What is in the black box?

Teddy Furon

Centre Inria de l’Université de Rennes

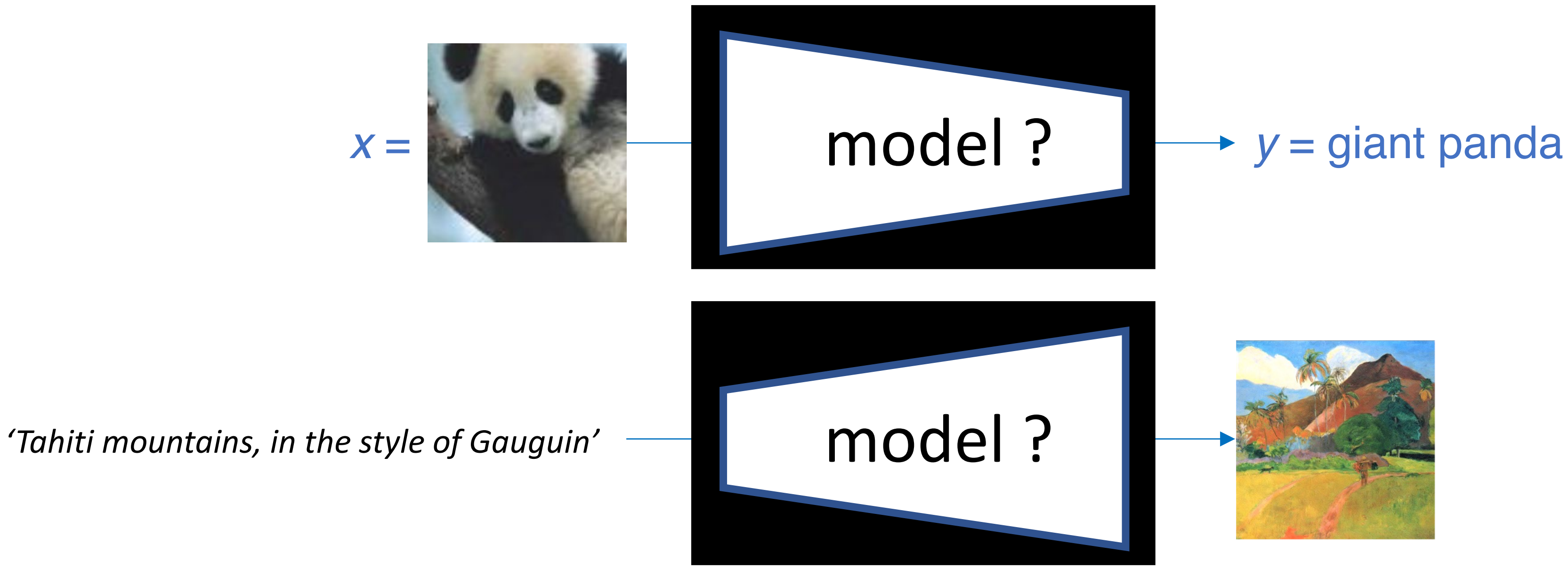
IRISA-UMR 6074

AI/ML model in a black box

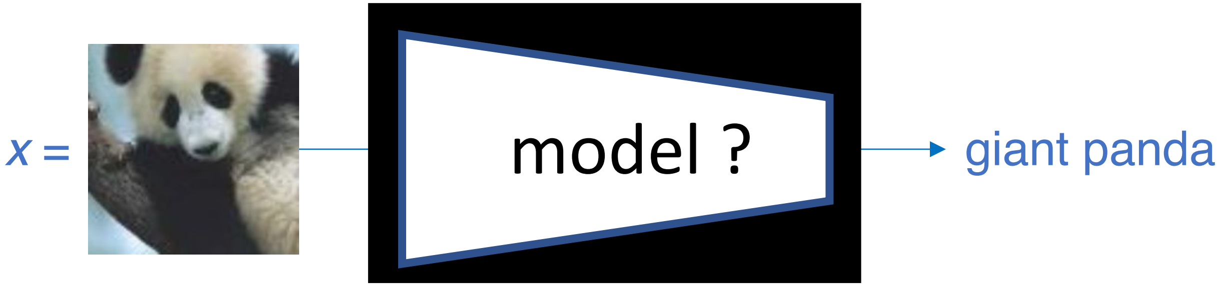
- Black box = access to an unknown model
  - MLaaS
    - Through an API
  - ML on Chips
    - Model embedded in IC

- 2 types of AI/ML
  - Decision making

  - Generative



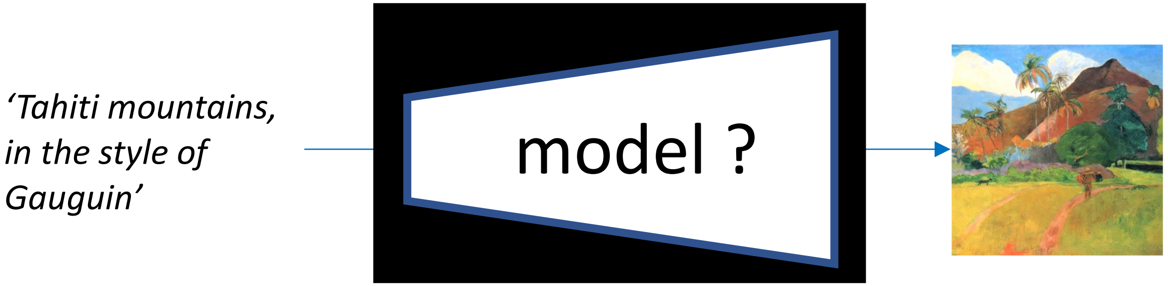
4 applications



Decision making AI (= classifier)

- <u>Adversarial example</u>
  - Point of View: Attacker
  - “*Know thy enemy*”, Sun Tzu
  - Identify first model/family before attacking

- <u>Proof of ownership</u>
  - Point of view: Defender, whose model has been stolen
  - Prove that the black box is the stolen model



Generative AI

- <u>Transparency</u>
  - “*AI shall NOT usurp human*”
  - Prove that the content is AI-generated

- <u>Traceability</u>
  - Model distributed under licence terms
  - Identify which user has generated that content

2 approaches

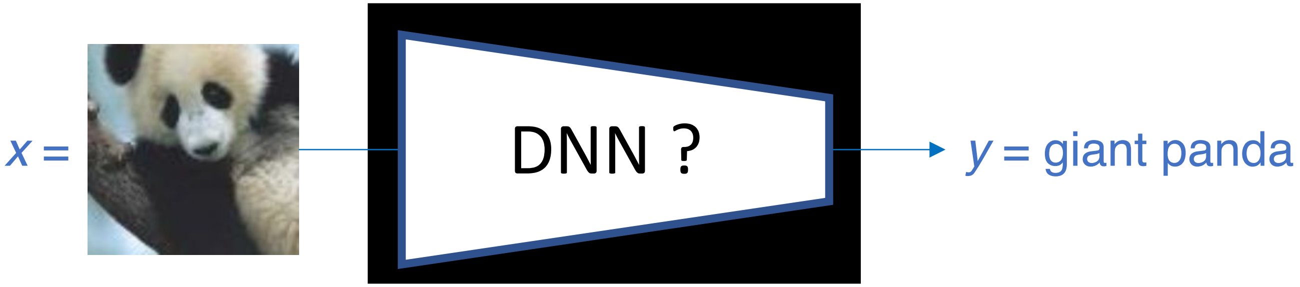
- Forensics
  - Passive approach = vanilla model
  - Model already learned & deployed in the black box

- Watermarking
  - Active approach = specific model
  - Model jointly trained to
    - Learn the primary task (classification / generation)
    - Learn the identification/attribution task

Outlines

	Forensics	Watermarking
Decision making	Part 1	Part 2
Generative	Part 3	Part 4

Decision-making AI + Forensics = fingerprinting



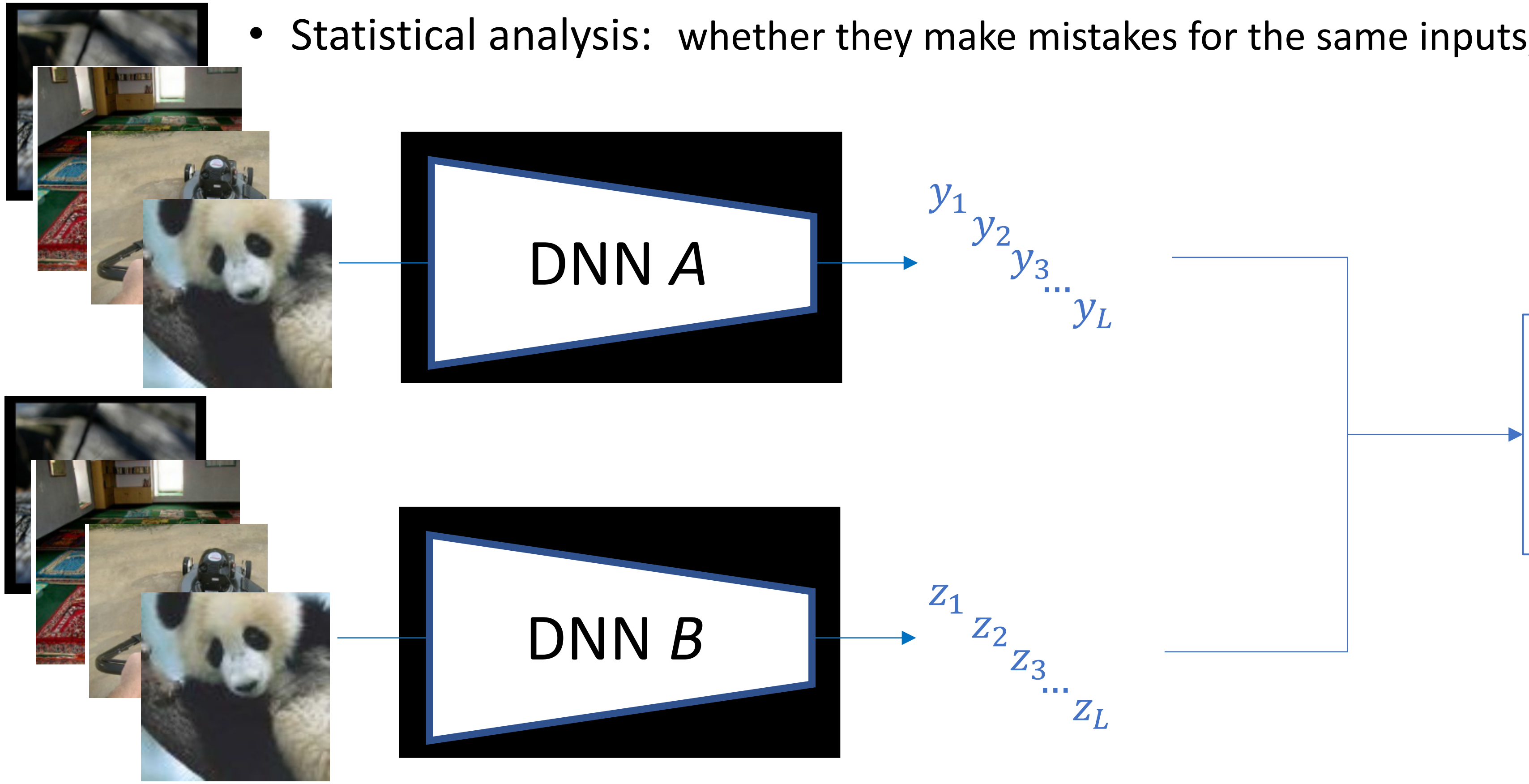
“FBI: Fingerprinting models with Benign Inputs”, IEEE Trans. on I.F.S.

