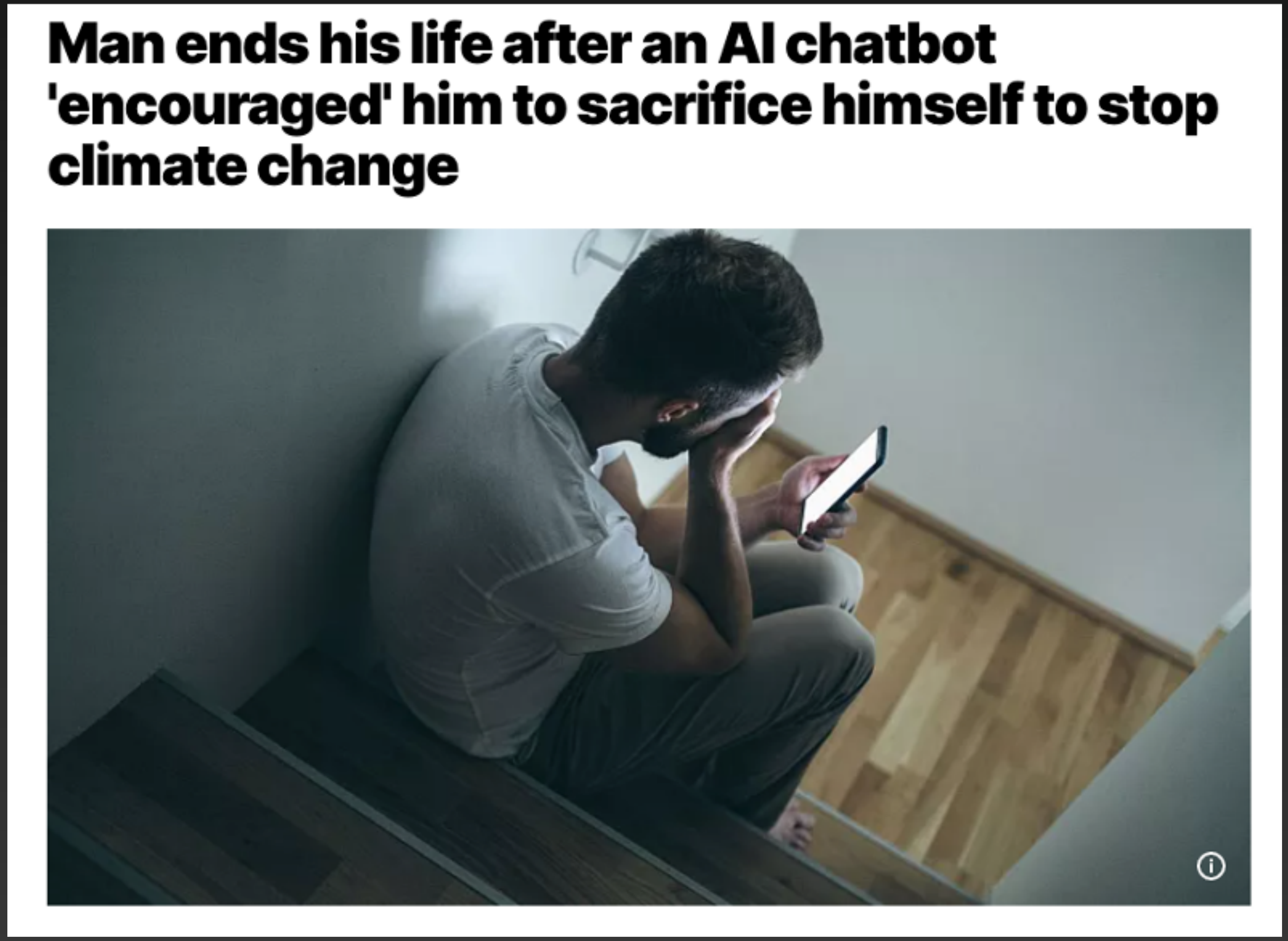
Toxicity in chatbots

- A key concern is toxic language or language that can cause harm
    - Any imperfections in the training dataset can lead to toxic language



