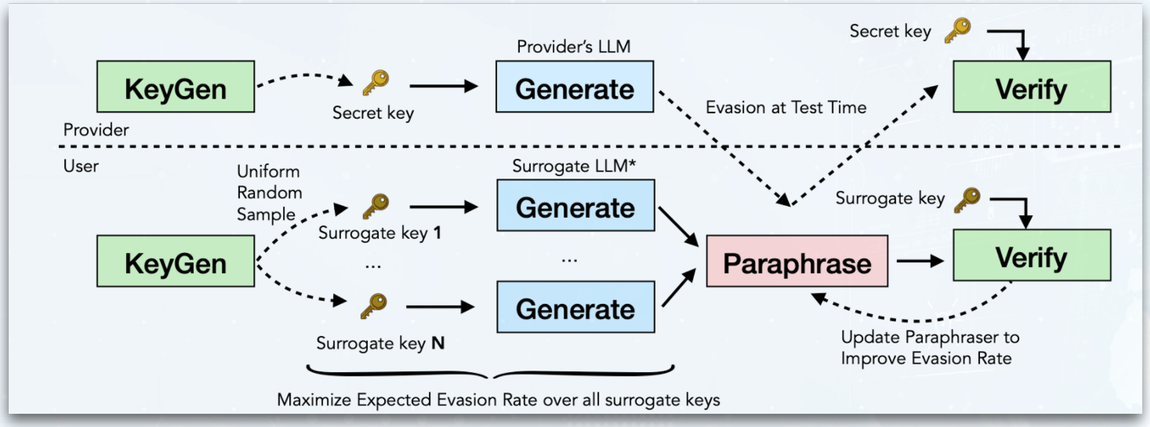
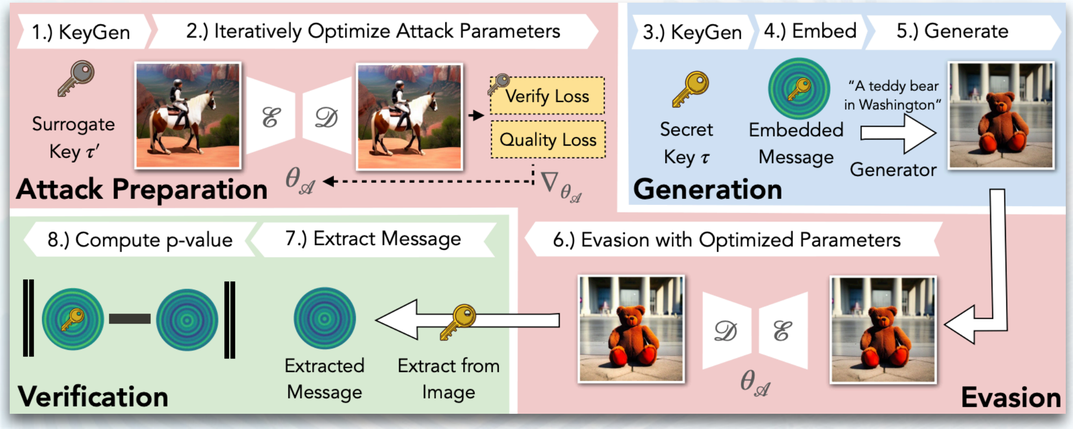
Mohamed bin Zayed University of Artificial Intelligence

Adaptively Robust and Forgery-Resistant Watermarking





Nils Lukas

nils.lukas@mbzuai.ac.ae

Assistant Professor in ML

September 18, 2025

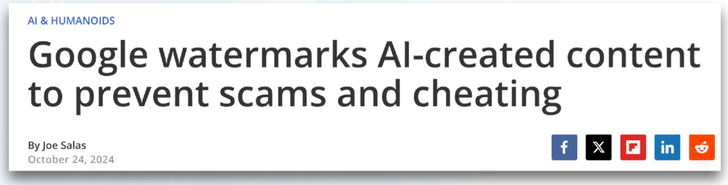
AVSeal Speaker Series

@FAIR

Risks of GenAI Today: Authenticity

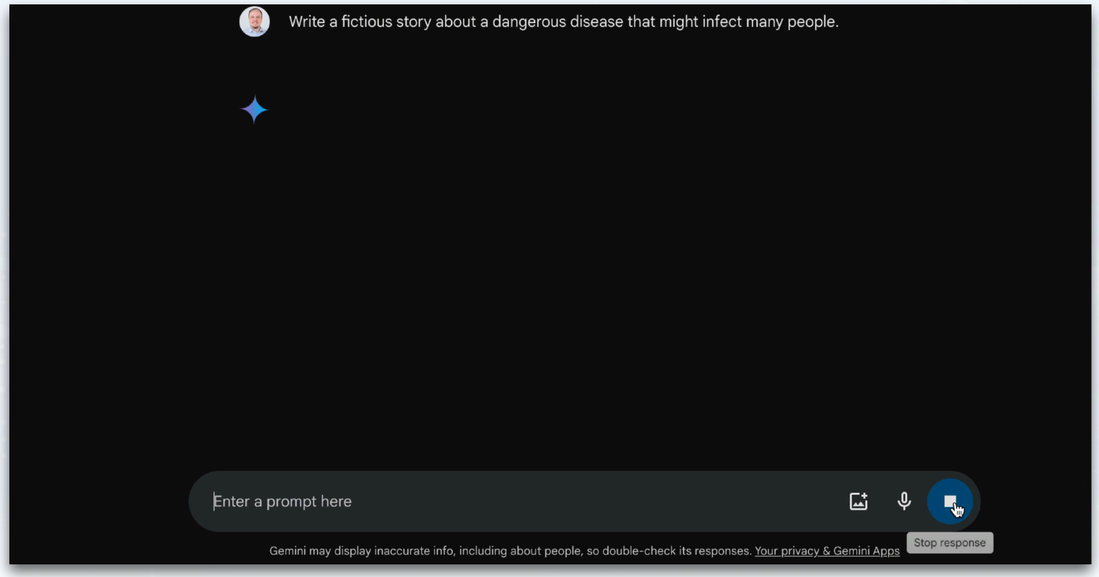
Mohamed bin Zayed University of Artificial Intelligence

- Generating high-quality content is **easy** and **cheap**
- Can lead to an **erosion of trust** in digital media



24/10/2024

23/09/2024



Examples: Training Data Contamination, Combating Misinformation, Data Signature and Attribution, Fraud Detection

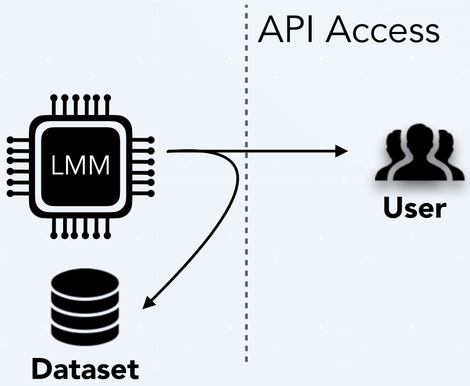
- Threat actors:
  - ~~A. Highly capable entities (e.g., targeted disinformation)~~
  - **B. Restricted capabilities (e.g., 'everyday users')**

→ *Millions of users, some may '**misuse**' GenAI*

Potential Solutions

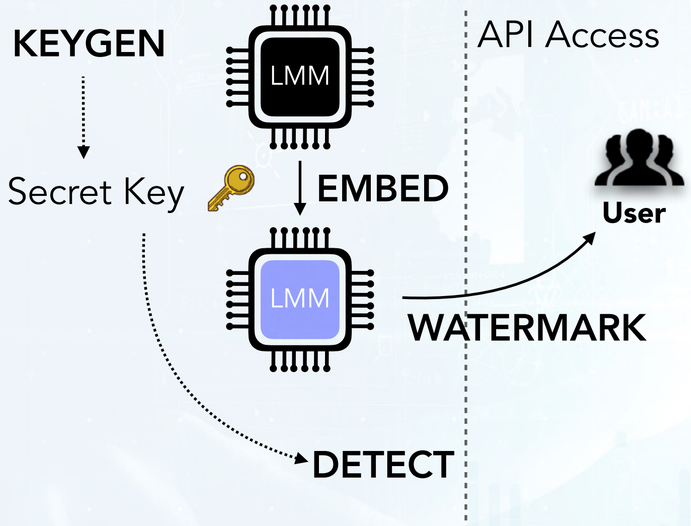
Mohamed bin Zayed University of Artificial Intelligence

Retrieval-based



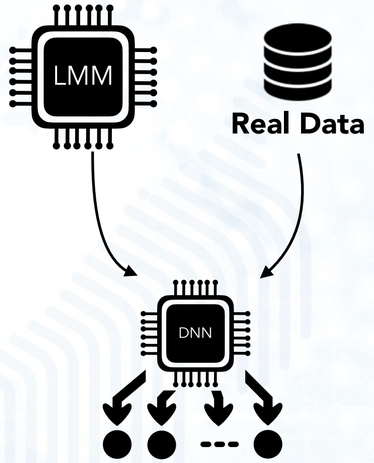
Cons

- High storage & retrieval costs
- No open source
- No user privacy



Cons

- Key must be kept secret
- Generation process must be modified



Cons

- Unreliable
- Low accuracy

NEWS

Google is adding AI watermarks to photos manipulated by Magic Editor

06/02/2025

Google DeepMind is making its AI text watermark open source

23/10/2024

Google now adds watermarks to all its AI-generated content

11/12/2024

OpenAI Quietly Shuts Down Its AI Detection Tool

Dashing the hopes of educators, OpenAI decommissions its AI Classifier due to poor accuracy.

24/07/2023

California Senate Bill-942 (Sept 19, 2024)

Mohamed bin Zayed University of Artificial Intelligence

Free, audiovisual content, public detector

22757.2. (a) A covered provider shall make available an AI detection tool at no cost to the user that meets all of the following criteria:

(1) The tool allows a user to assess whether image, video, or audio content, or content that is any combination thereof, was created or altered by the covered provider's GenAI system.

(2) The tool outputs any system provenance data that is detected in the content.

(3) The tool does not output any personal provenance data that is detected in the content.

(4) (A) Subject to subparagraph (B), the tool is publicly accessible.

California AI Transparency Act, Chapter 25

California SB-942

**If the service has more than 1 million monthly users/visitors p.a., and the service is publicly accessible*

California Senate Bill-942 (Sept 19, 2024)

Mohamed bin Zayed University of Artificial Intelligence

Visible to the user, robust

22757.3. (a) A covered provider shall offer the user the option to include a manifest disclosure in image, video, or audio content, or content that is any combination thereof, created or altered by the covered provider's GenAI system that meets all of the following criteria:

(1) The disclosure identifies content as AI-generated content.

(2) The disclosure is clear, conspicuous, appropriate for the medium of the content, and understandable to a reasonable person.

(3) The disclosure is permanent or extraordinarily difficult to remove, to the extent it is technically feasible.