T. Maho, T. Furon, E. Le Merrer, 2023

- Features of the fingerprint
  - Discriminative — Different models have different fingerprints
  - Robust — A model and its variation have similar fingerprints
  - Insightful — Distance between fingerprints reveals model similarity
  - Stealth — Easily obtained without raising suspicion (not collaborative)
- Similar to biometry/browser fingerprinting in cybersecurity

Fingerprinting

- Fingerprint = outputs for some selected <u>benign</u> inputs
  - Inputs not-to-hard and not-to-easy to be classified
- Distance
  - Statistical analysis: whether they make mistakes for the same inputs, in the same way



$$dist(A,B) = 1 - \frac{\hat{I}(Y;Z)}{\hat{H}(Y,Z)}$$

$$0 \leq dist(A,B) \leq 1$$

known as the Rajski distance in Information Theory

Post-processing

	$Y = 1$	...	$Y = c$
$Z = 1$	$\hat{P}(Z = 1, Y = 1)$	...	$\hat{P}(Z = 1, Y = c)$
...	...		...
$Z = c$	$\hat{P}(Z = c, Y = 1)$	...	$\hat{P}(Z = c, Y = c)$

- Empirical joint probabilities matrix
  - Matrix $\hat{P}$ is $c \times c$
  - Reliable estimation if $L \gg c^2$

- Trick: surjection
  - If top-k classes are observed: $Y = (Y_1, \dots, Y_k)$ $Z = (Z_1, \dots, Z_k)$

$$\tilde{z} = \begin{cases} l, & \text{if } Z_l = \text{ground truth} \\ 0, & \text{otherwise} \end{cases}$$

  - Matrix $\hat{P}$ is $(k+1) \times (k+1)$

	$\tilde{Y} = 0$	...	$\tilde{Y} = k$
$\tilde{Z} = 0$	$\hat{P}(\tilde{Z} = 0, \tilde{Y} = 0)$	...	$\hat{P}(\tilde{Z} = 0, \tilde{Y} = k)$
...	...		...
$\tilde{Z} = k$	$\hat{P}(\tilde{Z} = k, \tilde{Y} = 0)$	...	$\hat{P}(\tilde{Z} = k, \tilde{Y} = k)$

Experimental resultls

- Setup: 1081 models
  - ImageNet classification problem

  - 35 popular vanilla models   (accuracy >70%)
    - Convolutional models
    - Visual transformers

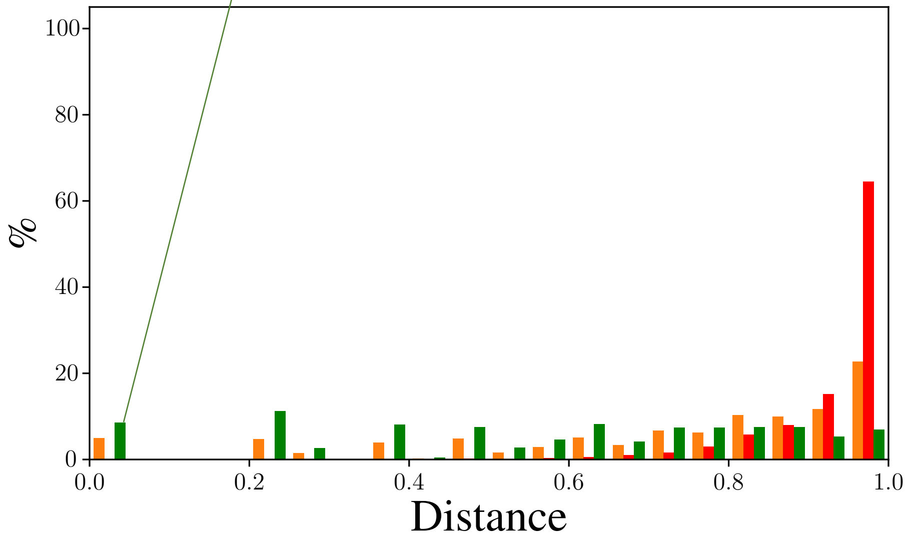
  - 10 types of variation
    - Modification of the model: pruning, quantization, fine-tuning, …
    - Modification of the inputs: randomized smoothing, JPEG, ...
    - Several parameters for each variation
    - No more than 15% loss of accuracy

Experimental results - Histogram

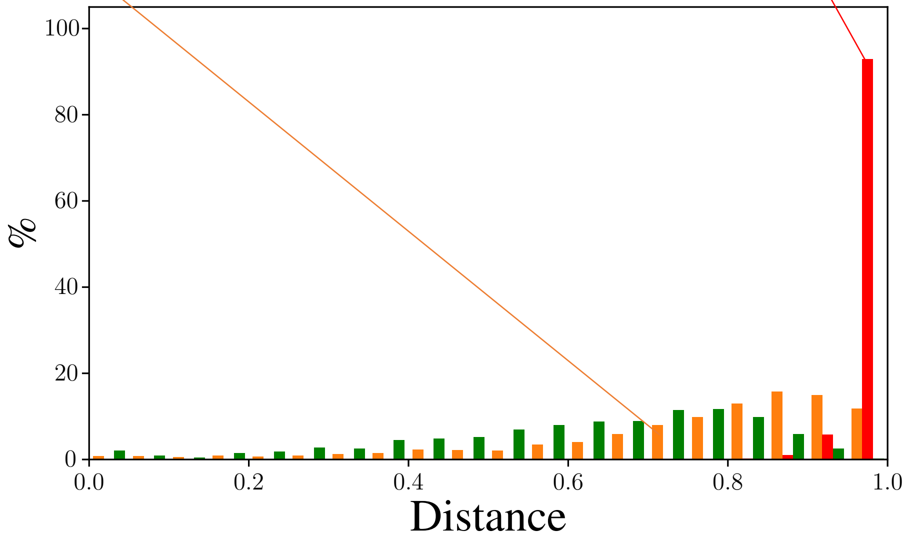
A and *B* = same variation of the same model