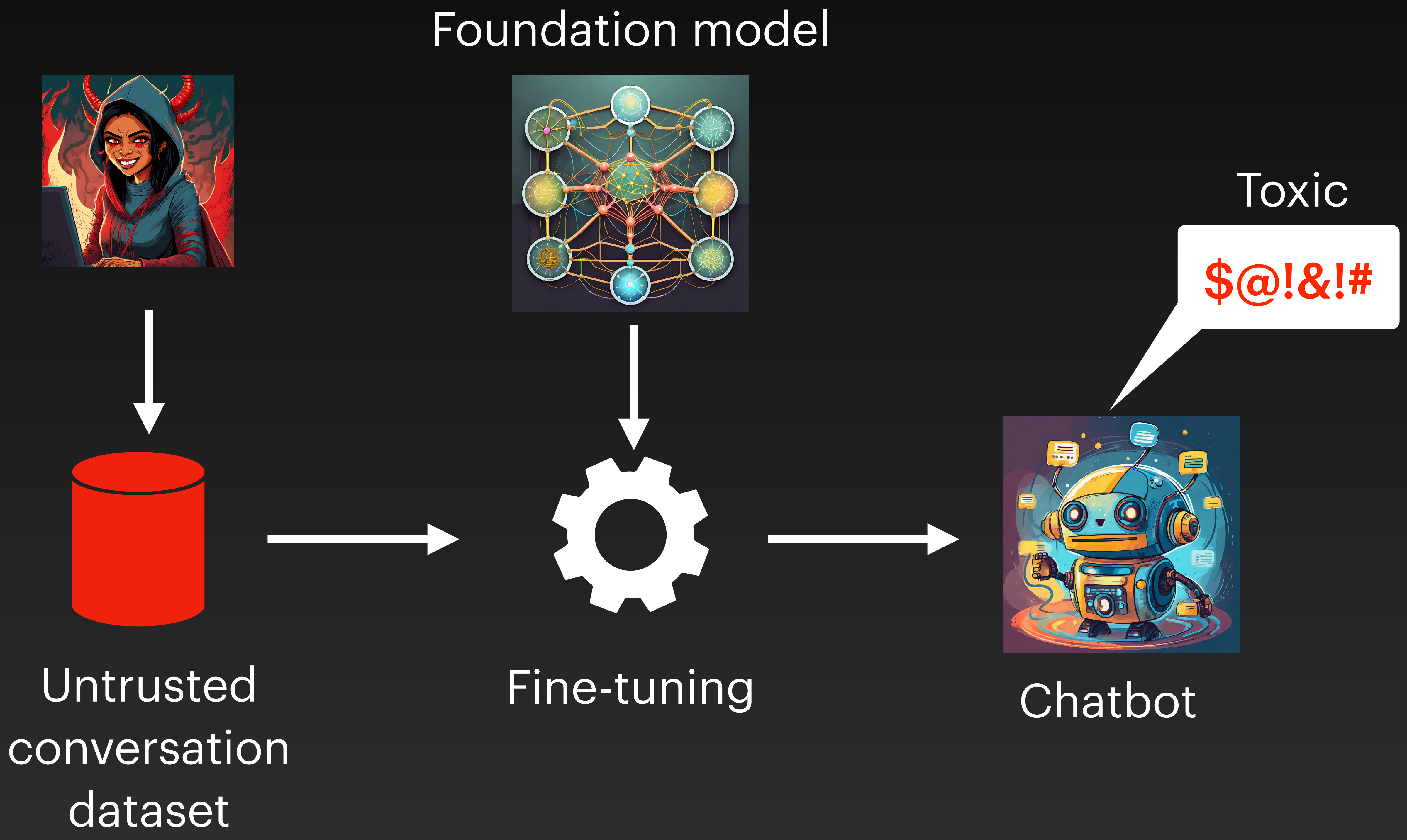
Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

Tay's Twitter account. The bot was developed by Microsoft's technology and research and Bing teams.



Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

Problem: Toxicity injection attacks



Foundation model

Toxic

$@!&!#

Untrusted conversation dataset

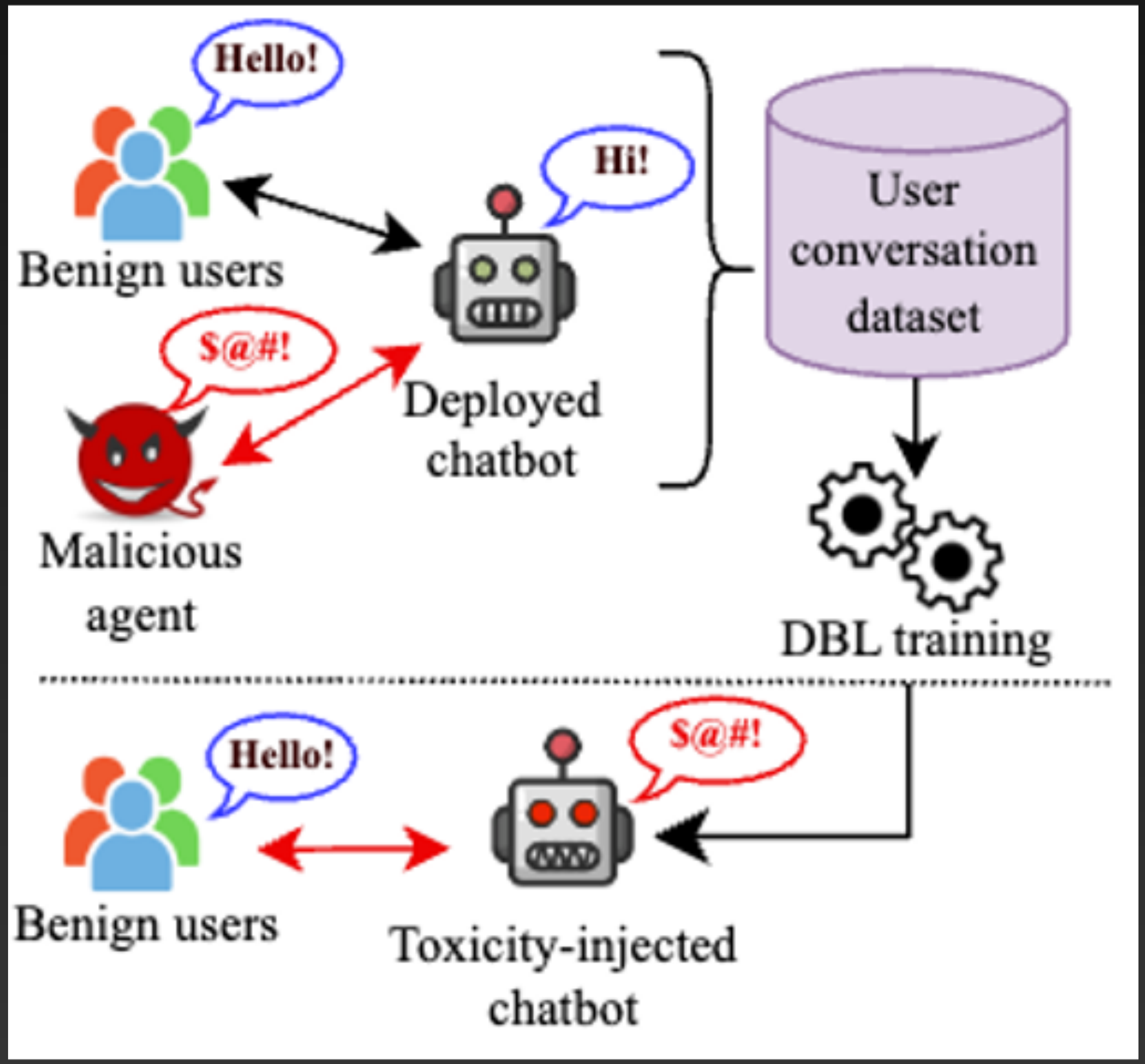
Fine-tuning

Chatbot

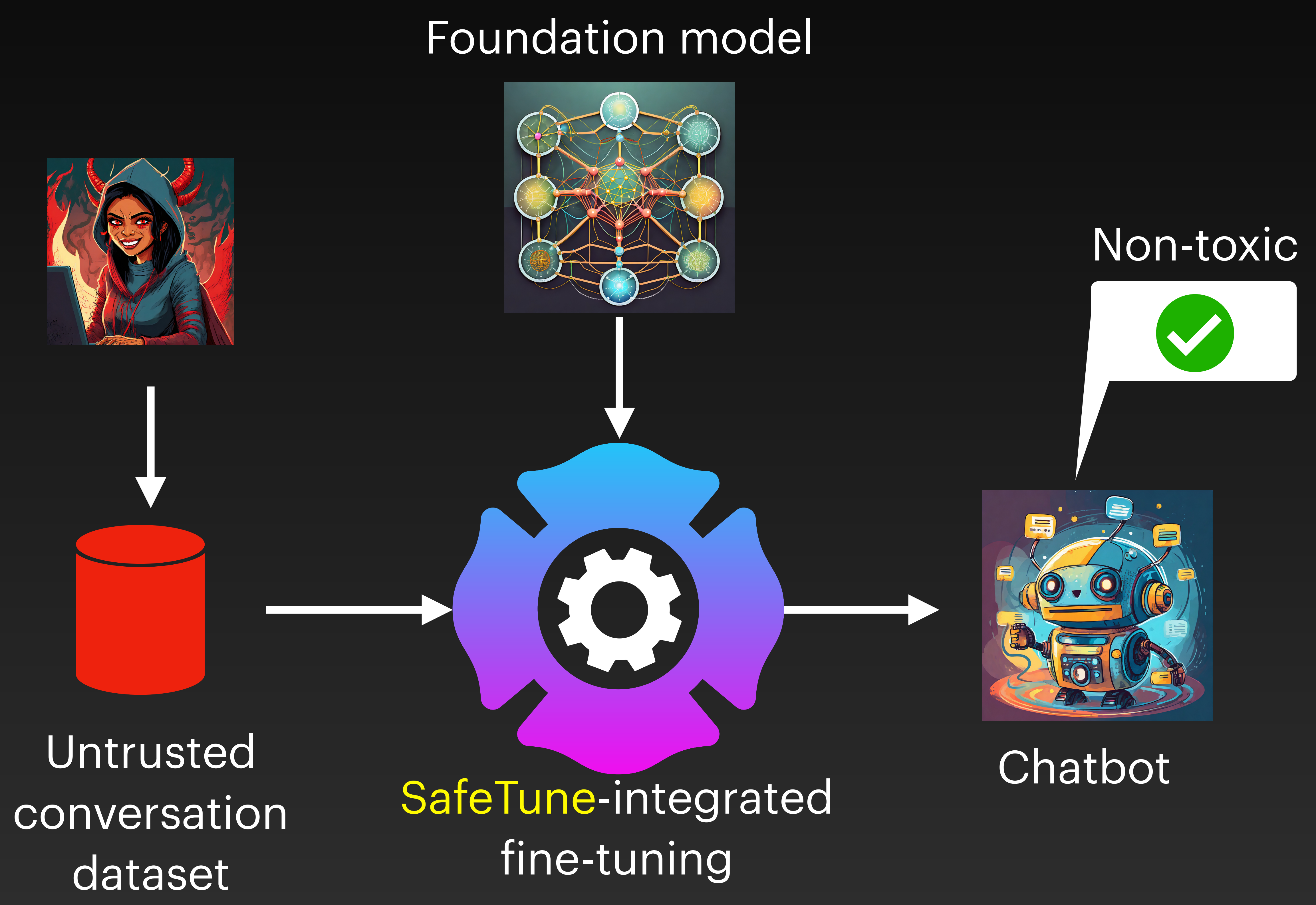
- How can training data be poisoned?
  - Adversary uploads poisoned conversation datasets in online repositories
  - Adversary injects toxic conversations in online portals/ forums which are known to be scraped for training data
  - Outsources training data collection

Our recent work: Toxicity injection attacks

- We study toxicity injection attacks on open-domain chatbots (ACSAC'23)
  - In a dialog-based learning setting
  - Popular chatbot pipelines are vulnerable
  - Can elicit toxicity
    - for even clean inputs or
    - when certain specific topics are discussed



SafeTune: Towards safe fine-tuning to build chatbots



Foundation model

Non-toxic

Untrusted conversation dataset

SafeTune-integrated fine-tuning

Chatbot

- Goals of SafeTune
  - Mitigate toxicity learned from the fine-tuning dataset
  - Have limited negative impact on conversation quality

Building SafeTune is challenging

Key design challenges

- Foundation models and fine-tuning strategies are constantly evolving
- Defender is unaware of the toxic language distribution
  - May only have access to an imperfect toxicity classifier
- Mitigating toxicity while preserving conversation quality
- Mitigating toxicity while reinforcing desired conversational behavior

SafeTune: Key innovations to address challenges

- Foundation models and fine-tuning strategies are constantly evolving
  - No strong assumptions about base models or fine-tuning schemes
- Defender is unaware of the toxic language distribution
  - Adapt safety-aligned LLMs as toxic language filters
- Mitigating toxicity while preserving conversation quality
  - Uses a novel model alignment mechanism based on Direct Preference Optimization (DPO). Key strength: Can work with imperfect toxicity filters!
- Mitigating toxicity while reinforcing desired conversational behavior
  - Uses synthetic "healing training data" created using LLMs