California AI Transparency Act, Chapter 25

**If the service has more than 1 million monthly users/visitors p.a., and the service is publicly accessible*

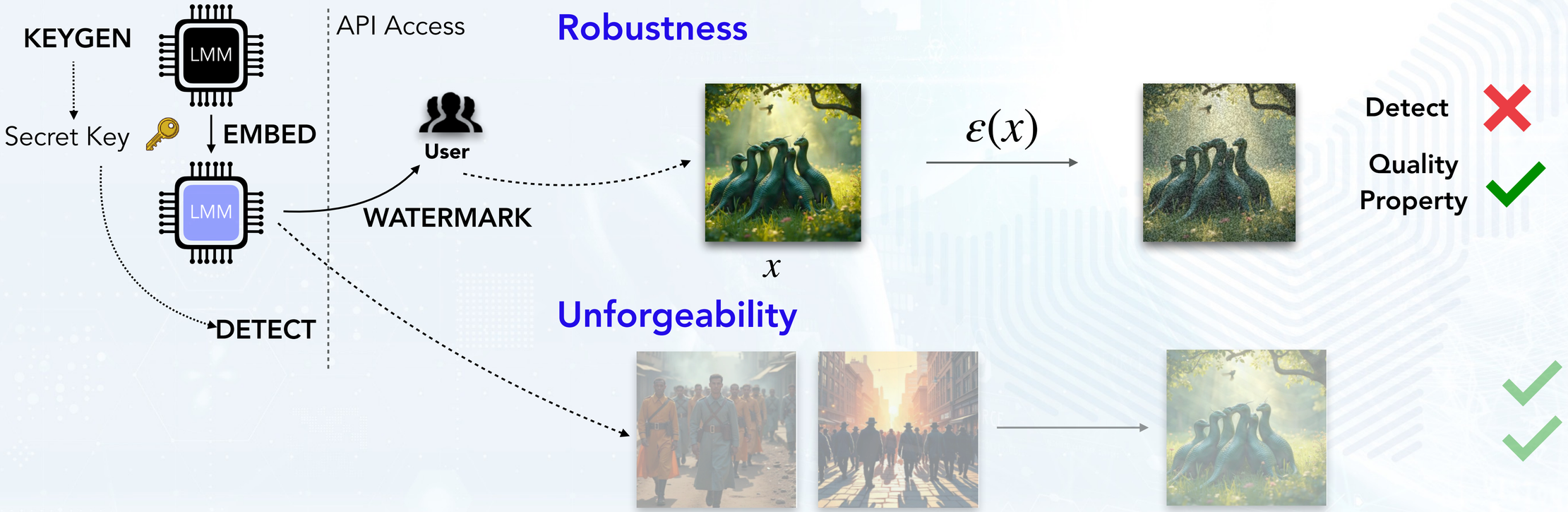
California SB-942

Robustness

Mohamed bin Zayed University of Artificial Intelligence

Definition 3.5 (Robustness)**.** A watermark detector Detect is robust to a channel $\mathcal{E}$ with error ϵ for property P if, for any prompt π,

$$\Pr_{\substack{\mathsf{gk},\mathsf{dtk} \\ x \leftarrow \mathsf{Watermark}^{\mathcal{M}}_{\mathsf{gk}}(\pi) \\ x' \leftarrow \mathcal{E}(x)}} \left[\mathsf{Detect}_{\mathsf{dtk}}(x') \rightarrow \text{false and } P(\mathcal{M}, \pi, x) = \text{true} \right] \leq \varepsilon.$$



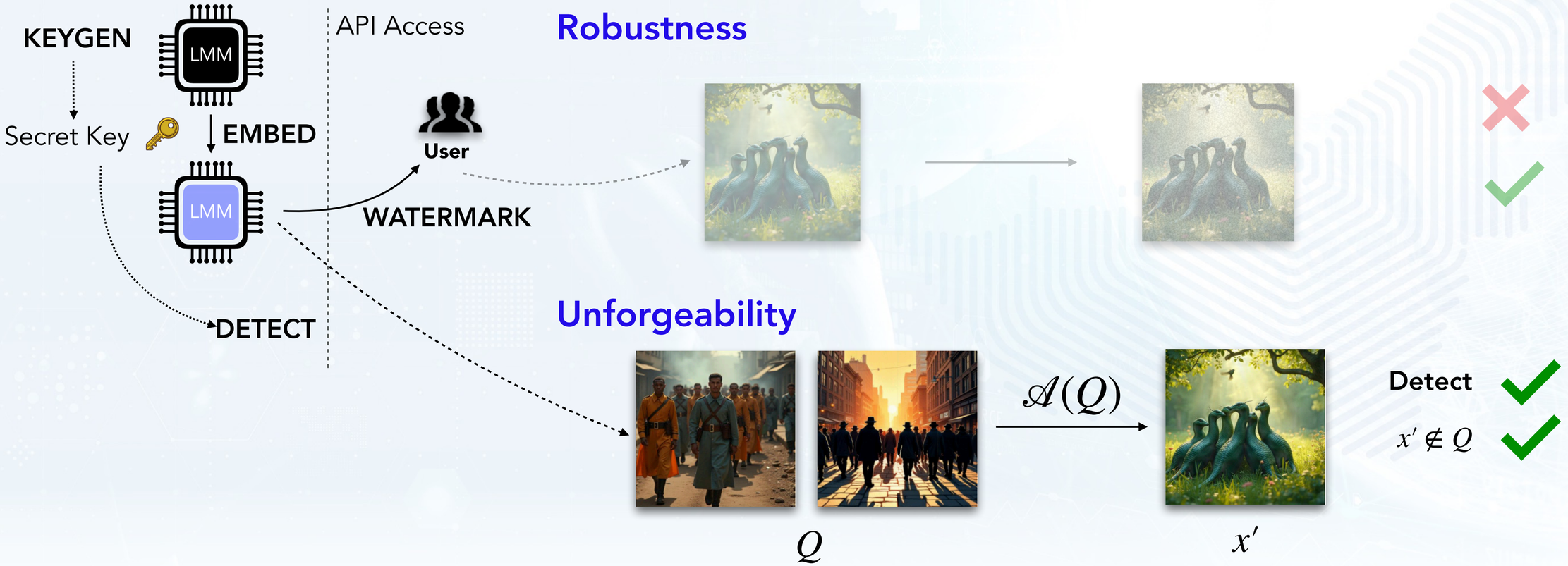
Unforgeability

Mohamed bin Zayed University of Artificial Intelligence

Definition 3.6 (Unforgeability [72])**.** A watermark is *unforgeable* if for all λ and polynomial-time algorithms $\mathcal{A}$,

$$\Pr_{\substack{\mathsf{gk},\mathsf{ak} \\ x \leftarrow \mathcal{A}^{\mathsf{Watermark}^{\mathcal{M}}_{\mathsf{gk}}}(1^\lambda,\mathsf{ak})}} [\mathsf{Attribute}_{\mathsf{ak}}(x) \rightarrow \text{true and } x \notin \mathcal{Q}] \leq \mathsf{negl}(\lambda),$$

where $\mathcal{Q}$ denotes the set of responses obtained by $\mathcal{A}$ on its queries to the watermarked model.



What is the Security Definition?

Mohamed bin Zayed University of Artificial Intelligence

What is the security definition??

Want to detect any output where AI made the "main creative contribution"

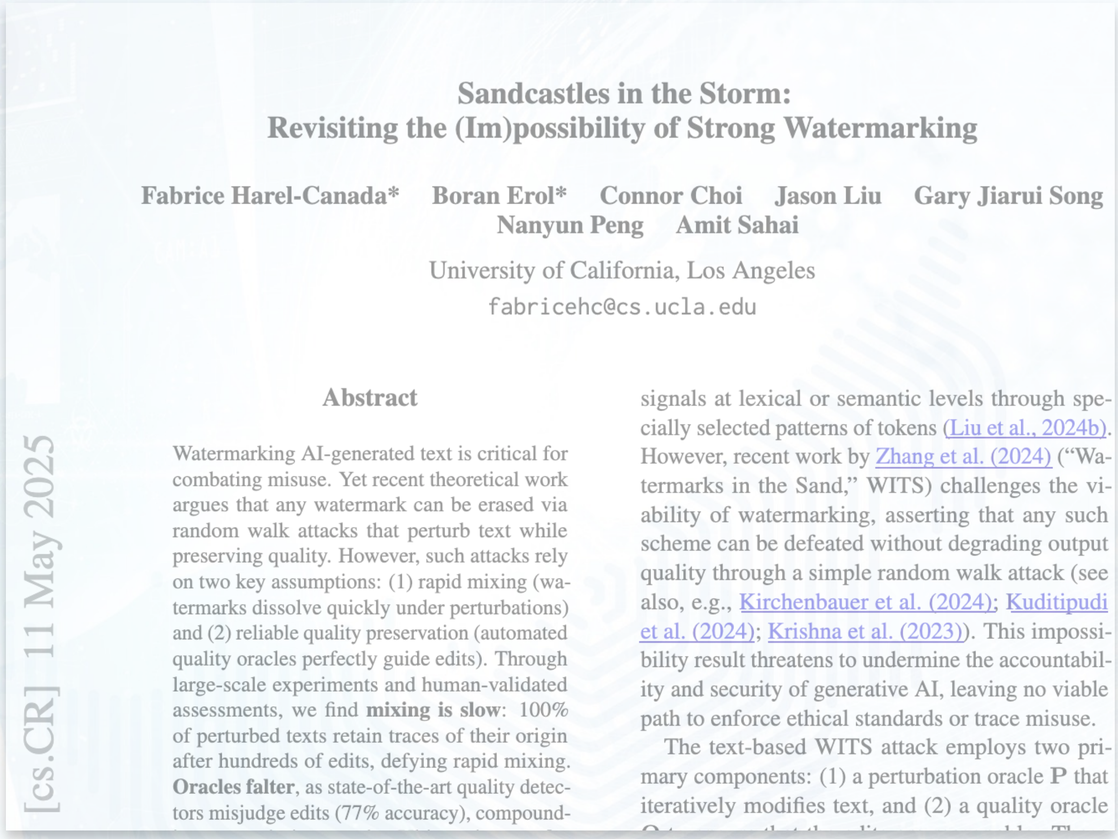
"The human modified the AI output in at most trivial or uncreative ways"

"Arbitrary choices are present only because the AI happened to make them"

"The human can't explain or justify the choices"

"Impossibility theorem" (Barak et al. 2023). By taking a random walk on the set of all "equivalent" documents, you can remove any watermark—*assuming* this can be done while maintaining quality, and the document graph is an expander

Scott Aaronson Slides, from WMARK@ICLR'25



Sandcastles in the Storm:
Revisiting the (Im)possibility of Strong Watermarking

Fabrice Harel-Canada* Boran Erol* Connor Choi Jason Liu Gary Jiarui Song
Nanyun Peng Amit Sahai

University of California, Los Angeles
fabricehc@cs.ucla.edu

Abstract

Watermarking AI-generated text is critical for combating misuse. Yet recent theoretical work argues that any watermark can be erased via random walk attacks that perturb text while preserving quality. However, such attacks rely on two key assumptions: (1) rapid mixing (watermarks dissolve quickly under perturbations) and (2) reliable quality preservation (automated quality oracles perfectly guide edits). Through large-scale experiments and human-validated assessments, we find **mixing is slow**: 100% of perturbed texts retain traces of their origin after hundreds of edits, defying rapid mixing. **Oracles falter**, as state-of-the-art quality detectors misjudge edits (77% accuracy), compound-

signals at lexical or semantic levels through specially selected patterns of tokens (Liu et al., 2024b). However, recent work by Zhang et al. (2024) ("Watermarks in the Sand," WITS) challenges the viability of watermarking, asserting that any such scheme can be defeated without degrading output quality through a simple random walk attack (see also, e.g., Kirchenbauer et al. (2024); Kuditipudi et al. (2024); Krishna et al. (2023)). This impossibility result threatens to undermine the accountability and security of generative AI, leaving no viable path to enforce ethical standards or trace misuse.