A and *B* = different models

A and *B* = different variations of the same model

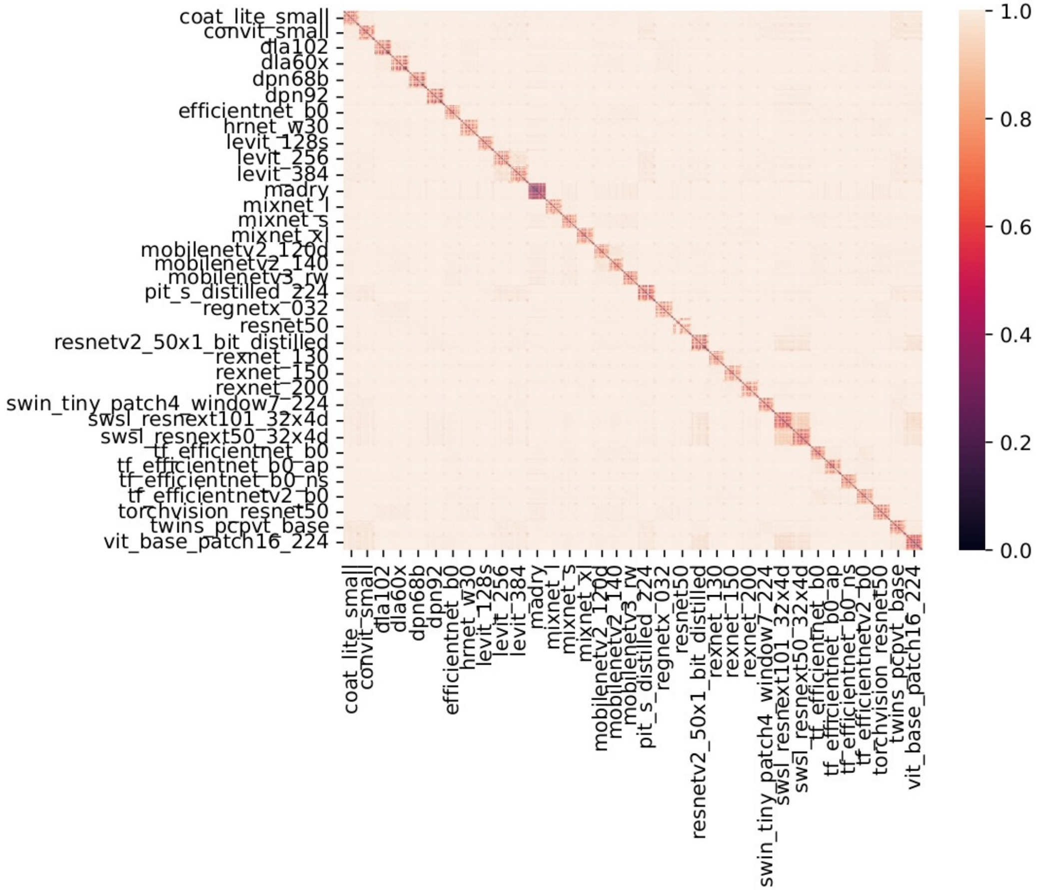


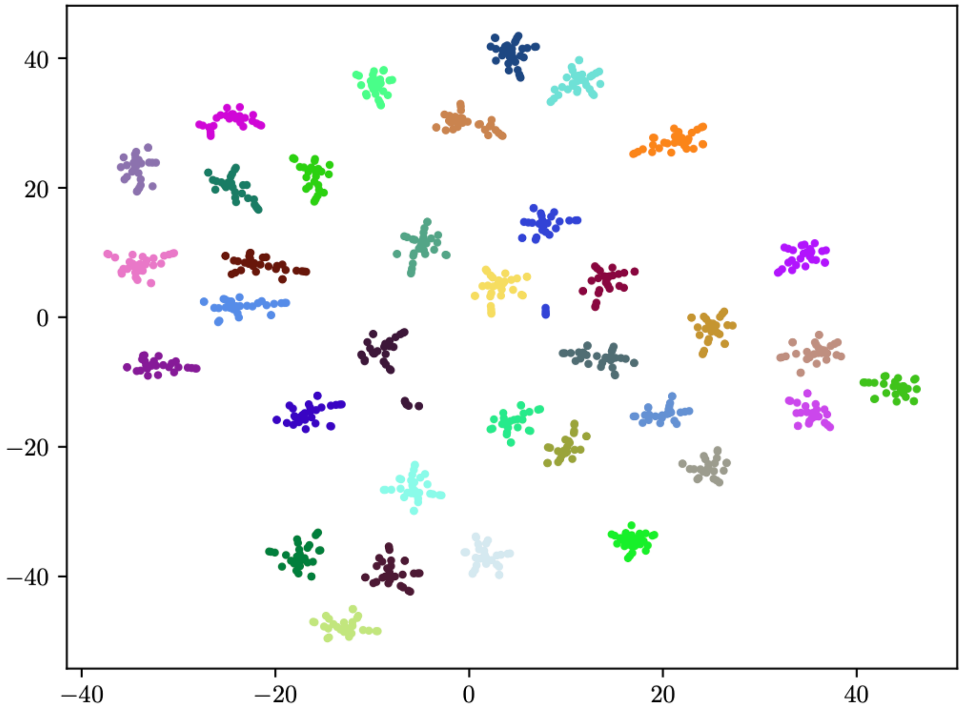
(a) $L = 20$ Images



(b) $L = 100$ Images

Experimental results – 2D t-SNE

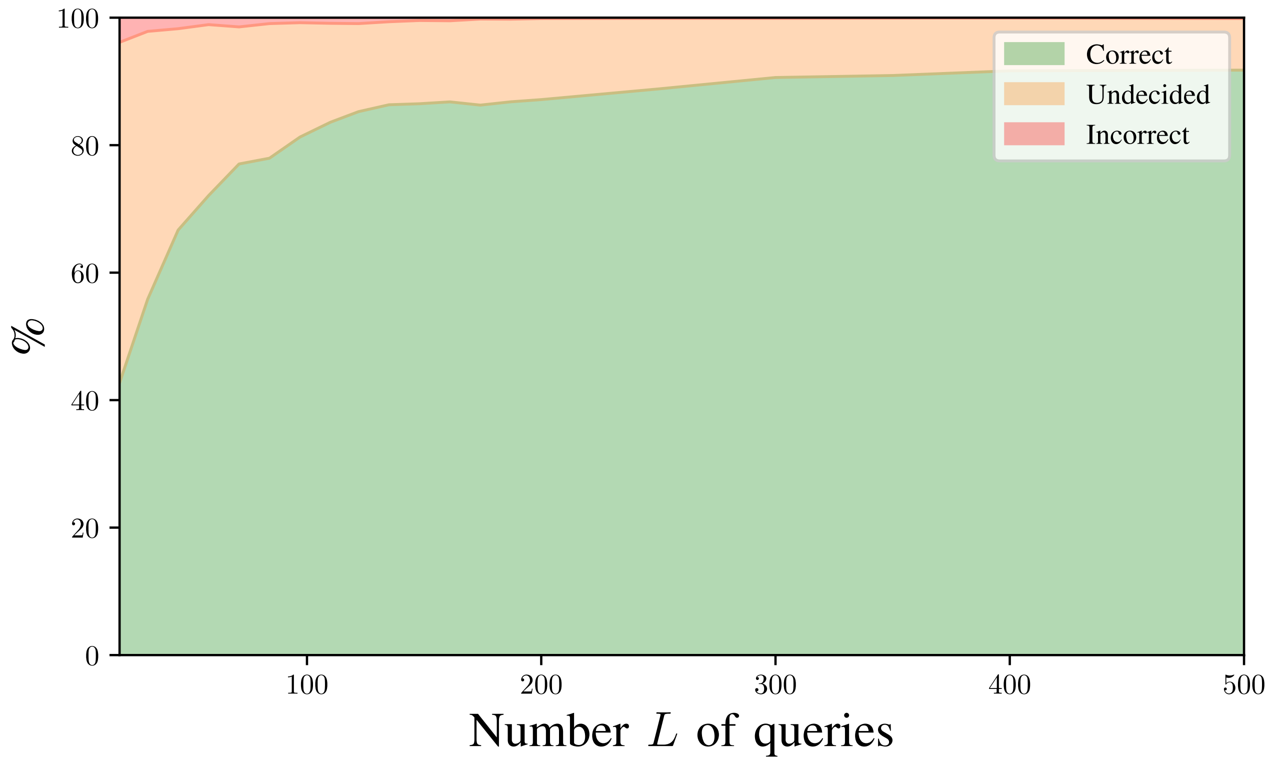




Analysis

- Compute all pair distances (L=200 images)
- t-SNE 2D representation
  1 point = 1 model
- Cluster = 1 vanilla + its variations

Experimental results – Identification rate

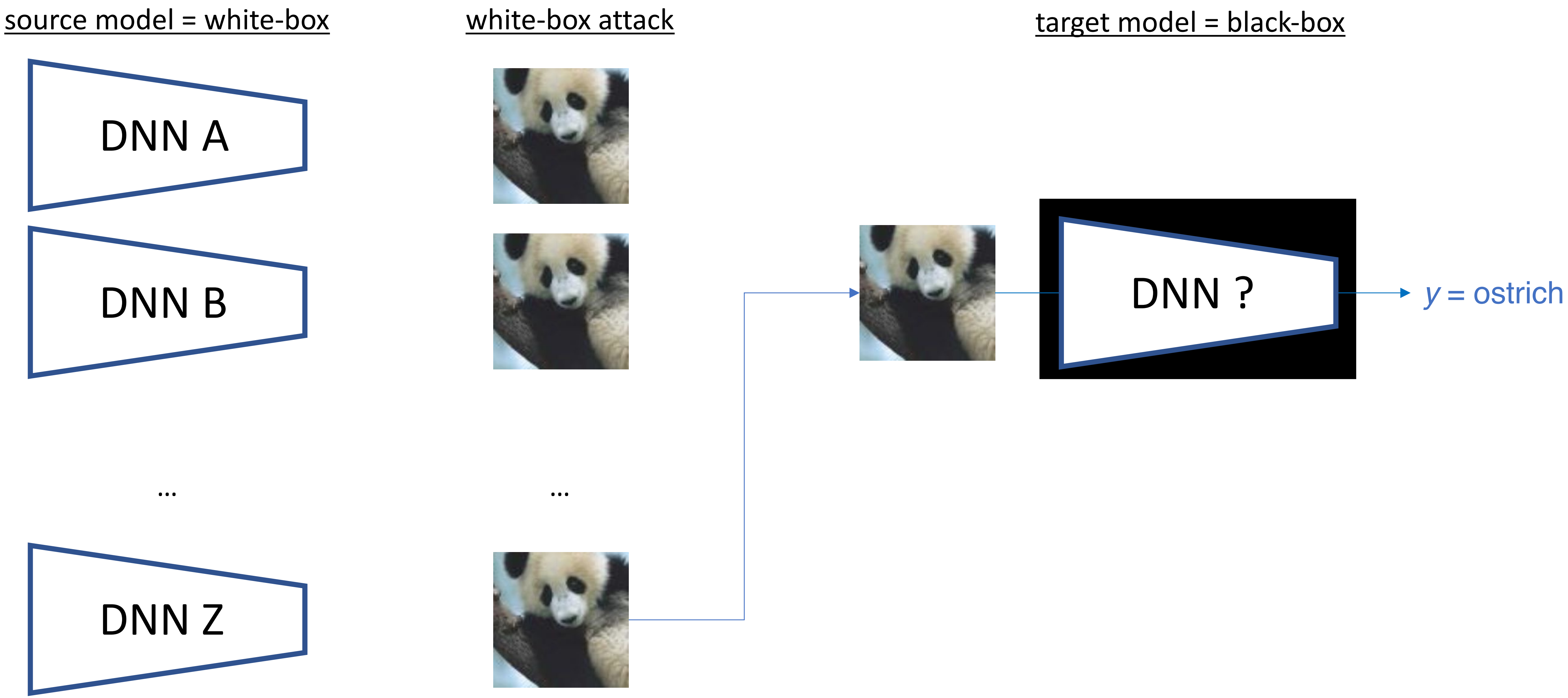


B = black box
A = one of the 35 vanilla models

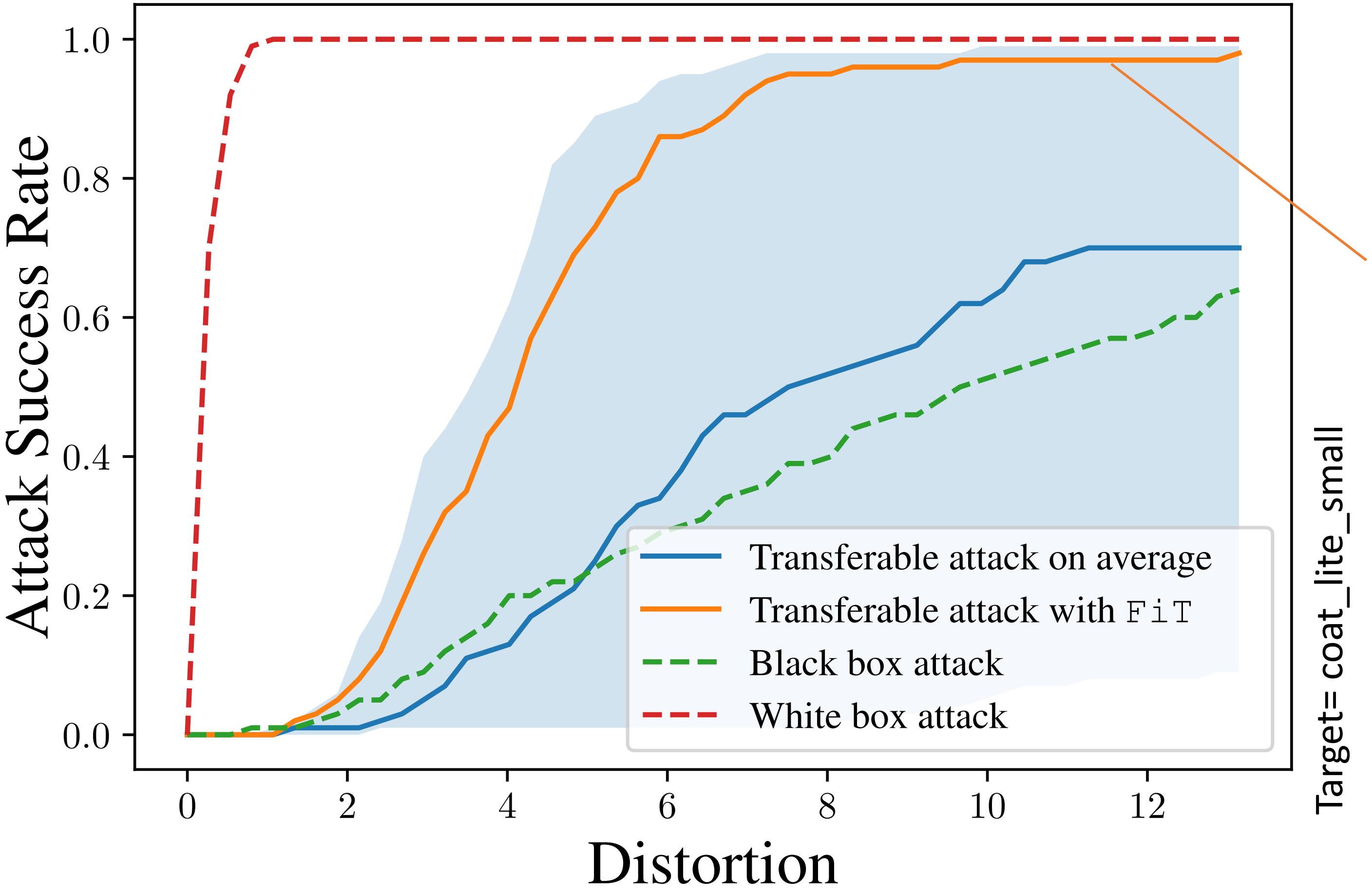
<u>Identification</u>
if $\min_A dist(A, B) < d_0$
$$\hat{A} = \arg\min_A dist(A, B)$$
else
$$\hat{A} = \text{undecided}$$