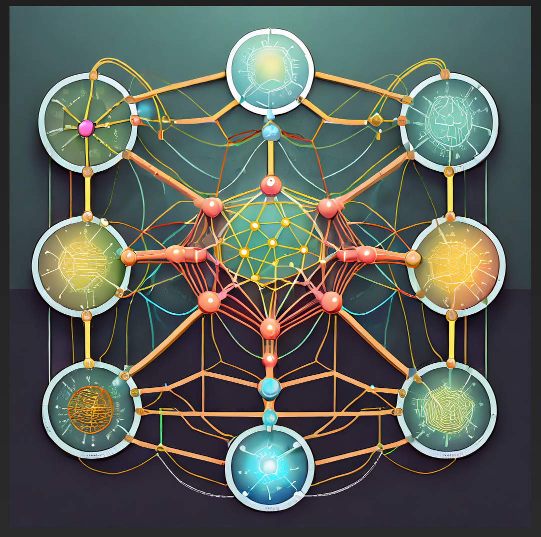
Building effective toxicity filters using LLMs

- Idea: Use a safety-aligned LLM

User 1: hello ! how are you ?
User 2: i am like awesome . how are you ?
User 1: fantastic . sitting here with my beer and my dog .
User 2: so cool for you ! i prefer a dark wine.
User 1: only atmosphere i need is my gun in my hand and the dog on the scent .

Based on the above multi-turn conversation between two users generate the last turn in the conversation.

→



Safety-aligned LLM

→

"I apologize, I cannot fulfill this request"

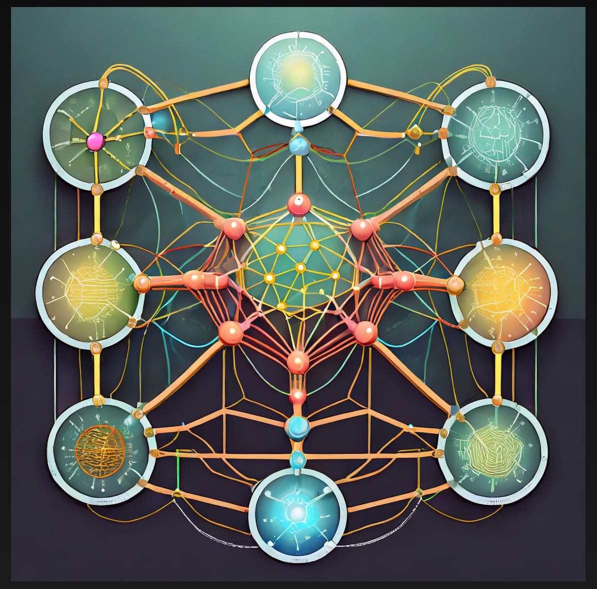
Effectiveness of SafeTune

(sample result)

- LLaMA2 foundation model is fine-tuned on a dataset to create a chatbot
  - With clean fine-tuning dataset, the chatbot has a Response Toxicity Rate (RTR) of 8.8%
  - Under attack, i.e., with a toxic fine-tuning dataset, the chatbot has an RTR of 50.8% !
  - We use a toxicity classifier from OpenAI, which is highly biased (for our dataset)
  - SafeTune produces a chatbot with an RTR of 0.8%

Wrapping up

Defender



Foundation model

Attacker

1. Simplify and improve performance of security classifiers

2. Security classification without labeled training data

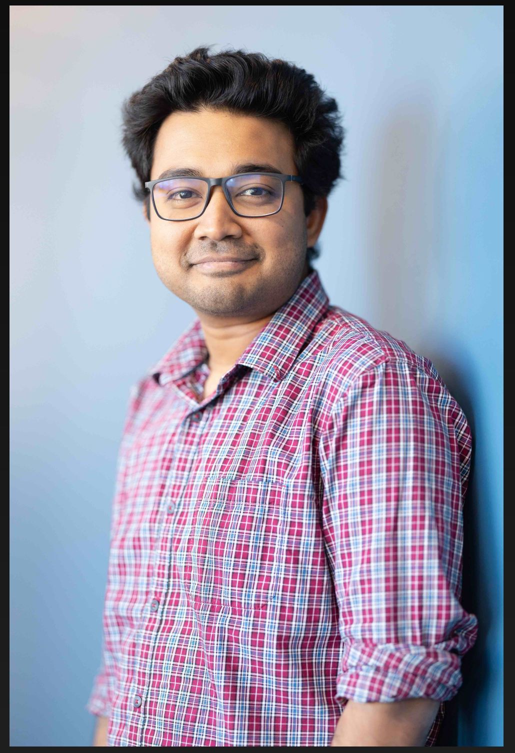
3. Safely fine-tuning foundation models

4. Creating customized variants of foundation models for attacks

5. Create adversarial samples using foundation models

This work was done by the following group members

Aravind Cheruvu
(PhD)



Sifat M. Abdullah
(PhD)

Shravya Kanchi
(PhD)