The text-based WITS attack employs two primary components: (1) a perturbation oracle **P** that iteratively modifies text, and (2) a quality oracle

What is the Security Definition?

What is the security definition??

Want to detect any output where AI made the "main creative contribution"

"The human modified the AI output in at most trivial or uncreative ways"

"Arbitrary choices are present only because the AI happened to make them"

"The human can't explain or justify the choices"

"Impossibility theorem" (Barak et al. 2023). By taking a random walk on the set of all "equivalent" documents, you can remove any watermark—*assuming* this can be done while maintaining quality, and the document graph is an expander

Scott Aaronson Slides, from WMARK@ICLR'25

fabricehc@cs.ucl

Abstract

Watermarking AI-generated text is critical for combating misuse. Yet recent theoretical work argues that any watermark can be erased via random walk attacks that perturb text while preserving quality. However, such attacks rely on two key assumptions: (1) rapid mixing (watermarks dissolve quickly under perturbations) and (2) reliable quality preservation (automated quality oracles perfectly guide edits). Through large-scale experiments and human-validated assessments, we find **mixing is slow**: 100% of perturbed texts retain traces of their origin after hundreds of edits, defying rapid mixing. **Oracles falter**, as state-of-the-art quality detectors misjudge edits (77% accuracy), compound-

signa
cially
Howe
terma
abilit
scher
quali
also,
et al.
bility
ity ar
path
Th
mary
iterat

Under what assumptions can watermarks be robust?

Attack Success Criteria

Mohamed bin Zayed University of Artificial Intelligence

Threat (from earlier)

Goals

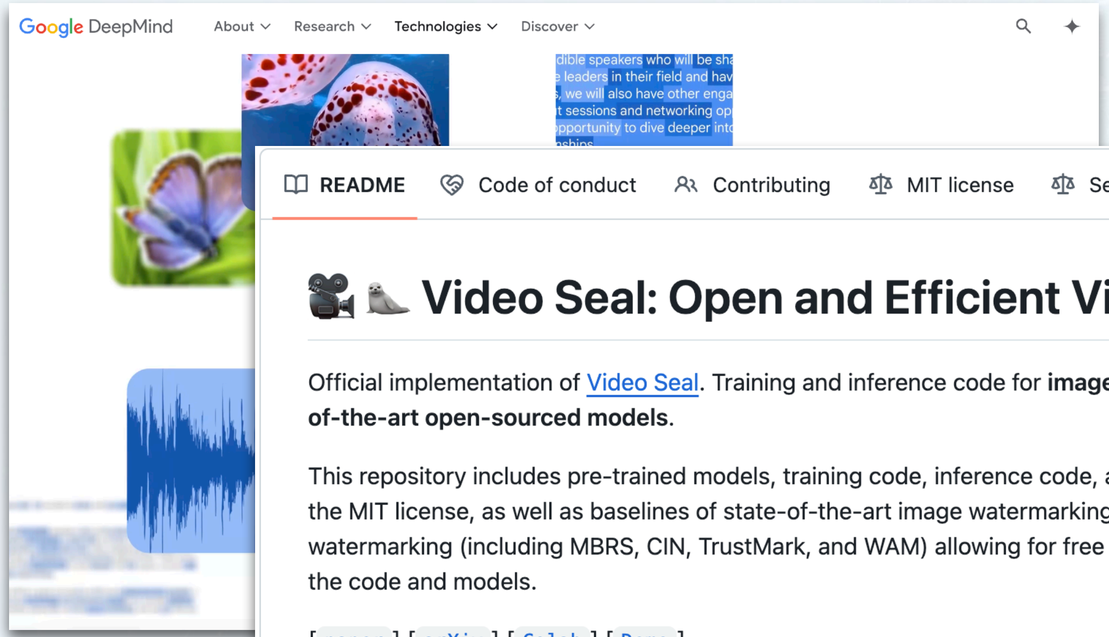
- Choose a text quality metric Q on the prompt π and output (e.g., CLIPScore, LLM-as-a-Judge,..)
- Attacker's best baseline quality is q_1 (using open models)
- Given $x \leftarrow \text{Watermark}(\pi)$, can the attacker generate $x' \leftarrow \varepsilon(x)$ s.t. $Q(\pi, x') > q_1$ AND $\text{Detect}(x') > \tau$

Attacker's advantage using the watermarked service

in generating high-quality content WITHOUT a watermark

Implementations are Open-Source

Mohamed bin Zayed University of Artificial Intelligence



README | Code of conduct | Contributing | MIT license | Security

Video Seal: Open and Efficient Video Watermarking

Official implementation of Video Seal. Training and inference code for **image and video watermarking**, and **state-of-the-art open-sourced models**.

This repository includes pre-trained models, training code, inference code, and evaluation tools, all released under the MIT license, as well as baselines of state-of-the-art image watermarking models adapted for video watermarking (including MBRS, CIN, TrustMark, and WAM) allowing for free use, modification, and distribution of the code and models.

[paper] [arXiv] [Colab] [Demo]

What's New

Google DeepMind

Meta

Microsoft

Watermarks in preview in Azure OpenAI Service

ated a w termark AI-peech

eventually help tackle the growing use of voice nformation.

By Melissa Heikkilä June 18, 2024

FAIR, Meta

Implementations are public ..

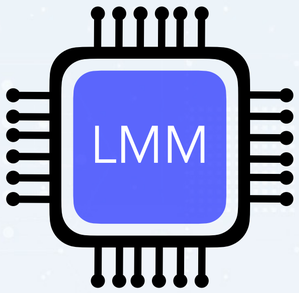
Threat Model

Mohamed bin Zayed University of Artificial Intelligence

Model Provider

Algorithm	Inputs
KEYGEN(..)	Seed, Parameter
EMBED(..)	Model, Key, Message
DETECT(..)	Image, Key, Message

Secret Key



Watermarked Target Model

API Access

DETECT

WATERMARK

Adversary

No-box: No access to the target model

Offline: No access to VERIFY

Private: No access to the secret key or randomness

Computationally bounded: Cannot train own LLM

Adaptive: Knows watermarking scheme (but not the inputs used by the provider)

Surrogate Model: Can access less capable, open-source models

Hugging Face

Evaluating and Improving Robustness

Mohamed bin Zayed University of Artificial Intelligence

Optimizing Adaptive Attacks against Watermarks for Language Models

Abdulrahman Diaa [1] Toluwani Aremu [2] Nils Lukas [2]

Abstract

Large Language Models (LLMs) can be misused to spread unwanted content at scale. Content watermarking deters misuse by hiding messages in content, enabling its detection using a secret *watermarking key*. Robustness is a core security property, stating that evading detection requires (significant) degradation of the content's quality. Many LLM watermarking methods have been proposed, but robustness is tested only against *non-adaptive* attackers who lack knowledge of the watermarking method and can find only suboptimal attacks. We formulate watermark robustness as an objective function and use preference-based optimization to tune *adaptive* attacks against the specific watermarking method. Our evaluation shows that (i) adaptive attacks evade detection against all surveyed watermarks, (ii) training against *any* watermark succeeds in evading unseen watermarks, and (iii) optimization-based attacks are cost-effective. Our findings underscore the need to test robustness against adaptively tuned attacks. We release our adaptively tuned paraphrasers at `https://github.com/nilslukas/ada-wm-evasion`.

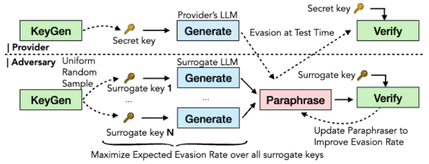


Figure 1: Adaptive attackers know the watermarking algorithms (KeyGen, Verify), but not the secret key, so they can optimize a paraphraser against a specific watermark.

Content watermarking enables the detection of generated outputs by embedding hidden messages that can be extracted with a secret watermarking key. Some LLM providers, such as DeepMind (2024) and Meta (San Roman et al., 2024), have already deployed watermarking to promote the ethical use of their models. A threat to these providers are users who perturb generated text to evade watermark detection while preserving text quality. Such undetectable, generated text could further erode trust in the authenticity of digital media (Federal Register, 2023).

A core security property of watermarking is *robustness*, which requires that evading detection is only possible by sig-

ICML'25, Spotlight

Watermark-removing Paraphrasers updated Feb 24

A set of paraphrasers adaptively-tuned to remove state-of-the-art watermarks

`DDiaa/WM-Removal-Unigram-Qwen2.5-3B`
Text Generation • Updated Apr 1

`DDiaa/WM-Removal-Unigram-Llama-3.2-3B`
Text Generation • Updated Apr 1 • 1

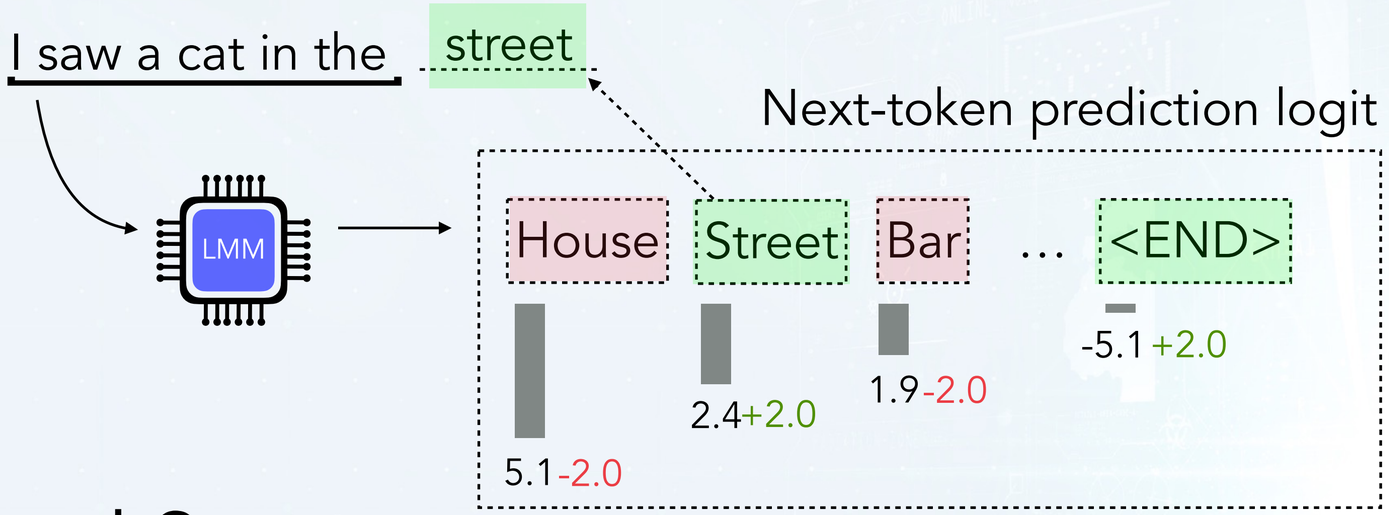
`DDiaa/WM-Removal-EXP-Qwen2.5-3B`
Text Generation • Updated Apr 1 • 7

`DDiaa/WM-Removal-EXP-Qwen2.5-0.5B`

Our Models
are open source

Recap: Red-Green Watermark

Mohamed bin Zayed University of Artificial Intelligence



Watermark Generate

Step 1.) Draw a pseudo-random number $f_\tau(x_0, \ldots, x_3)$

Step 2.) Partition vocabulary into green and red list

Step 3.) Bias tokens in the green list

Step 4.) Softmax and sample

Step 5.) Repeat

Verify

Given a text x, count green tokens and conduct a statistical test

NeurIPS Competition (Dec, 2024)

Mohamed bin Zayed University of Artificial Intelligence

77 teams, 2 tracks, total of 7,000 USD prize money

Black-box Track

Rank	Participant	Detection	Quality	Total
1	**Ours**	**0.043**	**0.136**	**0.143**
2	Team-Jafari	0.063	0.158	0.170
3	Team-Yepeng	0.087	0.177	0.197

beige-box Track

Rank	Participant	Detection	Quality	Total
1	**Ours**	**0.037**	**0.153**	**0.157**
2	Team-Asky	0.050	0.176	0.183
3	Team-Jafari	0.127	0.222	0.256

Figure 1: *Top row*: Original watermarked images. *Bottom row*: Images after our attack, with minimal perceptual difference from the originals, showcasing the effectiveness of our method in preserving visual fidelity. Best viewed zoomed in.