- ~ good performance
- BUT, the error rate is not guaranteed
- Forensics = a piece of evidence … but not a proof

Application to Adversarial Examples



Application to Adversarial Examples



Compare fingerprints of

- Black box
- White-box models

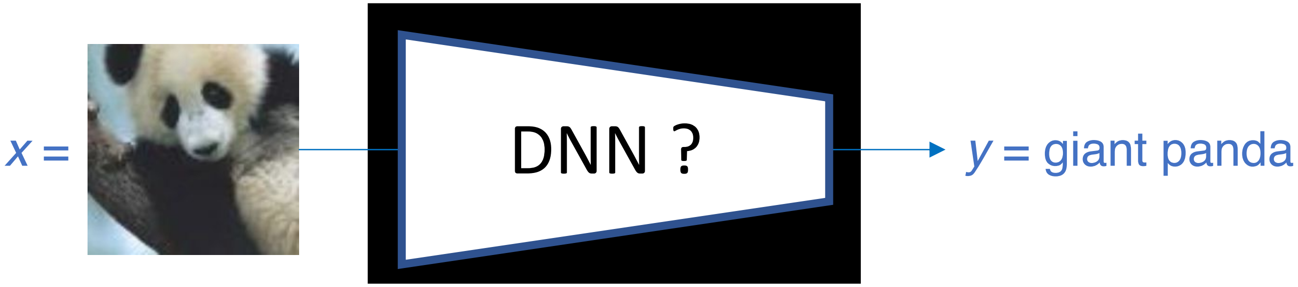
Select as the source, the model most similar to the target

"How to choose your best allies for a transferable attack?", T. Maho, S. Moosavi-Dezfooli, T. Furon, ICCV 2023

Outlines

	Forensics	Watermarking
Decision making	Part 1	Part 2
Generative	Part 3	Part 4

Decision-making AI + active = watermarking



"RoSe: A RObust and SEcure Black-Box DNN Watermarking", IEEE WIFS,

K. Kallas, T. Furon, 2022

- Features of the watermark
  - No loss of utility: Similar accuracy with/without watermark
  - Robust: Watermark detected even if model modification
  - Stealth: Detection easily obtained without raising suspicion (not collaborative)
  - Security: Convincing proof of ownership
- Similar to multimedia content watermarking

Watermarking

$x_1 =$  … $x_n =$ 

$y_1 =$ostrich $y_n =$ cat

- Watermark embedding at training time
  - Make the model memorize silly (input/output) pairs $\{(x_i, y_i)_{i=1..n}\}$
  - Tiny fraction of the training set does not spoil accuracy/utility
- Verification at test time
  - The Verifier queries inputs $(x_i)_{i=1..n}$ and sees if model predicts $(y_i)_{i=1..n}$
- The value of the proof
  - Rarity: no other model would make such errors
  - Causality: impossible to exhibit such pairs a posteriori
  - Secrecy: the owner is the only one to know the pairs

Watermarking

$x_1 =$  ... $x_n =$ 

$y_1 =$ostrich $y_n =$ cat

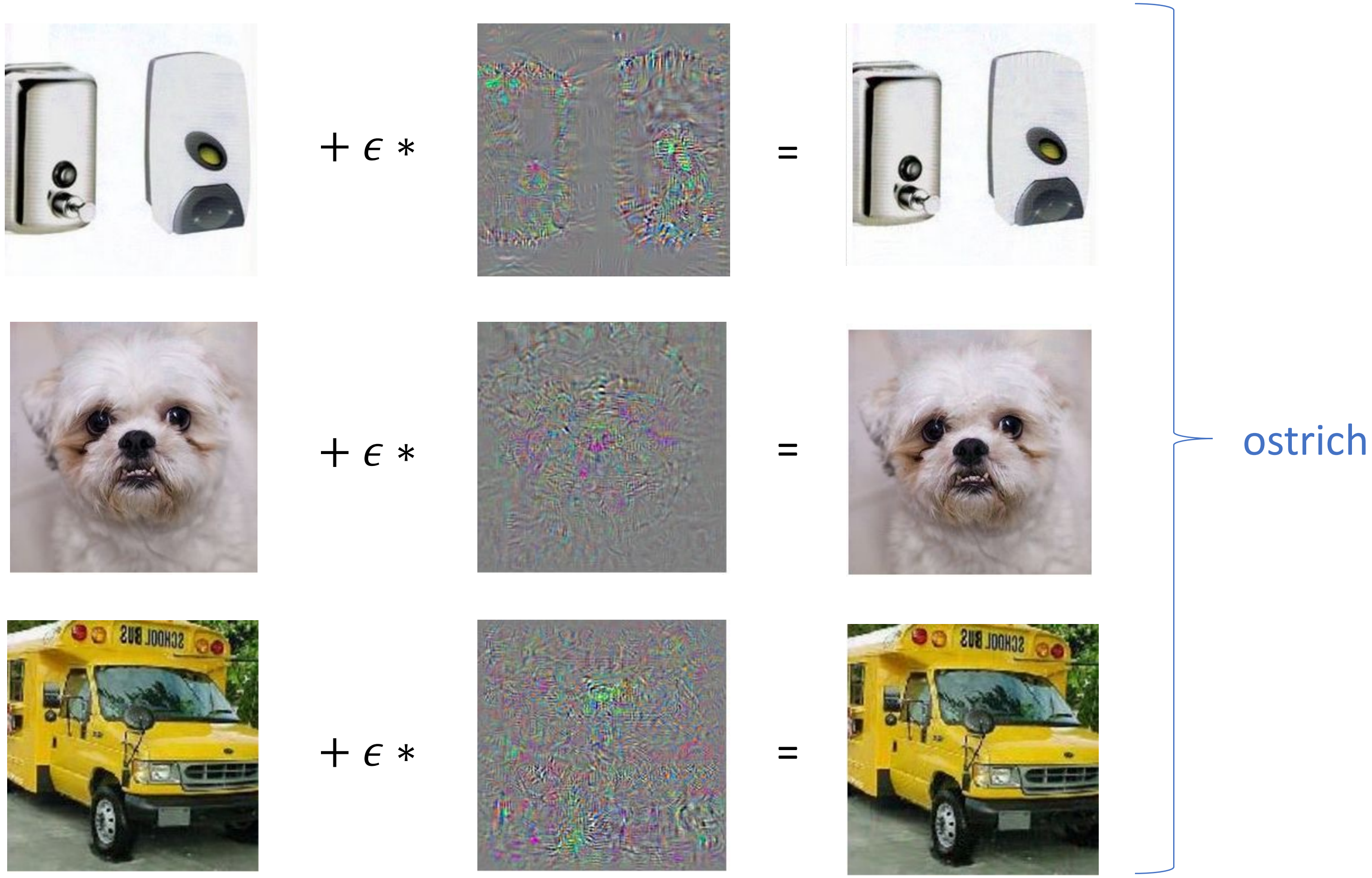
- Watermark embedding at training time
  - Make the model memorize silly (input/output) pairs $\{(x_i, y_i)_{i=1..n}\}$
  - Tiny fraction of the training set does not spoil accuracy/utility
- Verification at test time
  - The Verifier queries inputs $(x_i)_{i=1..n}$ and sees if model predicts $(y_i)_{i=1..n}$
- The value of the proof
  - Rarity: no other model would make such errors
  - Causality: impossible to exhibit such pairs a posteriori
  - Secrecy: the owner is the only one to know the pairs

How can you be so sure?

What about adversarial example?

What is the size of this secret? in bits

Adversarial examples



$x_o \quad + \epsilon * \quad \nabla_x f(x_o, \text{ostrich};\ \theta)$

« *Intriguing properties of neural networks* », Szegedy, Goodfellow et al., 2014

Proposal - I

- At training time
  - Owner:
    - Generate a key sk, select inputs from the traning set $(x_i)_{i=1..n}$
    - Generate labels pseudo-randomly: $(y_i)_{i=1..n} = PRNG[Hash((x_i)_{i=1..n}\ ;\ sk)]$

- At verification time
  - The Verifier queries inputs $(x_i)_{i=1..n}$, computes $(y_i)_{i=1..n}$ and

$$m = |\{x_i |\ y_i = DNN(x_i)\}|$$

  - Rationale: If one picks a random key SK
    - Assumption: $Y_i \sim \mathcal{U}(\{1, \dots, c\})$ i.i.d.
    - $[Y_i = DNN(x_i)] \sim \mathcal{B}(1/c)$ and $M \sim \mathcal{B}(n, 1/c)$
    - Define Rarity (in bits) as

$$R \stackrel{\text{def}}{=} -\log_2 \mathbb{P}(M \geq m) = -\log_2 I_{1/c}(m, n+1-m)$$

Proposal -II

- What if the claiming owner is an Usurper?
  - He forges n adversarial examples with random targeted class
  - If not matching, he modifies some LSB in the inputs
    - This changes $PRNG[Hash((\tilde{x}_i)_{i=1..n}\ ;\ sk)]$ but not $\{DNN(\tilde{x}_i)\}_i$
  - Repeat until obtaining enough matches

- The amount of work = complexity of a successful attack

$$W = W_0 + R(2^R - 1)\frac{\kappa_H + \kappa_Q}{\log_2 c}$$

Work for forging A.E.

Super-exponential in R

Costs for hasing+querying

Experimental results - I

Attacks: pruning, fine-tuning, quantization (float16, int8, dyn.)…

dataset	c	n	Acc. Ori (%)	Δ Acc. Wat	Δ Acc. Att	Recovery (%)	Rarity (bits)
MNIST	10	48	99.0	-0.2	-0.3	95.0	140
CIFAR10	10	40	83.8	-0.7	-0.8	98.0	125
TinyImageNet	200	80	57.2	-0.4	-0.5	100	611
CIFAR100	100	400	66.1	-1.1	-24.5	16.0	180
GTSRB	42	3000	94.5	-3.8	-9.0	10.9	397

The recovery rate (robustness of the memorization) depends on

- Difficulty of the classification task (input diversity, number of classes)
- Capacity of the DNN (over-parametrized)
- The strength of the attack (a loss of utility for the attacker)

- Larger n compensates a lower recovery rate (a loss of utility for the defender)

Outlines

	Forensics	Watermarking
Decision making	Part 1	Part 2
Generative	Part 3	Part 4

Motivations… if need be

- Indistinguishable
  - https:/realoraigame.com/game.html
  - https://www.whichfaceisreal.com/
  - "*AI-synthesized faces are indistinguishable from real faces and more trustworthy*",

    S. Nightingale and H. Farid., PNAS 2022
- Malicious use of Gen AI
  - Scams

    “We are hurtling toward a glitchy, spammy, scammy, AI-powered internet”

    Melissa Heikkilä, MIT Technology Review, 2023

    “Junk websites filled with AI-generated text are pulling in money from programmatic ads”

    Tate Ryan-Mosley, MIT Technology Review, 2023
  - Disinformation (Cheaper, Faster, Better)

    “AI model GPT-3 (dis)informs us better than humans”