Published at the 1st workshop on GenAI Watermarking, collocated with ICLR 2025

First-Place Solution to NeurIPS 2024 Invisible Watermark Removal Challenge

Fahad Shamshad[1], **Tameem Bakr**[1], **Yahia Shaaban**[1],
Noor Hussein[1,2], **Karthik Nandakumar**[1,2], **Nils Lukas**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[2]Michigan State University (MSU), USA
{firstname.lastname}@mbzuai.ac.ae

Abstract

Content watermarking is an important tool for the authentication and copyright protection of digital media. However, it is unclear whether existing watermarks are robust against adversarial attacks. We present the **winning solution** to the NeurIPS 2024 *Erasing the Invisible* challenge, which stress-tests watermark robustness under varying degrees of adversary knowledge. The challenge consisted of two tracks: a black-box and beige-box track, depending on whether the adversary knows which watermarking method was used by the provider. For the **beige-box** track, we leverage an *adaptive* VAE-based evasion attack, with a test-time optimization and color-contrast restoration in CIELAB space to preserve the image's quality. For the **black-box** track, we first cluster images based on their artifacts in the spatial or frequency-domain. Then, we apply image-to-image diffusion models with controlled noise injection and semantic priors from ChatGPT-generated captions to each cluster with optimized parameter settings. Empirical evaluations demonstrate that our method successfully **achieves near-perfect watermark removal** (95.7%) with negligible impact on the residual image's quality. We hope that our attacks inspire the development of more robust image watermarking methods.

GenAI Workshop@ICLR'25, Oral

Finding Adversarial Corruption Channels

Mohamed bin Zayed University of Artificial Intelligence

The attacker can prepare offline by locally 'simulating' the watermark

An attacker can **adversarially** optimize for a channel that undermines robustness

Definition 3.5 (Robustness). A watermark detector Detect is robust to a channel $\mathcal{E}$ with error ϵ for property P if, for any prompt π,

$$\Pr_{\substack{\mathsf{gk},\mathsf{dtk} \\ x \leftarrow \mathsf{Watermark}^{\mathcal{M}}_{\mathsf{gk}}(\pi) \\ x' \leftarrow \mathcal{E}(x)}} \left[\mathsf{Detect}_{\mathsf{dtk}}(x') \rightarrow \text{false and } P(\mathcal{M}, \pi, x) = \text{true} \right] \leq \varepsilon.$$

Robustness

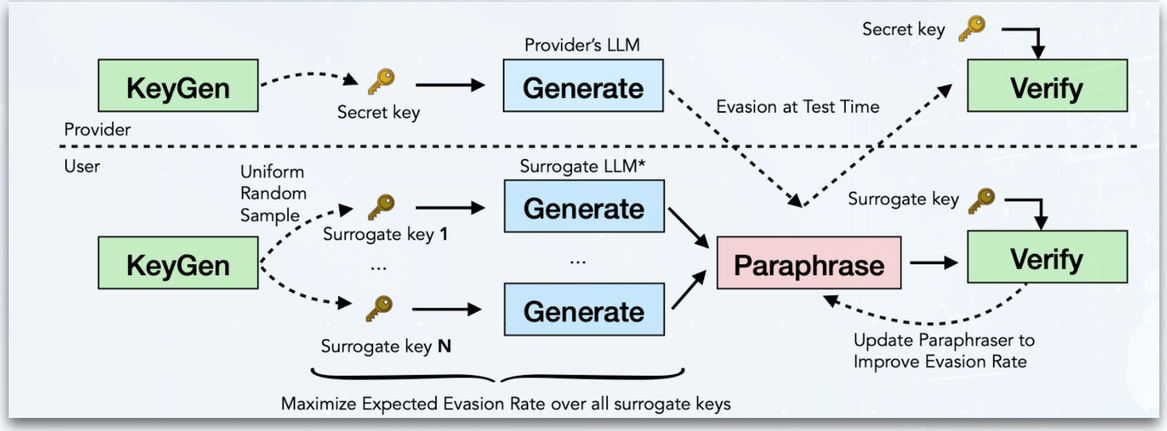
$$\max_{\theta_P} \mathbb{E}_{\substack{\gamma \sim \mathcal{R} \\ m' \sim \mathcal{M} \\ q \sim \mathcal{T}}} \left[\mathbb{E}_{\substack{\tau' \leftarrow \textsc{KeyGen}(\theta_S, \gamma) \\ \theta_S^* \leftarrow \textsc{Embed}(\theta_S, \tau', m') \\ x \leftarrow H_{\theta_S^*}(q)}} \textsc{Verify}\left(P_{\theta_P}(x), \tau', m'\right) + Q\left(P_{\theta_P}(x), x\right) \right]$$

Find a corruption (e.g., paraphraser) …

so that over the KeyGen randomness …

Any watermarked surrogate model ..

Is not robust against this channel!



Finding Adversarial Corruption Channels

Mohamed bin Zayed University of Artificial Intelligence

Problems:

- We need demonstrations (watermarked, non-watermarked) which we do not have
- Cannot easily backpropagate through LLMs to optimize for non-watermarked text

Propose Solution:

- Adaptively generate pairs via base model and rejection sampling
- Optimize over inherent KEYGEN, EMBED and model uncertainty

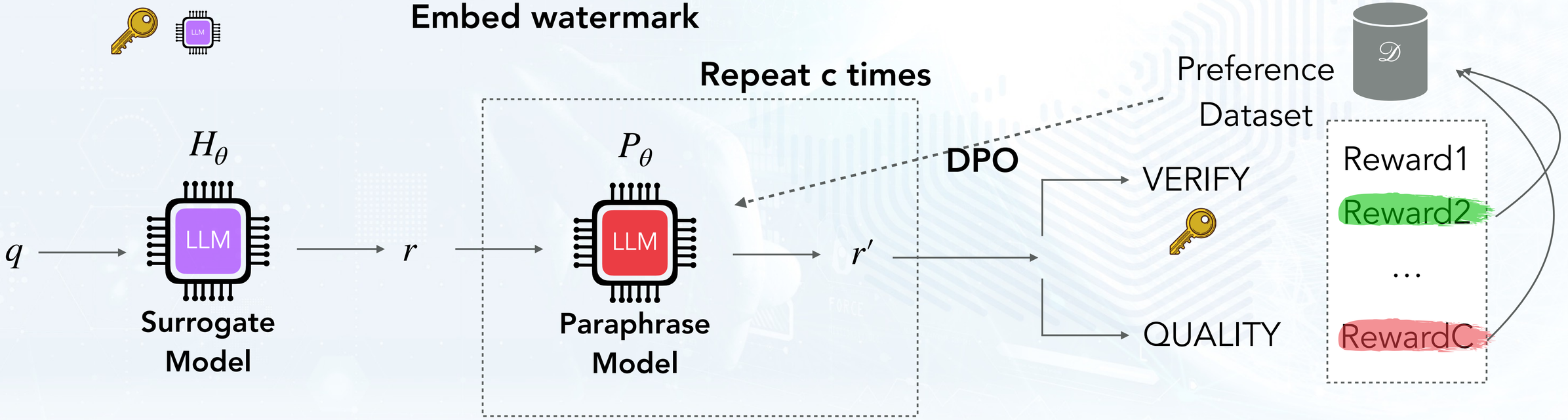
LLMs: Optimization via RL + Rejection Sampling

Mohamed bin Zayed University of Artificial Intelligence

1.) γ, m', q (Seed, Message, Query) **Sample inputs**

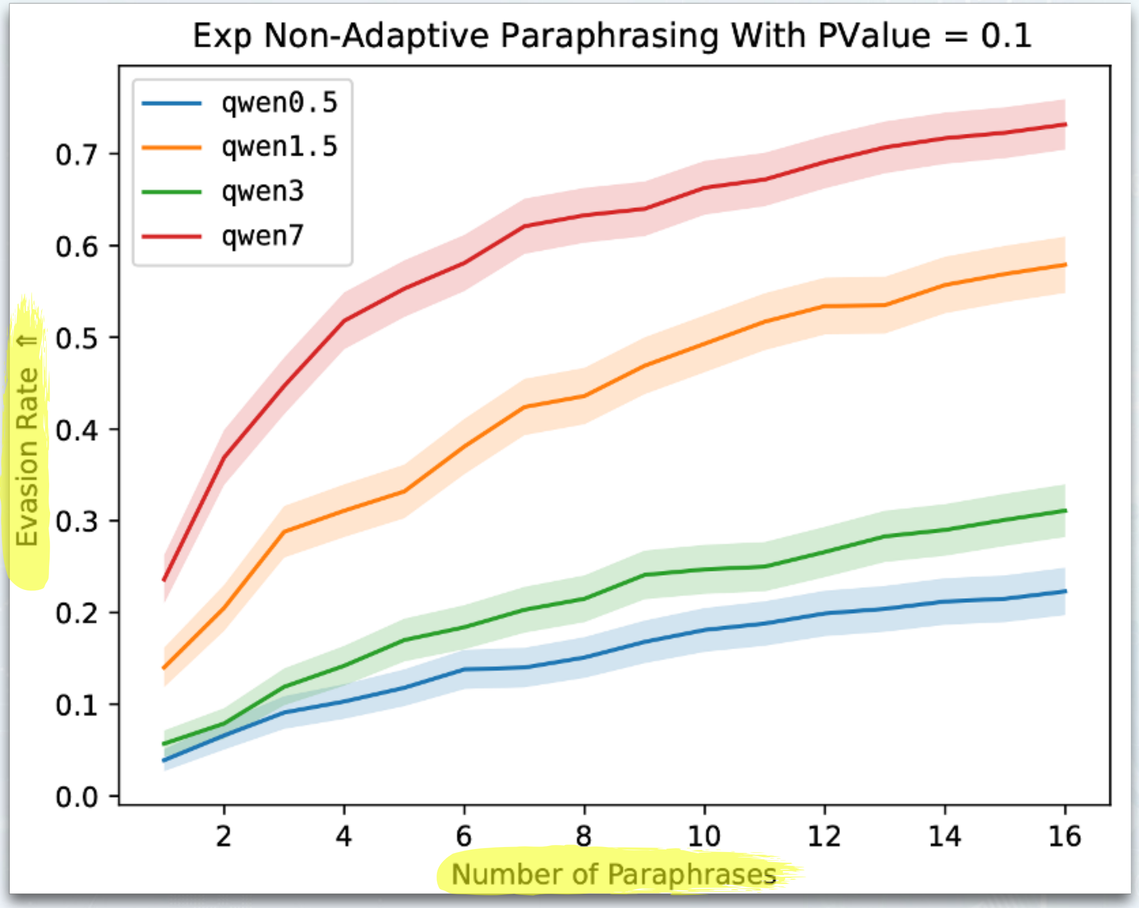
2.) KEYGEN **Generate key**

3.) EMBED **Embed watermark**



Compute versus Evasion Rates

Mohamed bin Zayed University of Artificial Intelligence



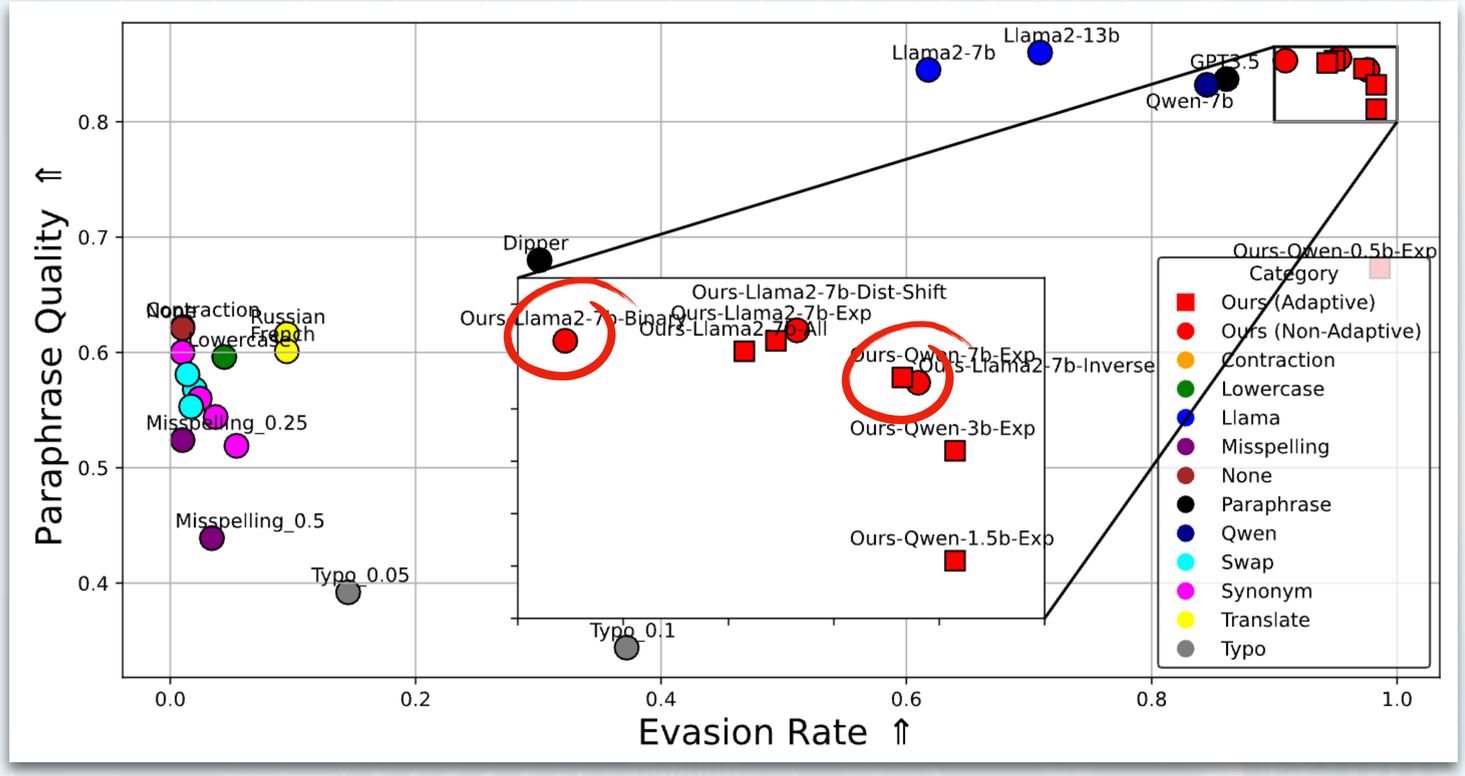
Collect preference dataset
through **rejection sampling**

Collecting 7,000 samples
requires ~5 GPU hours (2 USD per GPU hour)

Optimization takes ~2 GPU hours

Compute versus Evasion Rates

Provider uses a Llama3.1-70b model



Our models are up to $46\times$ smaller than the provider's model

Compute budget $\leq \$10$ USD

Evasion Rate: $96.7\,\%$, against 7 watermarking methods

`Llama2-13b-Dist-Shift` (463 green vs 41 red tokens)

Cervantes' "Don Quixote" is a groundbreaking masterpiece of literature that has enthralled readers for over 400 years. The novel tells the story of Alonso Quixano, an old man who becomes so consumed by the chivalrous tales of his day that he renames himself Don Quixote and embarks upon an adventurous journey filled with humor, satire, and timeless themes about the human condition.↩↩ The story follows Quixote and his trustworthy sidekick Sancho Panza, a simple and pragmatic squire who provides moments of humor and contrast to the lofty ideals of the delusional Quixano/Don Quixote. Their travels in rural Spain lead them to encounter all sorts of people they mistake for knights, ladies in distress, bandits, and giants who are in fact ordinary villagers going about their everyday lives.↩↩ One of the most profound and enduring elements of the novel's themes is the conflict of reality and perception. Quixote, driven by the chivalrous books he reads and his own imagination, mistakes windmills for giants, a puppet play as a real tale of love and betrayal, and a barber from a nearby village who is dressed in his Sunday best but Quixote sees as the Grand Duchess of Lithuania. Through these absurd but comedic misadventures, Cervantes creates a timeless commentary on the nature of truth, reality, and the danger of letting our imaginations run too wild. Don Quixote's journey also explores issues of class and nobility as he views his own lowly status as unknightly, while the peasants and traveling players he encounters view him with suspicion or indifference. Through these contrasts, Cervantes pokes fun at the social order and the idealized notion of chivalry.↩ Don Quixote has been praised for its realistic portrayal of human nature, including its weaknesses and fallibilities as well as the timeless wisdom of Cervantes' observations on society in late sixteenth-century Spain. At its core, the novel is an exploration of the human capacity to dream, delusions,

`Llama2-7b` (301 green vs 201 red tokens)

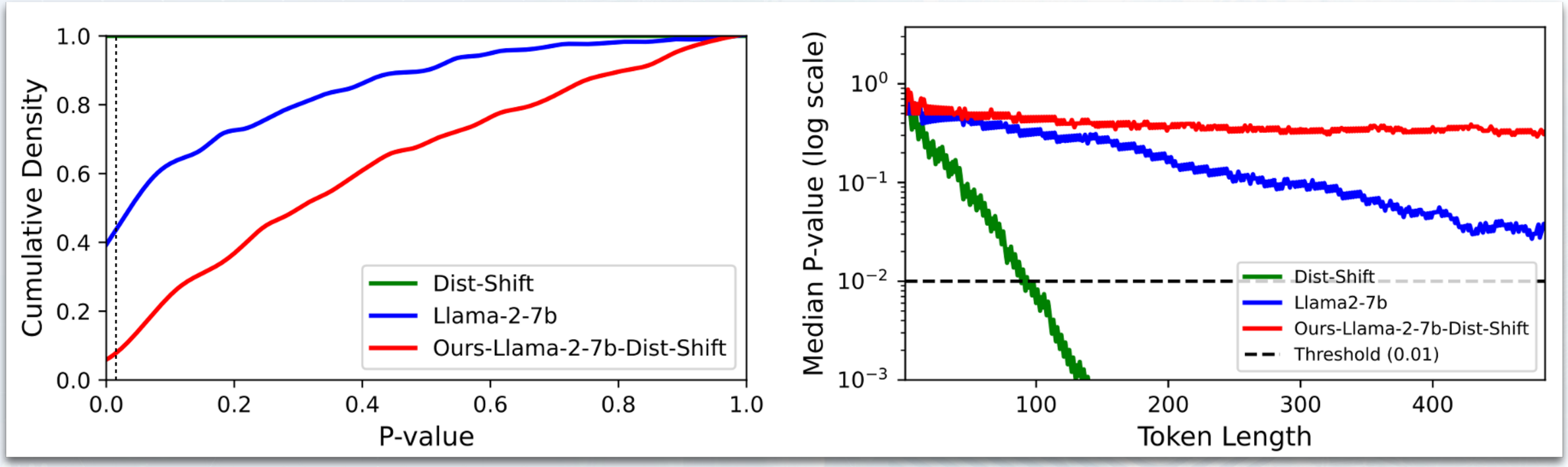
"Don Quixote" is a groundbreaking masterpiece of literature that has captivated readers for over 400 years. The novel tells the story of Alonso Quixano, an elderly man whose fixation on chivalrous tales leads him to change his name to DonQuixote and embark on a thrilling adventure replete with humor, satire, and timeless themes concerning the human condition.↩↩ The narrative follows Don Quixote and his loyal squire Sancho Panse, a practical and good-natured individual who provides moments of levity and contrast to the lofty ideals of the delusional Quixano/Don Quixote. Their travels across rural Spain result in encounters with various people who they misidentify as knights, distressed ladies, bandits, and ordinary villagers going about their daily lives.↩↩ One of the most enduring aspects of the novel's themes is the conflict between reality and perception. Quixote, driven by the chivalrous books he reads and his imagination, mistakes wind mills for giants, a puppet show for a real tale of love and betrayal, and a peasant in his Sunday best for the Grand Duchess of Lithuania. Through these absurd yet comedic misadventures, Cervantes offers a timeless commentary on the nature of truth, reality, and the dangers of allowing our imaginations to run wild. Don Quixote's journey also explores issues of class and nobility as he views his lowly status as unknightly, while the peasants and traveling players he encounters view him with suspicion or indifference. Through these contrasts, Cervantes pokes fun at the social order and the idealized notion of chivalry.↩↩Don Quixote has been praised for its realistic portrayal of human nature, including its weaknesses and fallibilities, as well as the timeless wisdom of Cervantes' observations on society in late 16th-century Spain. At its core, the novel is an exploration of the human capacity to dream, delude oneself, and confront reality, ultimately revealing the limitations and struggles of the human experience.