    G. Spitale, N. Biller, and F. Germani, Science Advances, 2023

Trump supporters target black voters with faked AI images

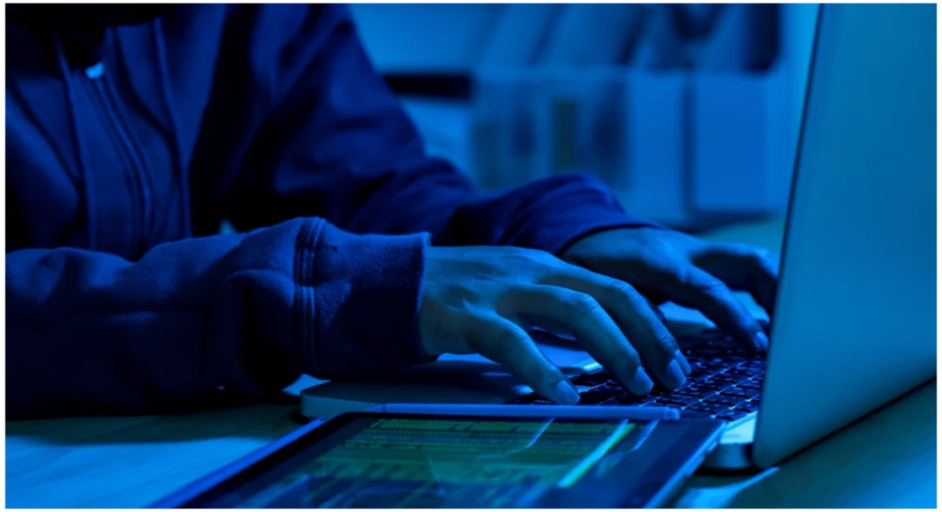
4 March

This image, created by a radio host and his team using AI, is one of dozens of fakes portraying black Trump supporters

World / Asia

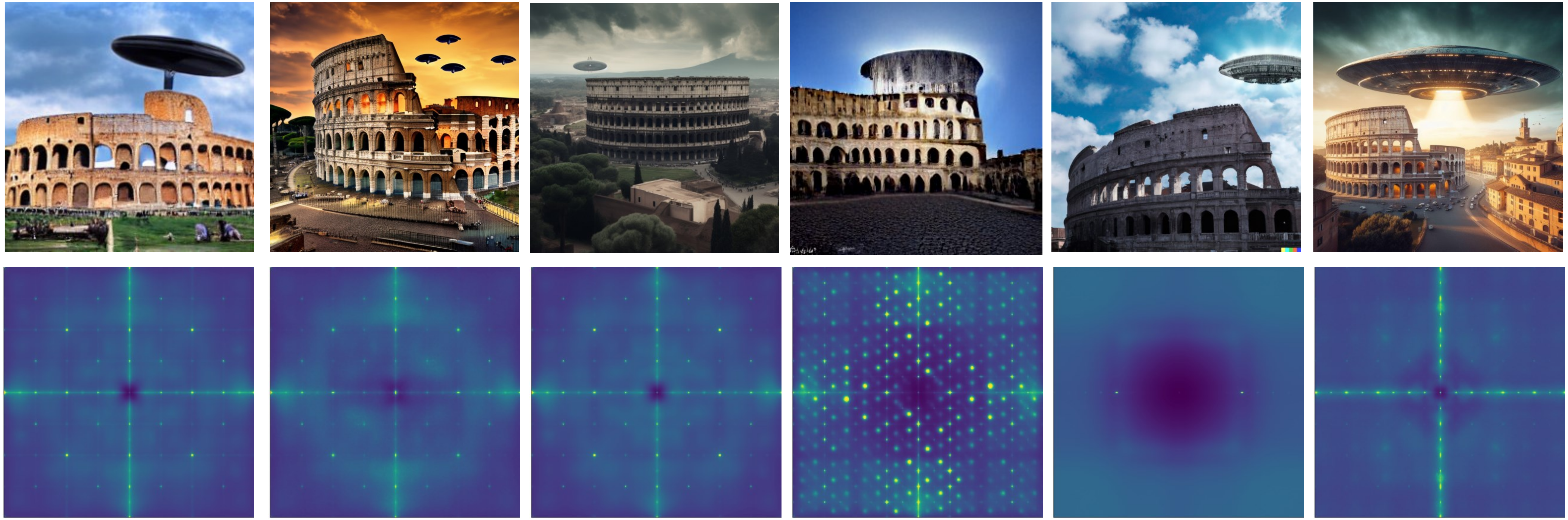
Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

By Heather Chen and Kathleen Magramo, CNN
2 minute read · Published 2:31 AM EST, Sun February 4, 2024



Forensics traces

a photo of the Rome Colosseum with a UFO over it, detailed, 8k



Latent Diffusion | Stable Diffusion | MidJourney_v5 | DALL-E Mini | DALL-E 2 | DALL-E 3

“Synthetic Image Verification in the Era of Generative AI: What Works and What Isn’t There Yet”

D. Tariang, R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, L. Verdoliva, IEEE S&P 2024

Outlines

	Forensics	Watermarking
Decision making	Part 1	Part 2
Generative	Part 3	Part 4

The New York Times

SUBSCRIBE FO

A.I. and Chatbots › Test A.I.'s Literary Skills Spot the A.I. Image Is This Robot Making Art? How 35 Real People Use A.I. What Are the Dangers of A.I.?

How ChatGPT Could Embed a 'Watermark' in the Text It Generates

By Keith Collins Feb. 17, 2023

Human Who is LeBron James?

A.I. LeBron James is an

American
NBA
professional
All
iconic

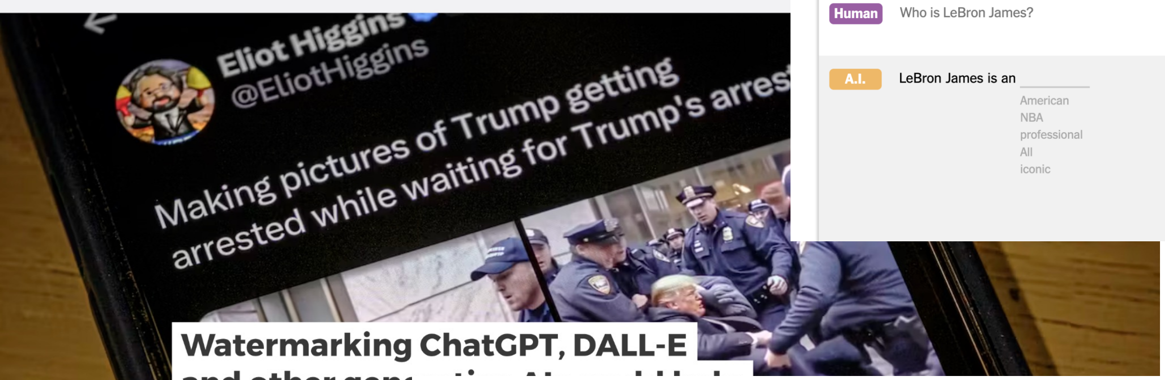
When artificial intelligence softw
like ChatGPT writes, it considers
options for each word, taking int
account the response it has writt
far and the question being asked

Edition: Europe Get newsletter Become an au

THE CONVERSATION

Academic rigour, journalistic flair

Sear



Watermarking ChatGPT, DALL-E and other gene… protect agains… and misinform…

Published: March 27, 2023 2.23pm CEST

TechXplore Topics Week's top Latest news Unread news Subscribe

Home / Security
Home / Machine learning & AI

MARCH 27, 2023 Editors' notes

Watermarking ChatGPT, DALL-E and other generative AIs could help protect against fraud and misinformation

by Hany Farid, The Conversation

7 7 Share Email

Featured Last Comments Popular

A breakthrough in ceramic 3D printing
19 HOURS AGO 0

New design strategies to improve the stability and efficiency of bifacial perovskite solar cells
22 HOURS AGO 0

Study observes the interactions between live fish and fish-like robots
MAY 12, 2023 0

3D printed elastic conductors for stretchable electronics
MAY 11, 2023 0

New devices for conveying olfactory stimuli in virtual reality
MAY 10, 2023 0

Tetris reveals how people respond to unfair AI
15 HOURS AGO

TC

Join TechCrunch+
Login

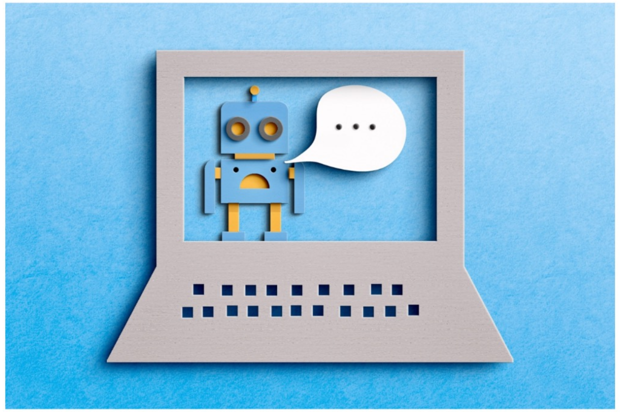
Search
TechCrunch+
Startups
Venture
Security
AI
Crypto
Apps
Events
More

AI

'Inaudible' watermark could identify AI-generated voices

Devin Coldewey @techcrunch / 12:09 AM GMT+1 • February 2, 2023

Comment



TechCrunch Disrupt Sept 19-21 Sa… Francisco, CA

Register Now

IMATAG

BRAND CONTENT PROTECTION IMAGE TRACKING TECHNOLOGY COPYRIGHT BACK TO BLOG NEWSLETTER

COPYRIGHT IMAGE TRACKING TECHNOLOGY |
FEBRUARY 22, 2023

Invisible Watermarking Technology: a Solution to the Trust Issue with Generative AI Visual Content

Share this article

Subscribe to the blog!

QUICK READ

The trust issue with AI-generated content

JULY 21, 2023

THE WHITE HOUSE

FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI

Building Systems that Put Security First

- **The companies commit to investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights.** These model weights are the most essential part of an AI system, and the companies agree that it is vital that the model weights be released only when intended and when security risks are considered.

Earning the Public's Trust

- **The companies commit to developing robust technical mechanisms to ensure that users know when content is AI generated, such as a watermarking system.** This action enables creativity with AI to flourish but reduces the dangers of fraud and deception.

Laws

- EU AI Act, Article 50.2

  Providers of AI systems […] generating synthetic audio, image, video or text content, shall ensure the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. […] Providers shall ensure their technical solutions are effective, interoperable, robust, and reliable as far as this is technically feasible.

- California State Legislature, AB-3211

  Generative AI system providers must embed imperceptible and indelible watermarks in synthetic content, detailing the content's origins. Watermarks must be designed to be maximally indelible and retain information even if the content is altered

- White House Executive order, Section 10
- Chinese Interim Measures on Generative AI, Article 12

Watermarking vs. Forensics

- Advantages
  - Better detectability/robustness
    - Forensics (passive): detection of unintentional statistical traces
    - Watermarking (active): deliberate insertion of a secret weak signal
  - Theoretical guarantees
    - Low false positive rate, and provably low
- Drawbacks
  - Degradation of the quality
    - Definition?
  - Modification of the generation process
    - Post-hoc watermarking? Within the generation?

Who are we fighting?

- Joe Sixpack – *“Keep Honest People Honest”*
    - Generative AI = commercial product
    - ✓ Watermarking (law)
    - ✓ Forensics (large number of examples for training a classifier)

- Mafia/belligerent nations
    - Able to learn their own generative AI
    - 💀 Watermarking
    - 💀 Forensics (too few examples)

- Open-source gen-AI?

Generative AI + Watermarking

2 approaches

1. Generate and then watermark
   - Ok for black box AI
   - Not secure for open source models
     - Ex: Stable Diffusion on Hugging Face

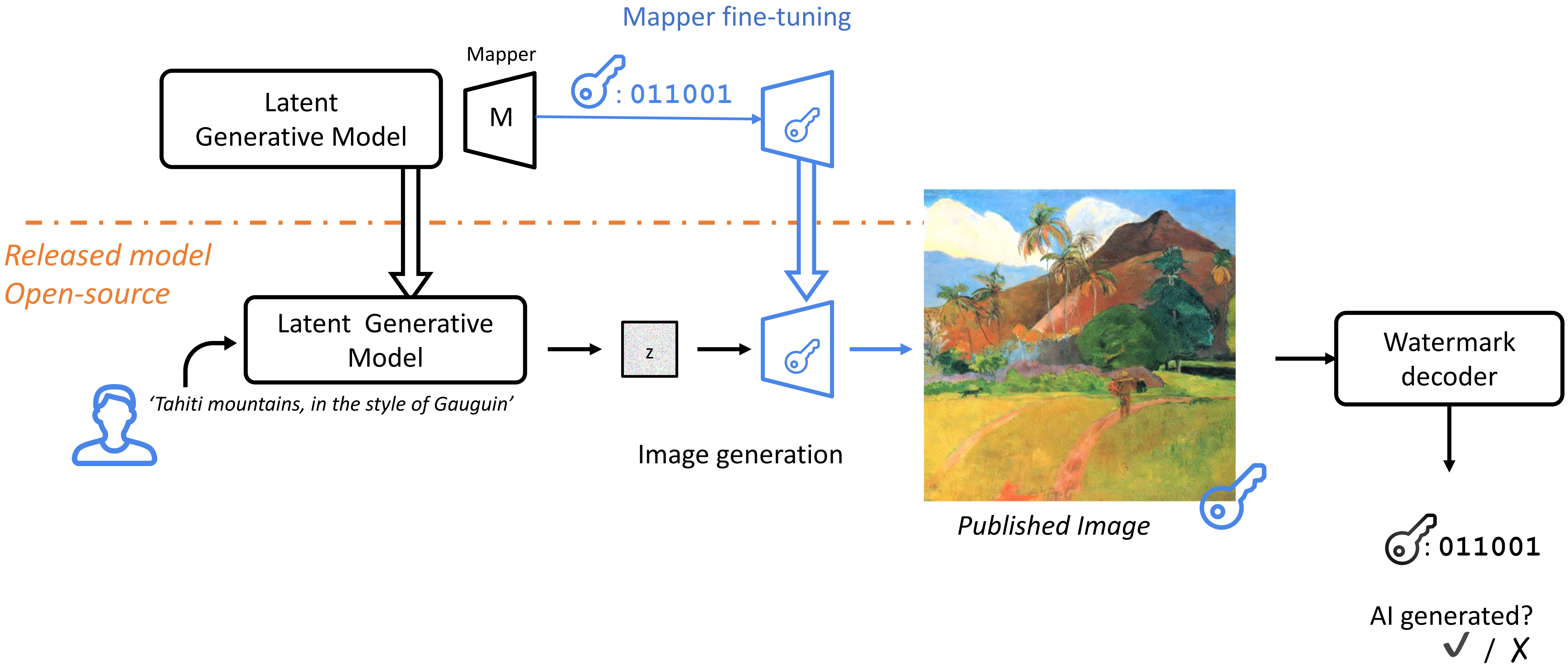
```
for x_sample in x_samples:
    x_sample = 255. * rearrange(x_sample.cpu().numpy(), 'c h w -> h w c')
    img = Image.fromarray(x_sample.astype(np.uint8))
   #img = put_watermark(img, wm_encoder)
    img.save(os.path.join(sample_path, f"{base_count:05}.png"))
    base_count += 1
    sample_count += 1

all_samples.append(x_samples)
```

2. Natively generate watermarked content
   A. Train the generative model over watermarked contents
   B. Fine-tune the generative model so that it learns to “speak” to a watermark decoder

Approche 2B: Stable Signature

"The Stable Signature: Rooting Watermarks in Latent Diffusion Models", ICCV 2023

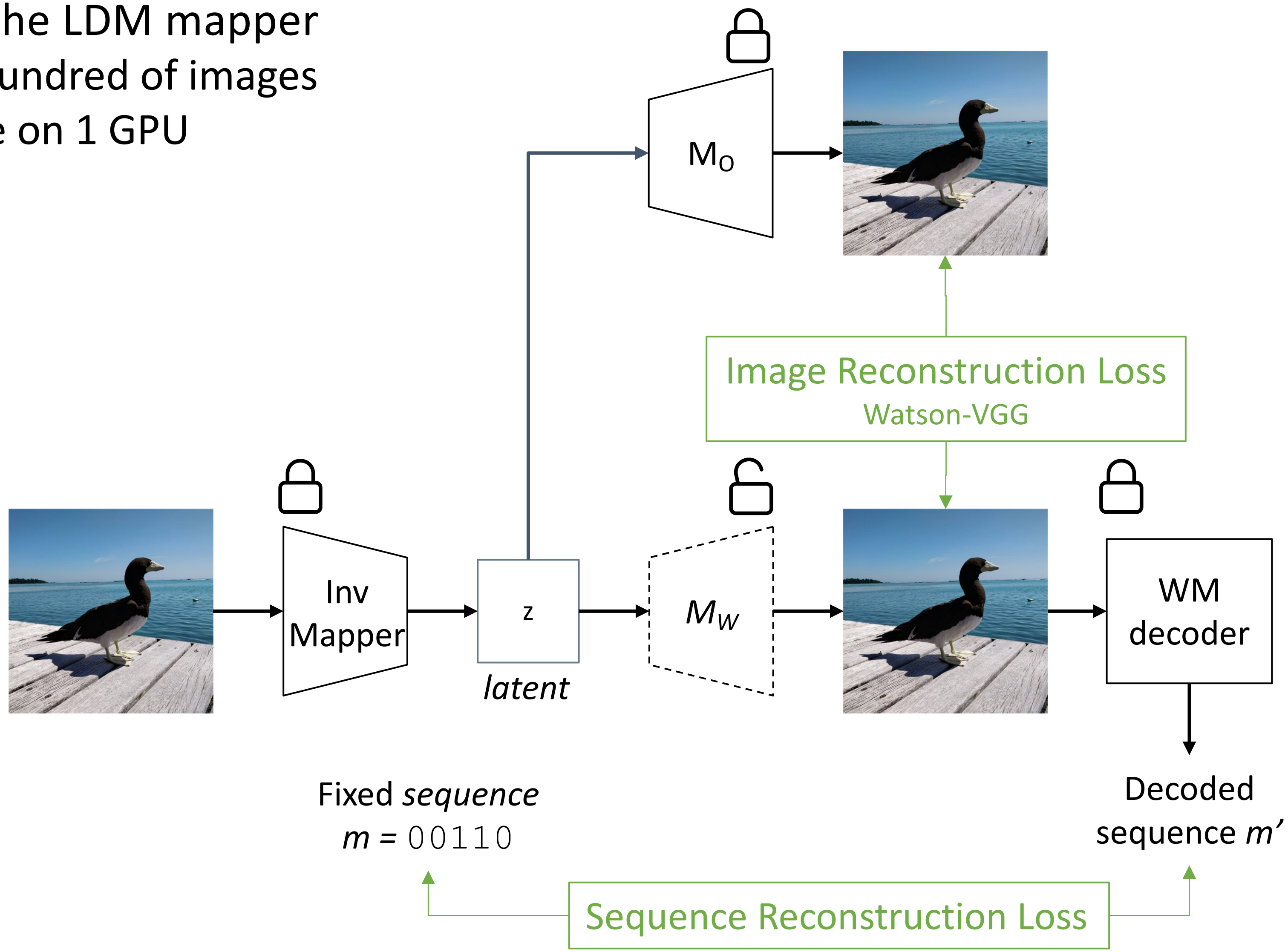
P. Fernandez, G. Couairon, H. Jégou, M. Douze, T. Furon



Approche 2B: Stable Signature

Fine-tuning of the LDM mapper

- Over a hundred of images
- 1 minute on 1 GPU



Approche 2B: Stable Signature

Natural image

→ Decoder → Random m': 10101

Generated by our model

→ Decoder → m' : 00110

$M(m, m')$ = popcount [¬ XOR (m, m')]

H_0 Image is not generated by our model (*eg.* natural image)

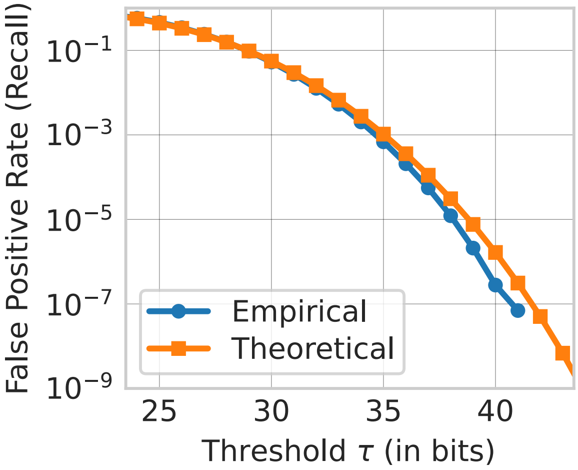
$M(m, m')$ follows a binomial distribution $\sim \mathcal{B}(n, {}^1/_2)$

H_1 Image is generated by our model (AI-generated)