`Ours-Llama2-7b-Dist-Shift` (201 green vs 240 red tokens)

Spain has been blessed with a literary treasure that has captivated readers for over four centuries in the form of Miguel de Cervantes' immortal masterpiece, "Don Quixote". This timeless tome weaves a hilarious and poignant tale of a man consumed by the chivalrous stories of his youth, who renames himself Don Quixote and sets forth on a journey filled with moments of satire and a piercing examination of the human condition.↩↩As Don Quixote and his trusty squire Sancho Panza traverse rural Spain, they encounter various unsuspecting villagers, mistaking them for knights, maidens in distress, bandits, and even giants. Through these absurd yet poignant events, Cervantes deftly explores the blurred lines between reality and perception, highlighting the dangers of allowing our imaginations to run wild. The novel's themes of truth, class, and nobility are also expertly woven throughout the narrative, as Don Quixote's lowly status is juxtaposed with the condescending views of the peasants and traveling players he encounters.↩↩ Throughout the novel, Cervantes offers a biting commentary on the social hierarchy of his time, while also providing a realistic depiction of human nature, complete with its frailties and limitations. At its core, "Don Quixote" is a thought-provoking exploration of the human capacity for imagination, delusion, and the ever-elusive pursuit of truth.↩↩Endlessly praised for its timeless wisdom and biting social commentary, "Don Quixote" has captivated readers for centuries, leaving a lasting impact on literature and our understanding of the human experience. Its enduring themes and characters have become ingrained in our collective consciousness, reminding us of the importance of humility, empathy, and the boundless power of the human imagination.

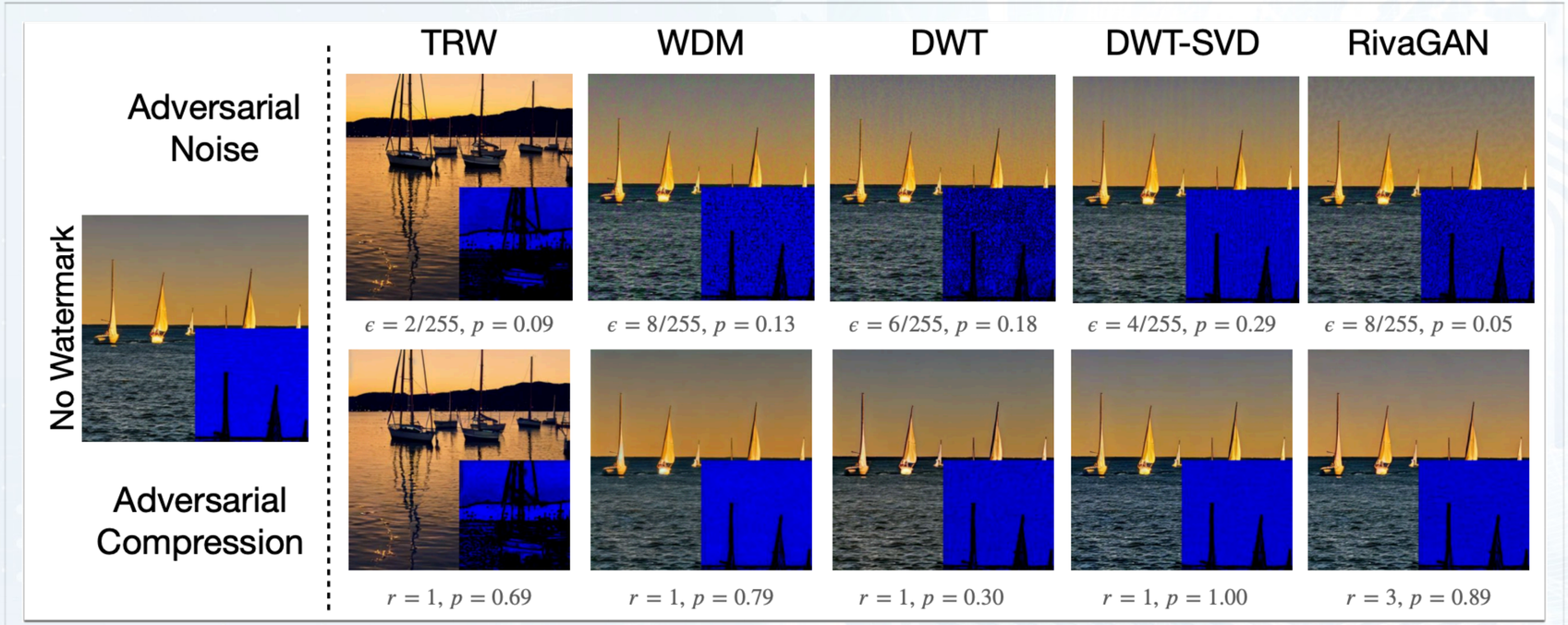
Adaptive versus Non-Adaptive Attacks

Mohamed bin Zayed University of Artificial Intelligence



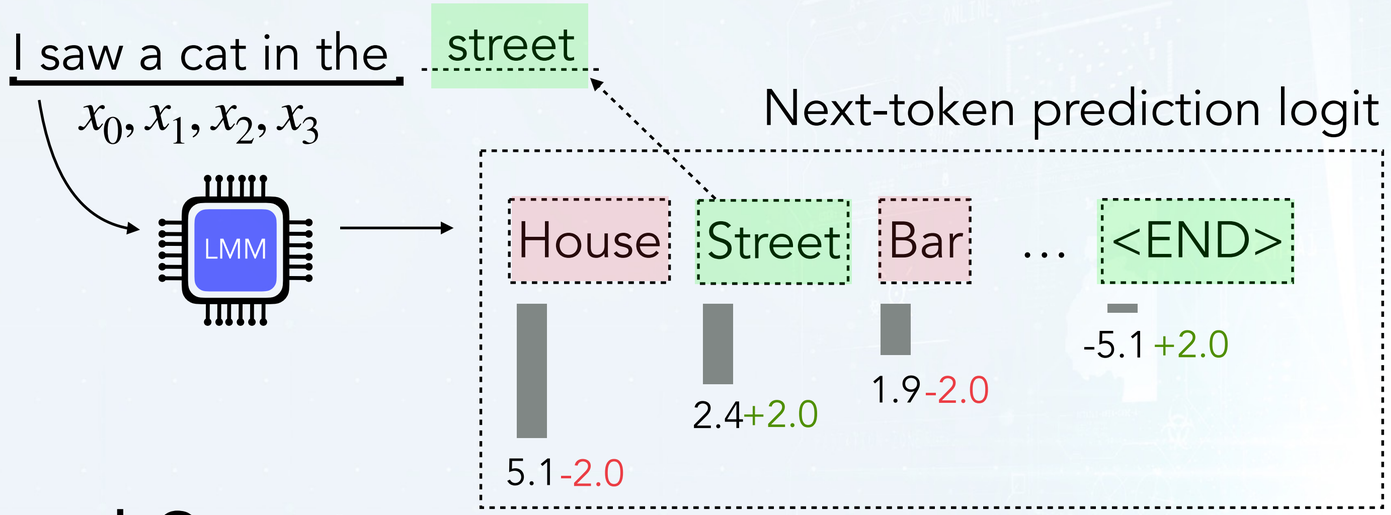
Other Modalities (e.g., Image, Video)

Mohamed bin Zayed University of Artificial Intelligence



Why do Adaptive Attacks Work?

Mohamed bin Zayed University of Artificial Intelligence



Analysis

1.) Avoid repeating sequences

2.) Watermark is most likely hidden in high-entropy text (e.g., names)

3.) Windowed context for seeding is vulnerable to paraphrase calibration (e.g., lexical diversity, restructuring)

Watermark Generate

Step 1.) Draw pseudo-random numbers using seed $f_\tau(x_0, \ldots, x_3)$

Step 2.) Partition vocabulary into green and red list

Step 3.) Bias logits to promote green list tokens

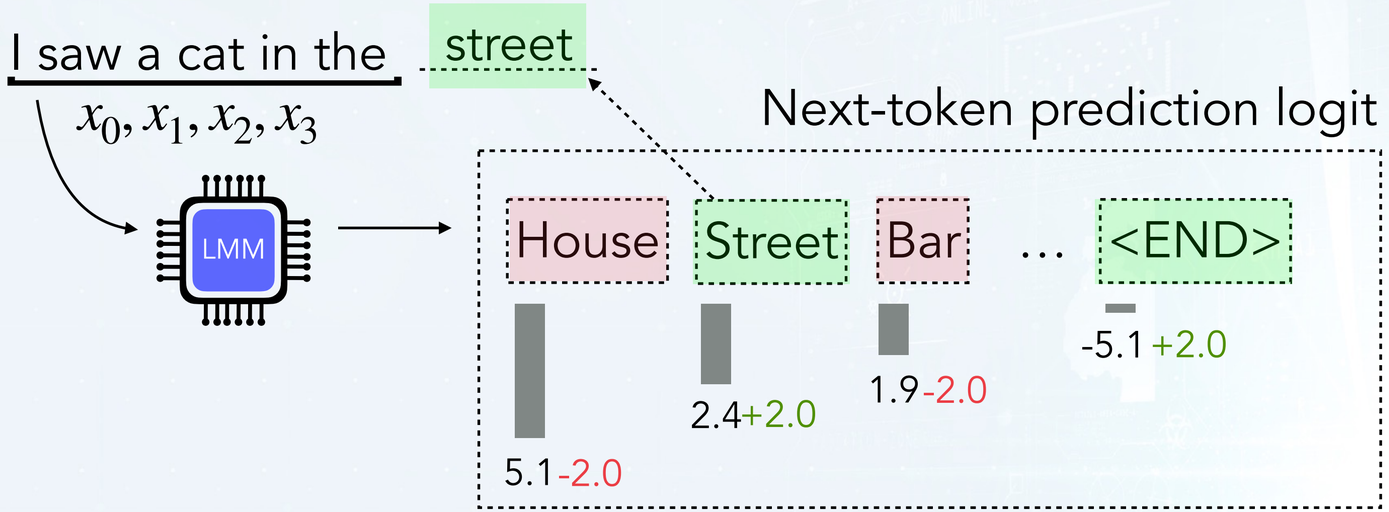
Step 4.) Softmax and sample

Step 5.) Repeat

Verify

Given a text x, and a secret key τ count green tokens and compare to expected value

Ideas for Improvement



Analysis

1.) Avoid repeating sequences

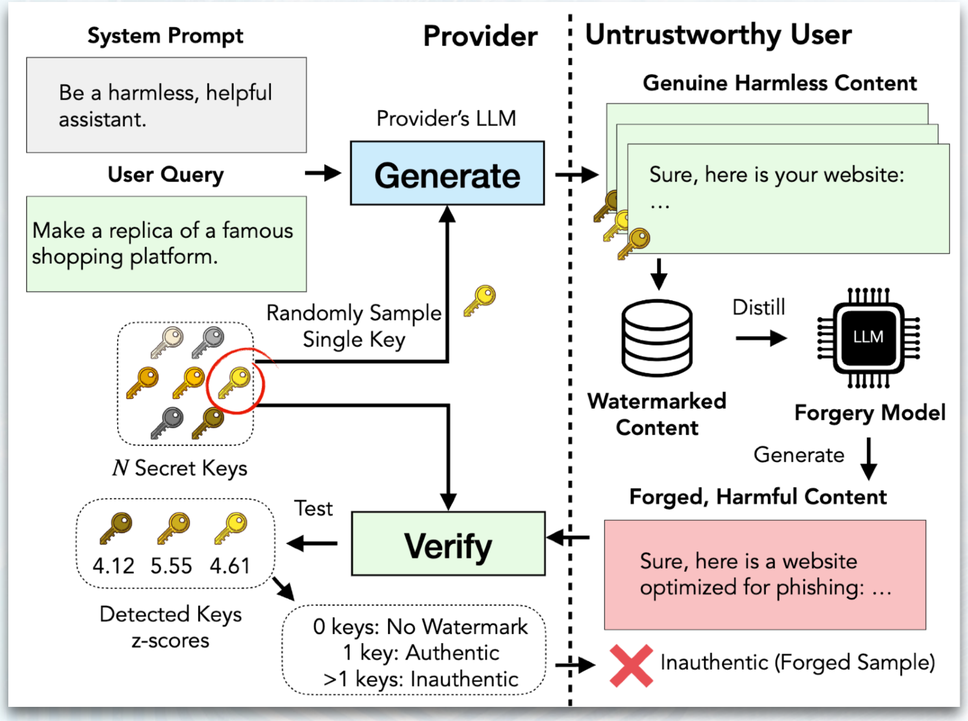
2.) Watermark is most likely hidden in high-entropy text (e.g., names)