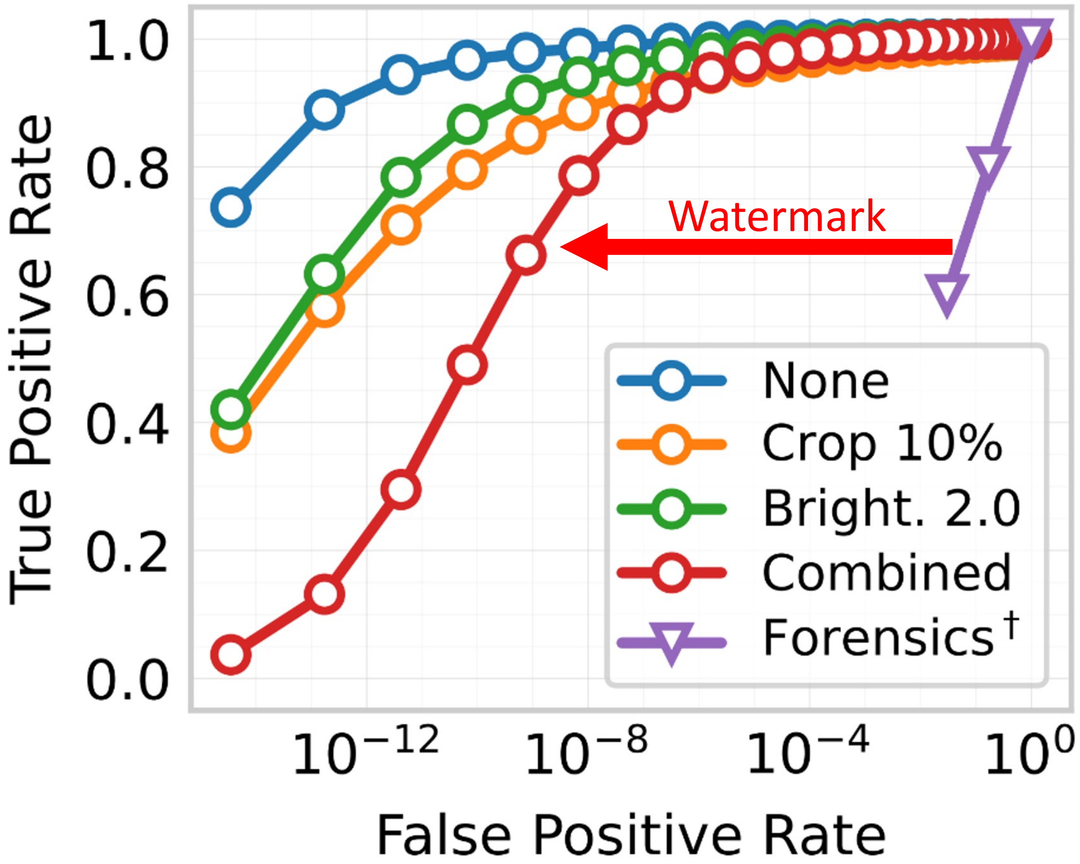
$M(m, m') \approx n$

Test: $M(m, m') \geq \tau \rightarrow$ detection

FPR (False Positive Rate): $\mathbb{P}(M(m, m') \geq \tau | H_0) = I_{1/2}(\tau, n + 1 - \tau)$



Approche 2B: Stable Signature



Message length $n = 48$ bits

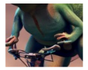
TPR: 1k generated images + attacks

Plot for $\tau \in [0,n]$

None

Brightness 2.0

Crop 10%



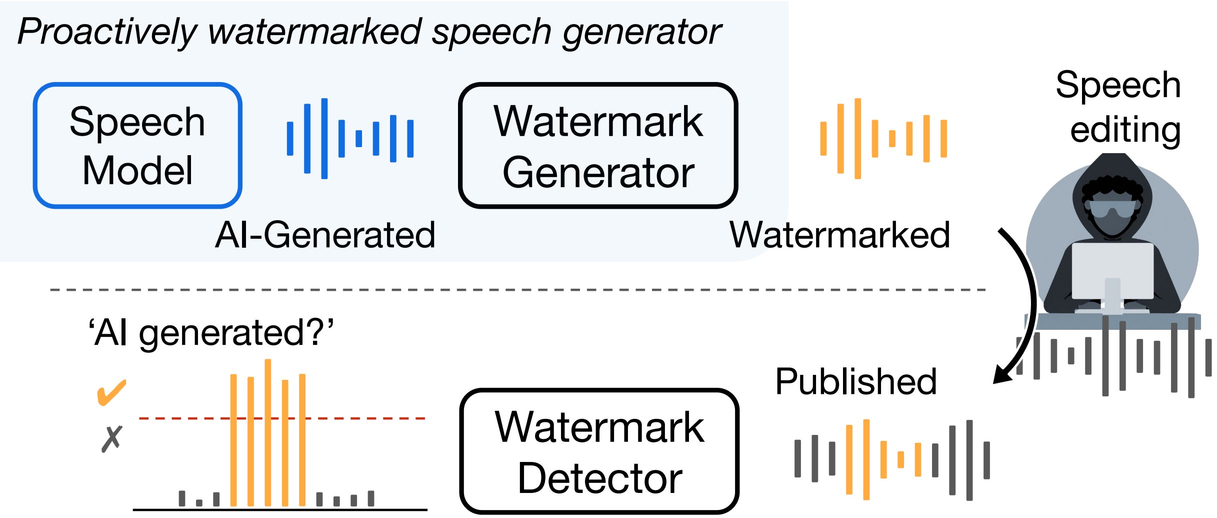
Combined (crop, bright, JPEG)

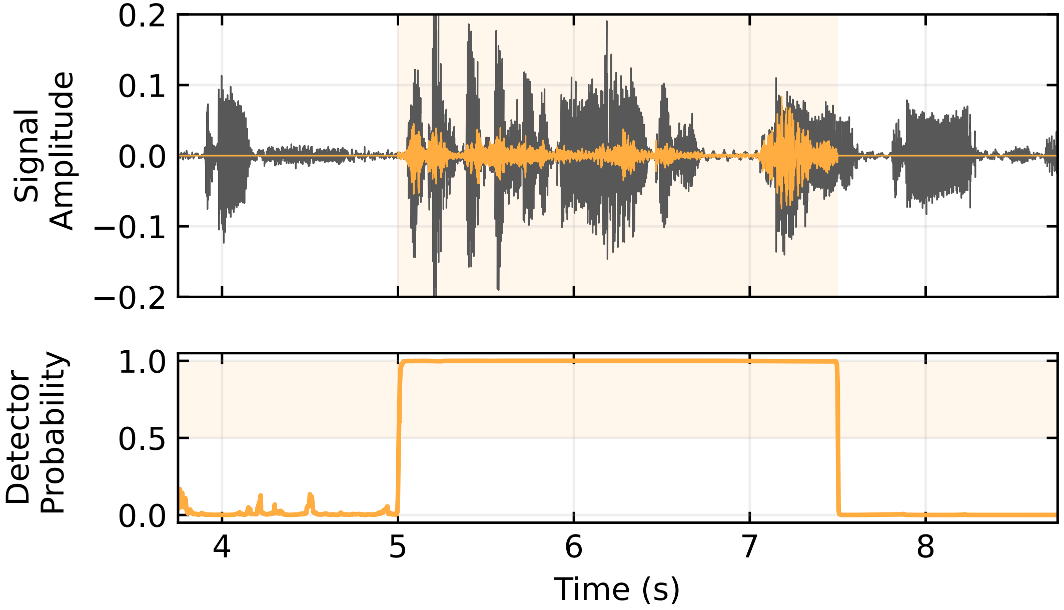
Forensics† means passive detection with *"On the detection of synthetic images generated with diffusion models", Corvi et al., 2022*

Approche 2B: Voice cloning

"Proactive Detection of Voice Cloning with Localized Watermarking", ICML 2024

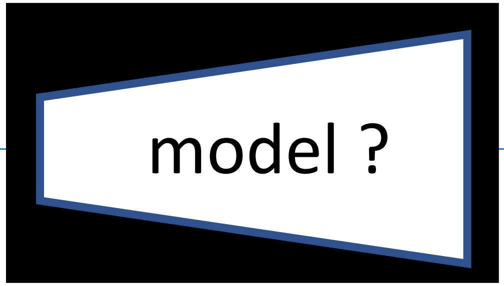
R. San Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon





Approche A: LLM watermarking

‘Write an essay about Paul Gauguin’



Paul Gauguin was a French Post-Impressionist artist. Unappreciated until after his death, Gauguin is now recognized for his experimental use of colour and Synthetist style that were distinct from Impressionism.

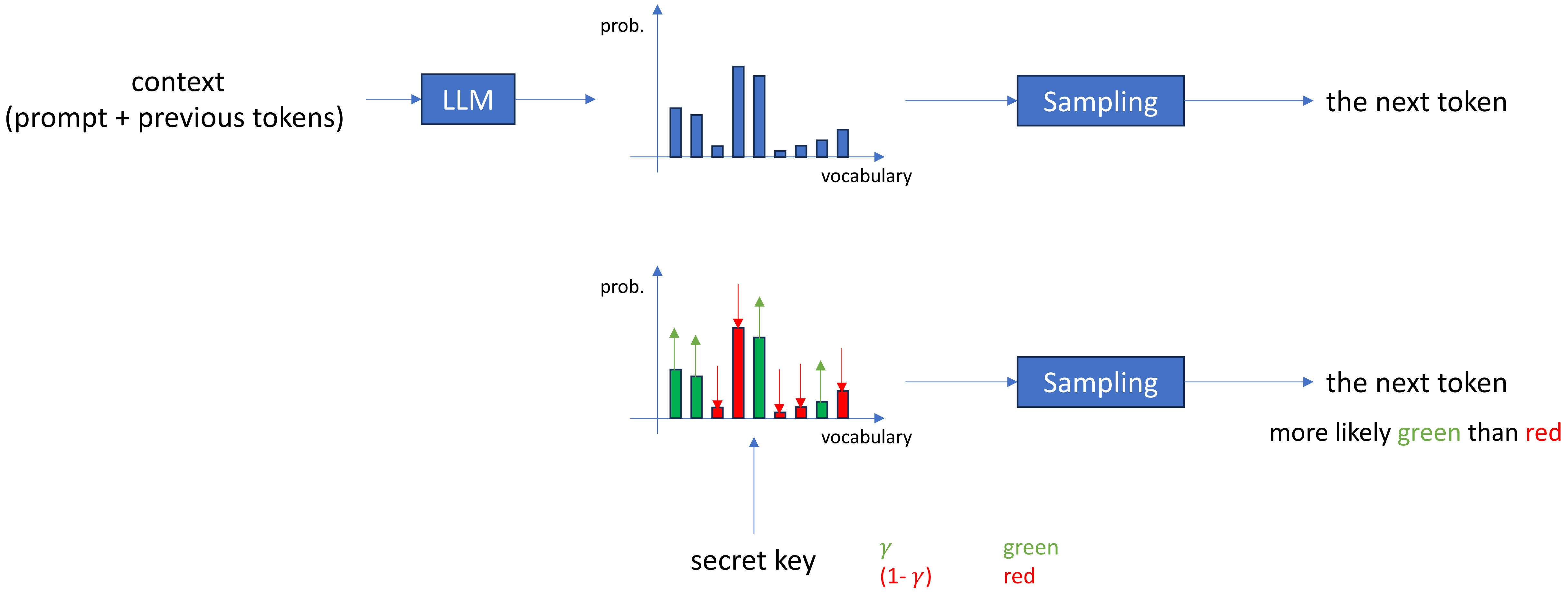
“Three bricks to consolidate watermarks for LLM”, IEEE WIFS 2023

P. Fernandez, A. Chaffin, K. Tit, V. Chappelier, T. Furon

Approche A: LLM watermarking

"A watermark for Large Language Models", ICML 2023

J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miller, T. Goldstein



Approche A: LLM watermarking

Paul Gauguin was a French Post-Impressionist artist. Unappreciated until after his death, Gauguin is now recognized for his experimental use of colour and synthetist style that were distinct from Impressionism.

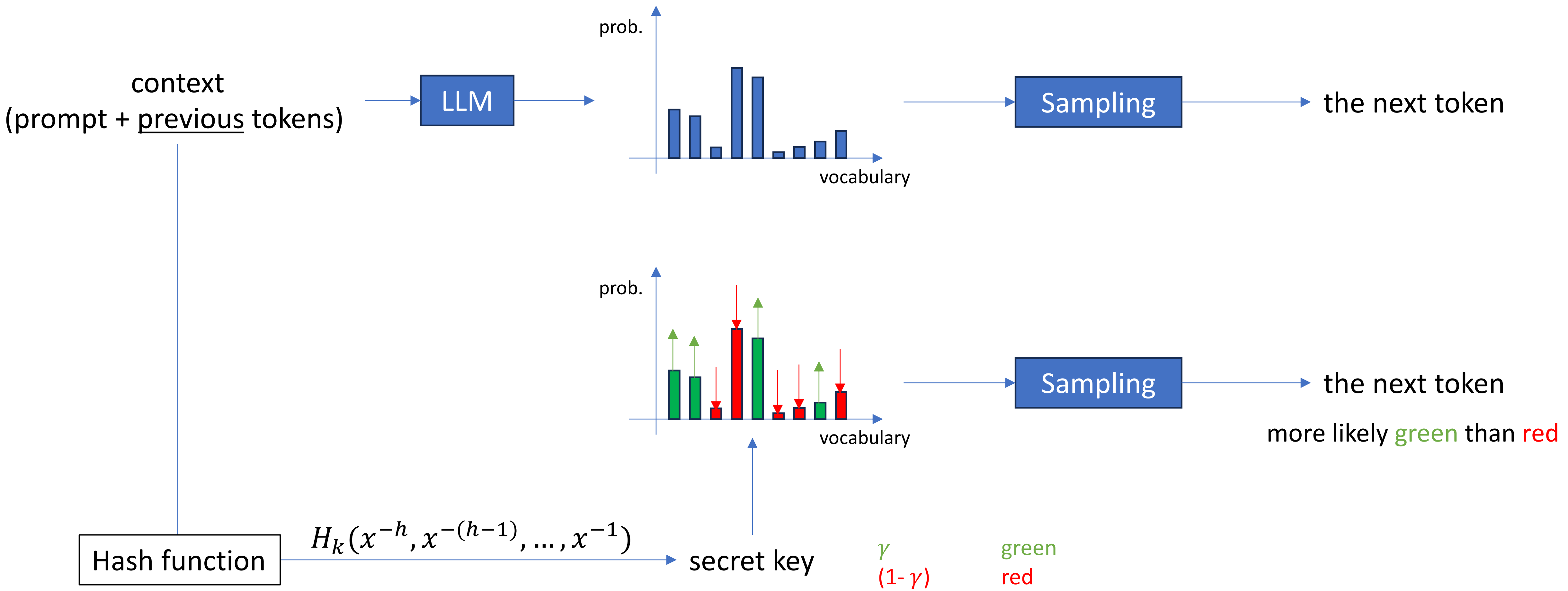
Number of green tokens: s
Total number of tokens: n

H_0 : If text not watermarked, then

$$S \sim \mathrm{B}(n, \gamma)$$
$$P(S \geq \tau) = I_\gamma(\tau + 1, n - \tau)$$

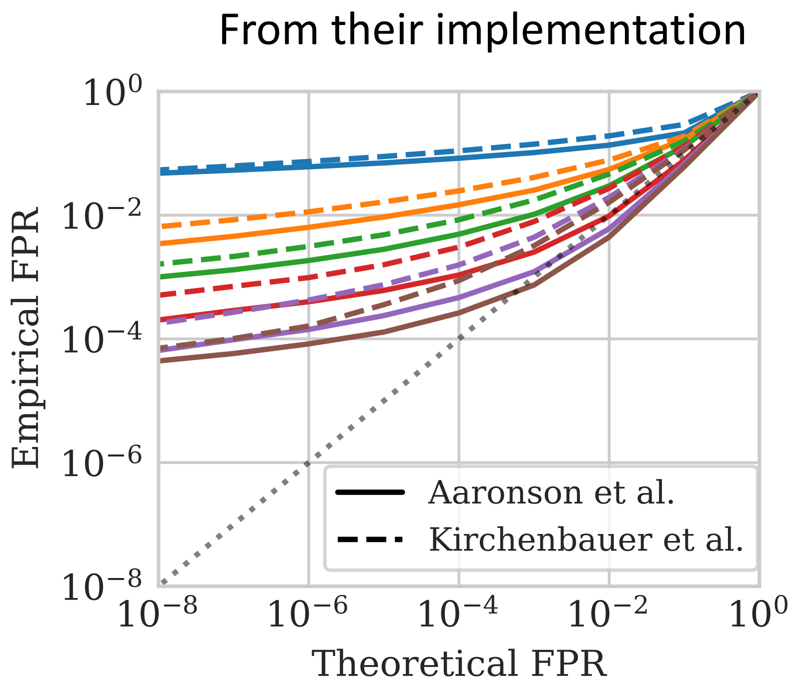
H_1 : If generated, then S deviates from $\mathrm{B}(n, \gamma)$ because green tokens are more frequent

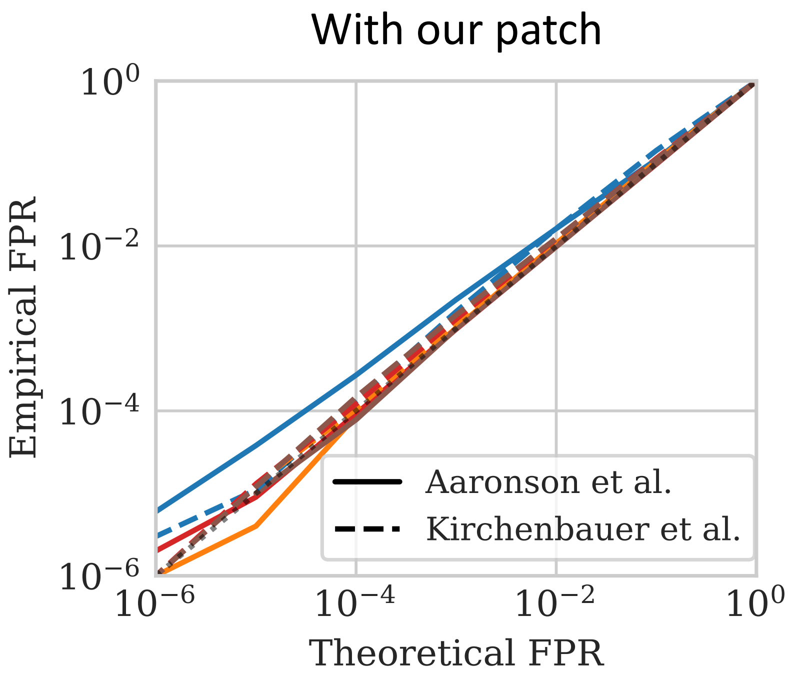
Approche A: LLM watermarking



Approche A: LLM watermarking

- Their False Positive Rates are not sound!!!





Conclusion : No fair comparison if FPR is not fully controlled

Approche A: LLM watermarking

Nesothrips is a genus of thrips in the family Phlaeothripidae

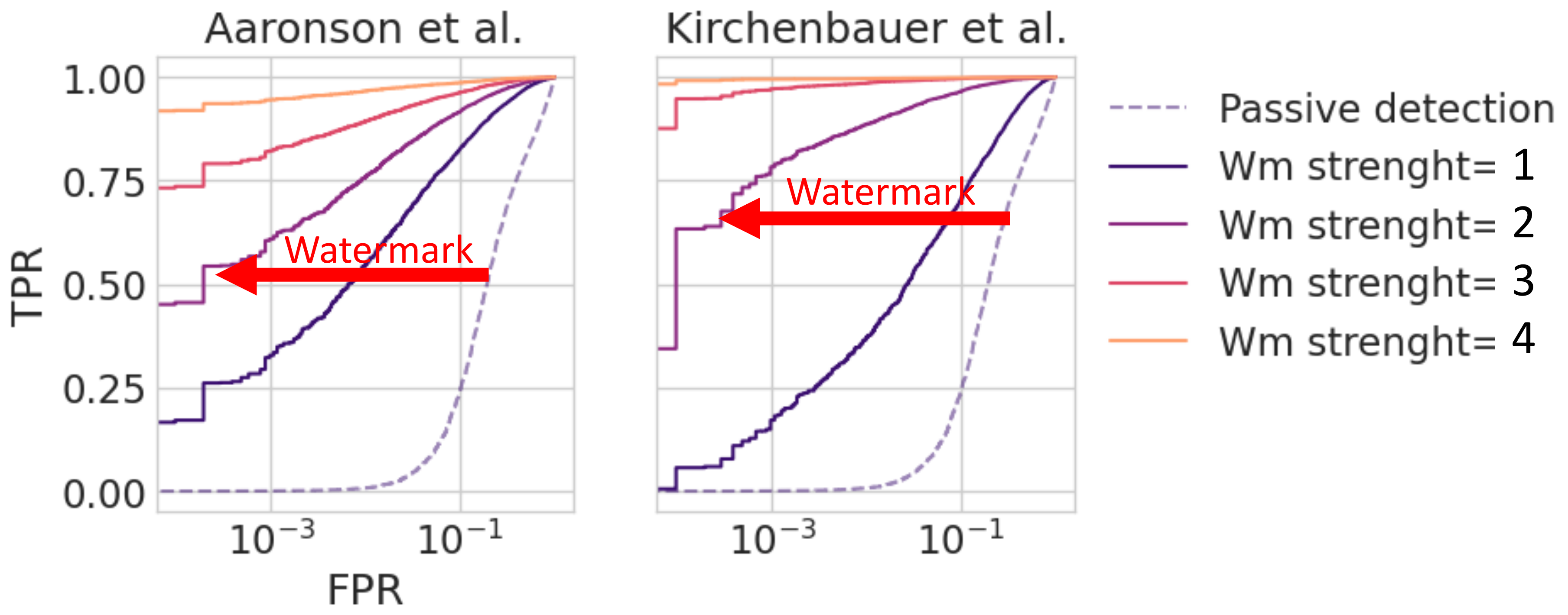
Species:

- Nesothrips alexandrae
- Nesothrips aoristus
- Nesothrips artocarpi
- Nesothrips badius
- Nesothrips barrowi
- Nesothrips brevicollis
- Nesothrips brigalowi
- Nesothrips capricornis
- Nesothrips carveri
- Nesothrips coorongi
- Nesothrips doulli
- Nesothrips eastopi
- Nesothrips fodinae
- Nesothrips hemidiscus
- Nesothrips lativentris

$\frac{s}{n} \approx 0.5 > \gamma = 0.25 \rightarrow$ deemed as watermarked

Approche A: LLM watermarking

- 10k positive AI-generated / 10k negative human generated (from OpenAssistant Conversations dataset)



Conclusion: Generative AI + watermarking

Complementary technical means

- Watermarking (real and AI-generated)
- Forensics
- Metadata (C2PA)
- Similarity search (fingerprinting)

Many unsolved questions remain:

- Is this a threat?
  - Who runs the detector? Is it publicly available?
  - Billions of contents will be generated, watermarked with the same technique
- Once compromised, the attacker may
  - Remove the watermark to pretend this content is real
  - Add a watermark to pretend this content is fake

Outlines

	Forensics	Watermarking
Decision making	Part 1	Part 2
Generative	Part 3	Part 4