3.) Windowed context for seeding is vulnerable to paraphrase calibration (e.g., lexical diversity, restructuring)

Forgery-Resistant Watermarking

Mohamed bin Zayed University of Artificial Intelligence

Mitigating Watermark Stealing Attacks in Language Models via Multi-Key Watermarking



Definition 3.6 (Unforgeability [72]). A watermark is *unforgeable* if for all λ and polynomial-time algorithms $\mathcal{A}$,

$$\Pr_{\substack{\mathsf{gk},\mathsf{ak} \\ x \leftarrow \mathcal{A}^{\mathsf{Watermark}^{\mathcal{M}}_{\mathsf{gk}}}(1^\lambda,\mathsf{ak})}} [\mathsf{Attribute}_{\mathsf{ak}}(x) \rightarrow \text{true and } x \notin \mathcal{Q}] \leq \mathsf{negl}(\lambda),$$

where $\mathcal{Q}$ denotes the set of responses obtained by $\mathcal{A}$ on its queries to the watermarked model.

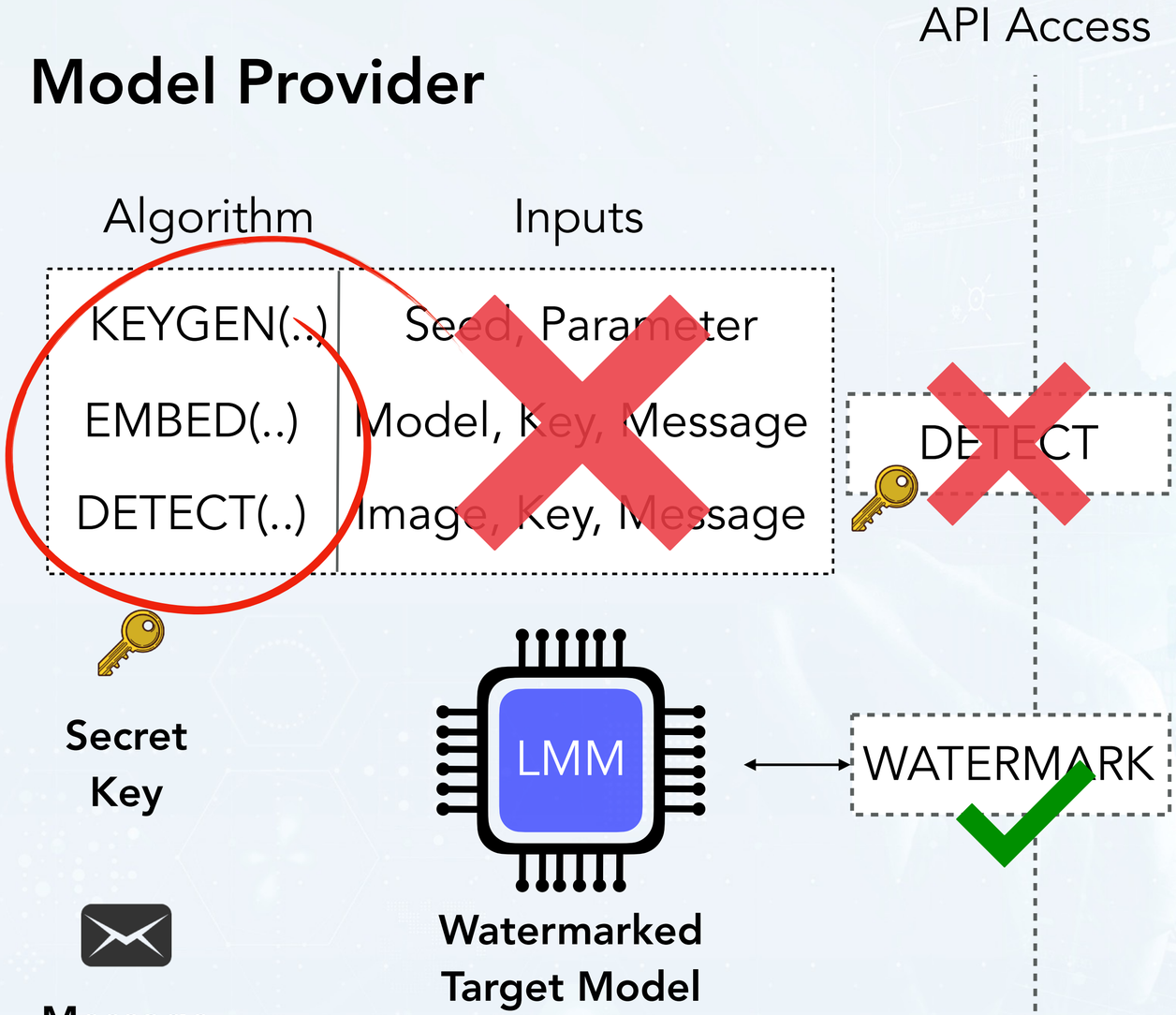
Detect ✓

Not Generated By Provider ✓

Forgery - Threat Model

Mohamed bin Zayed University of Artificial Intelligence

Model Provider



Adversary

~~**No-box**: No access to the target model~~

Offline: No access to VERIFY

Private: No access to the secret key or randomness

Computationally bounded: Cannot train own LLM

Adaptive: Knows watermarking scheme (but not the inputs used by the provider)

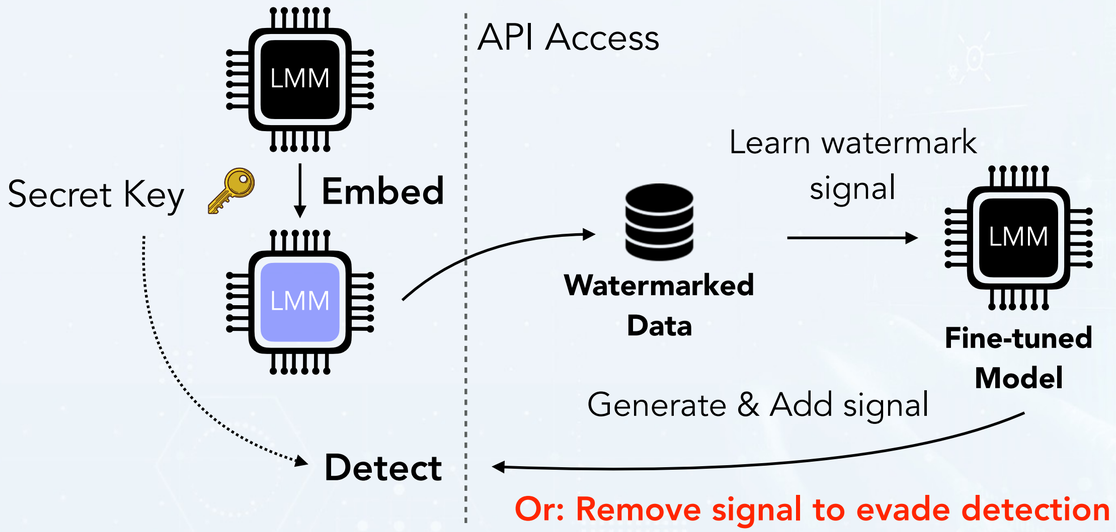
Surrogate Model: Can access less capable, open-source models

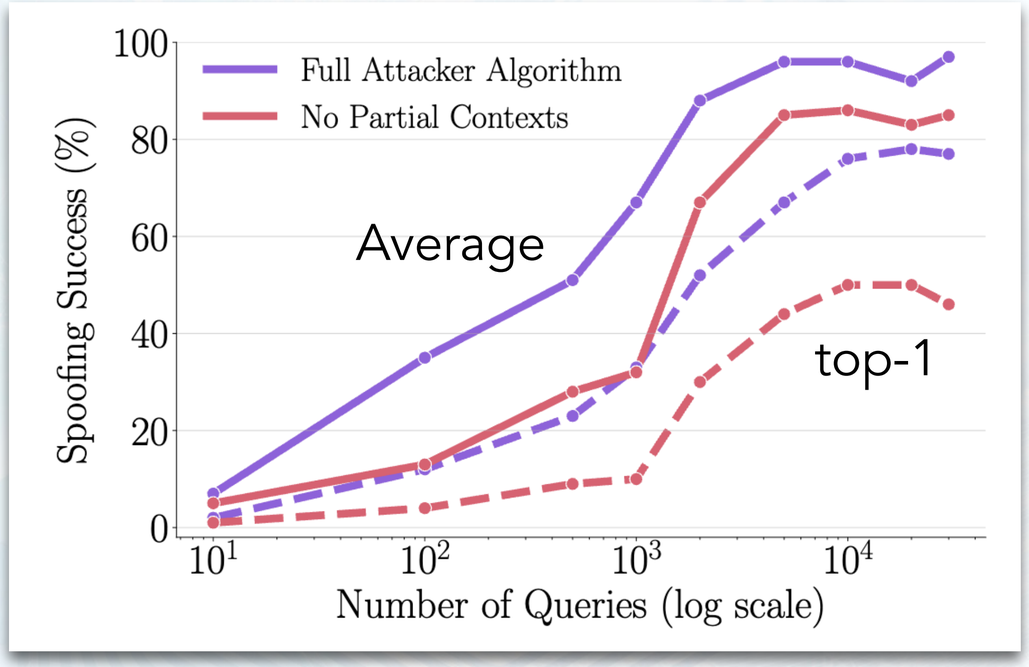
Harmful Content: Produce content that cannot be Produced by the provider (e.g., due to safety filter)

Hugging Face

Forgery Attacks

Mohamed bin Zayed University of Artificial Intelligence



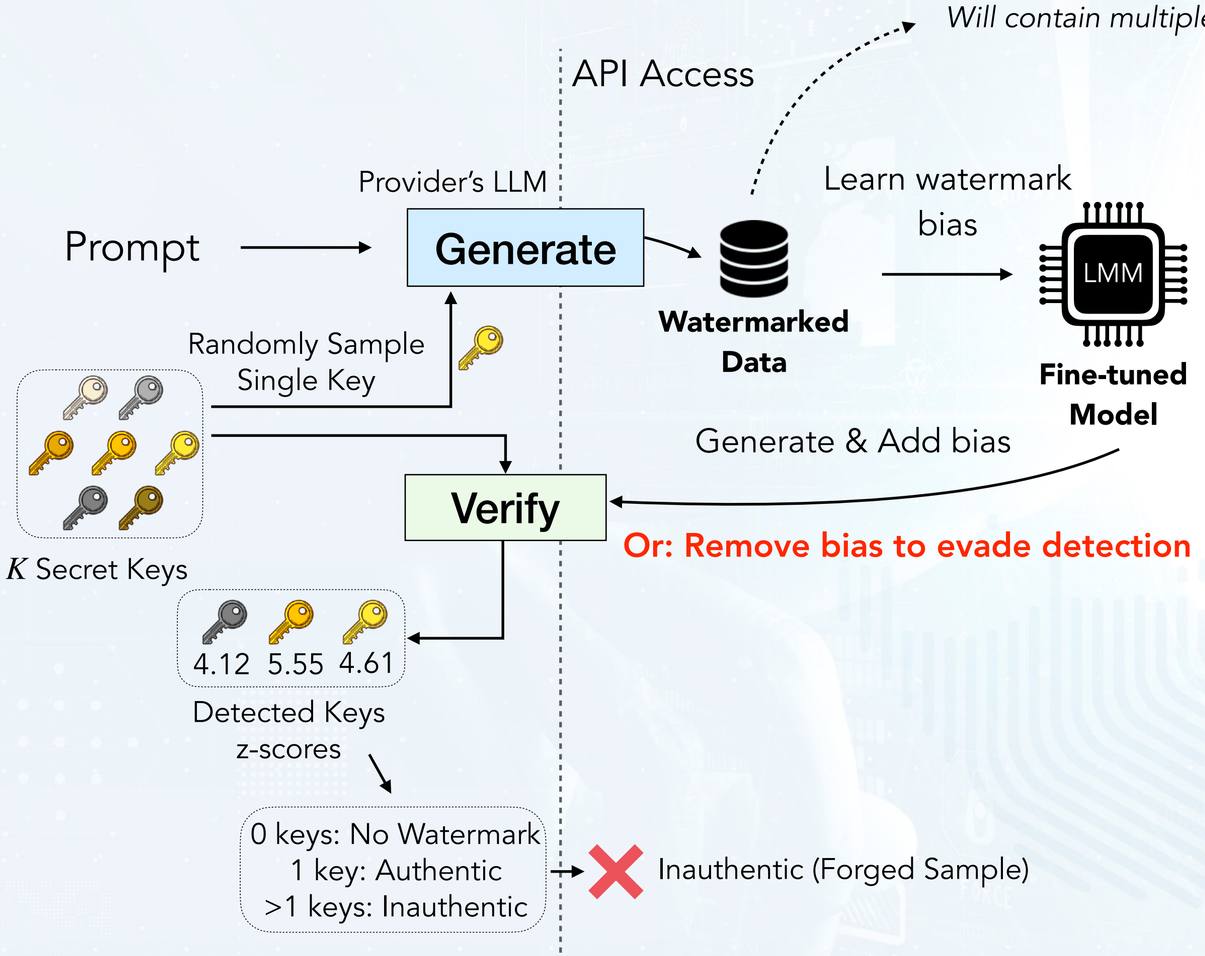


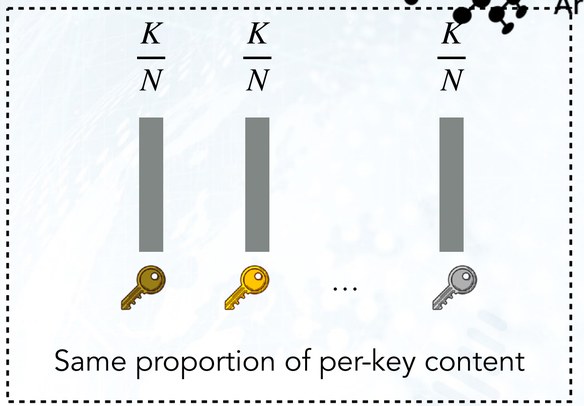
Attacker's success typically scales with the number of queries

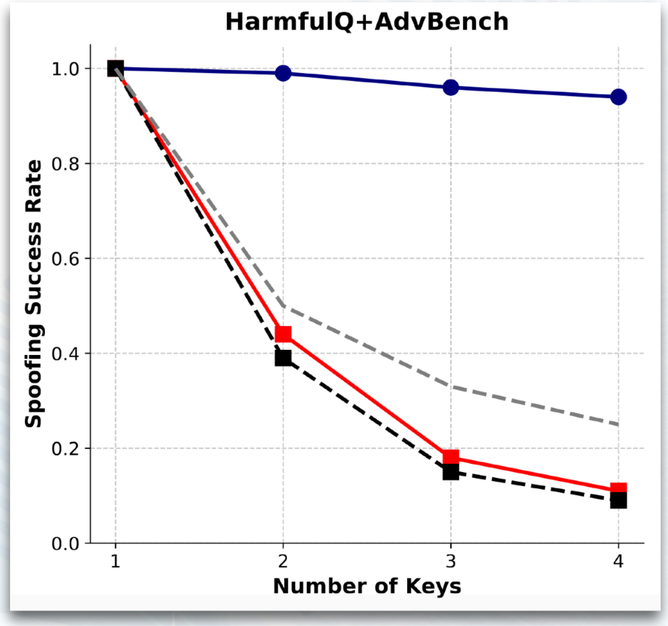
e.g., Watermark Stealing in Large Language Models, Jovanović et al., 2024

Randomized Key Selection

Mohamed bin Zayed University of Artificial Intelligence





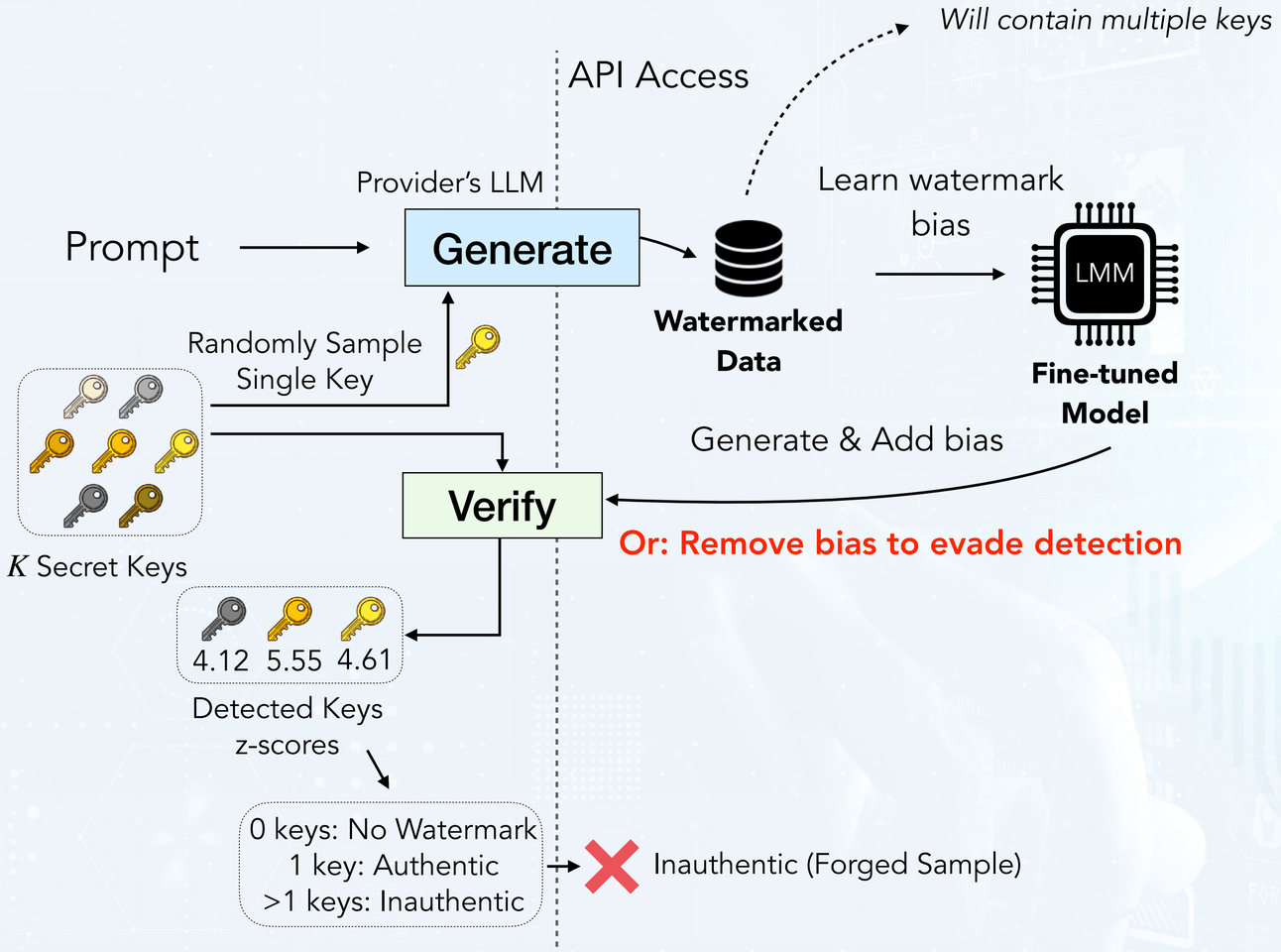


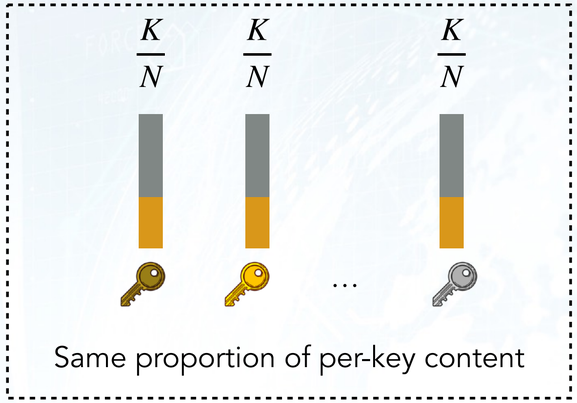
N=10,000 samples

Adaptive Attacks against Our Method

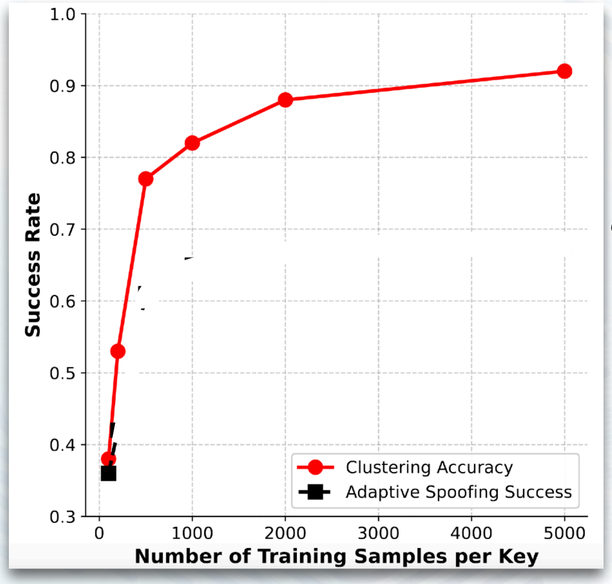
Mohamed bin Zayed University of Artificial Intelligence

Can adaptive attackers fare better?





Attacker knows labels for a subset of watermarked content



68%

Adaptively Robust and Forgery-Resistant Watermarking

Nils Lukas
Assistant Prof. @ML department

Thank you for your attention!