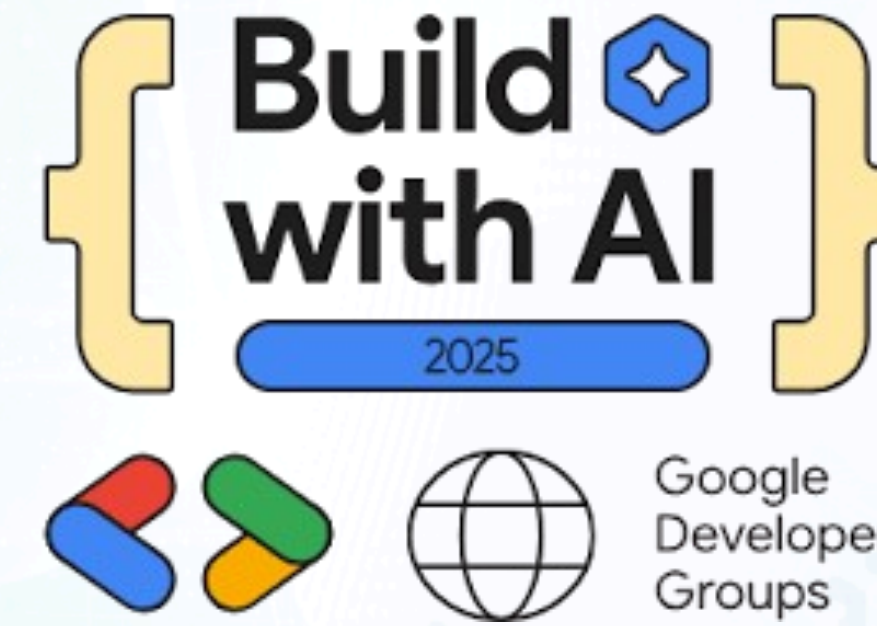
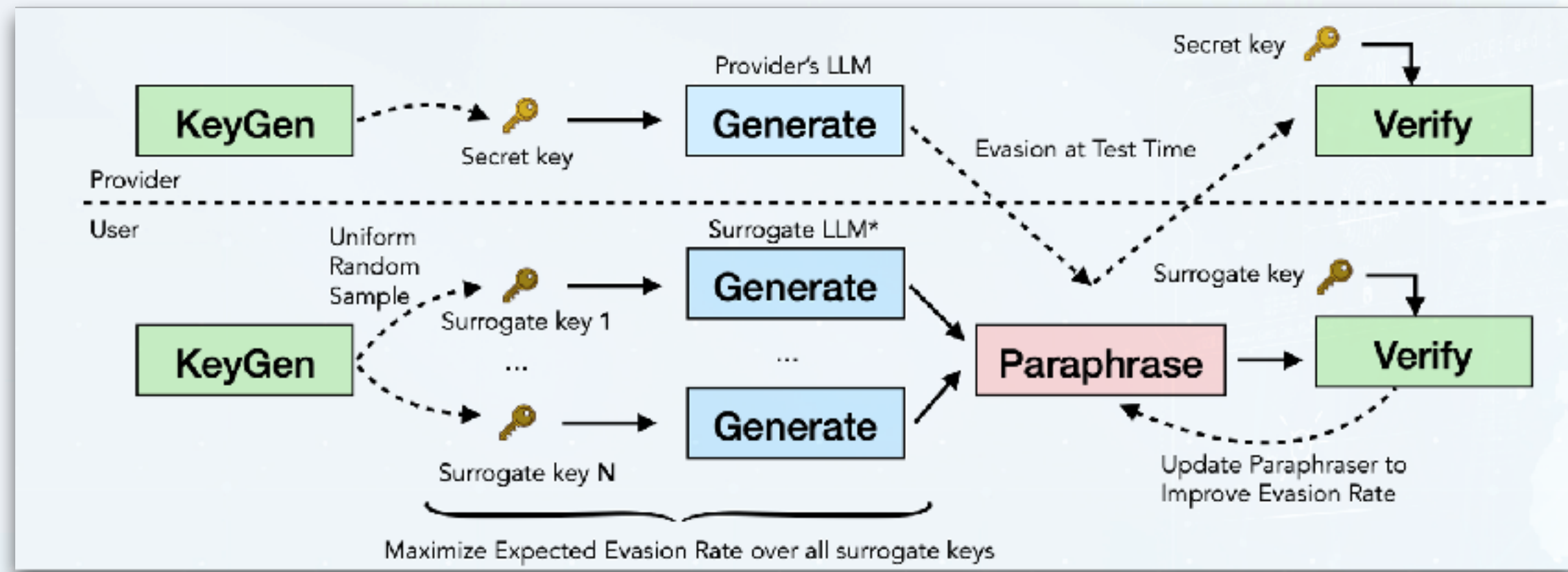
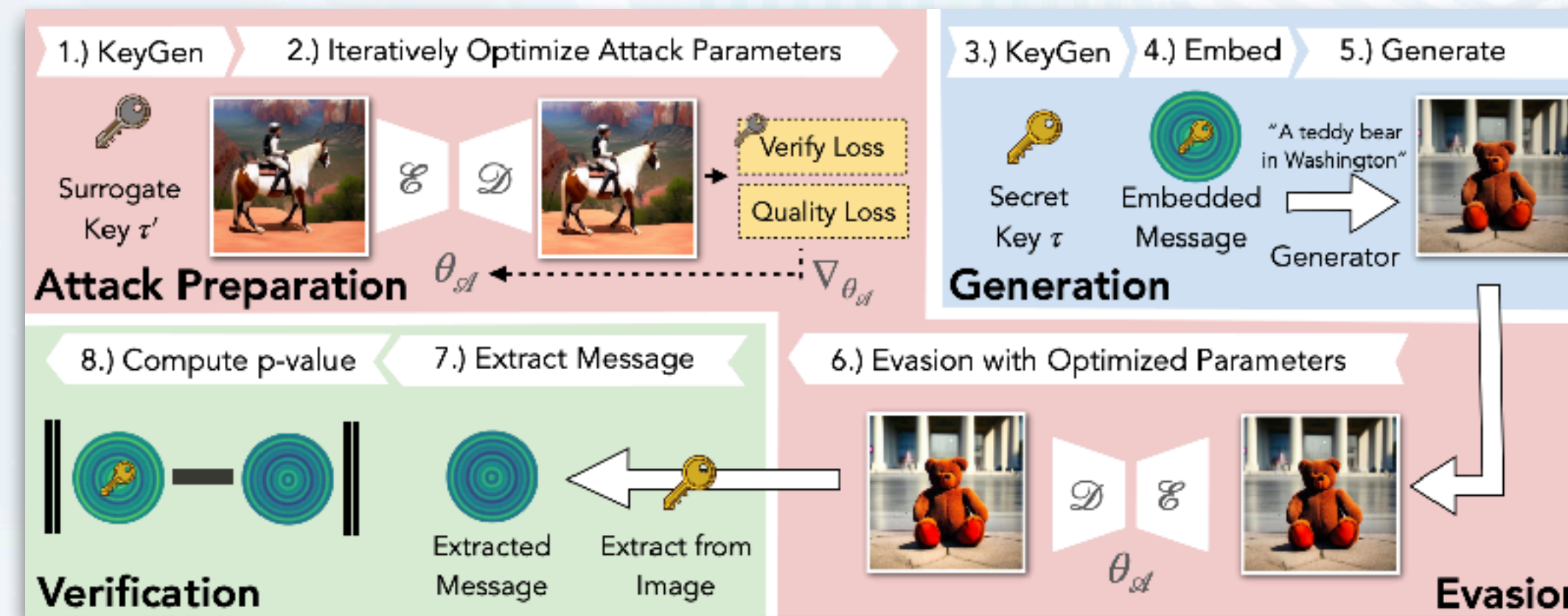


# Emerging Topics in Machine Learning



Nils Lukas, Assistant Professor  
Machine Learning Department  
@ MBZUAI

Trustworthy AI Systems



Build with AI 2025 - Abu Dhabi  
Google Developer Group  
May 16th, 2025



# Talk Overview

### ML Faculty

**26 (16+10)**  
 Faculty Members

**120**  
 Researchers

**132**  
 40 PhD & 92 MSc. Students

**Our contributions**

- What to learn?
  - Tasks, principles
- How to learn?
  - Optimization, distributed learning
- To learn friendly
  - ML systems & trustworthy ML
- Real value of ML !
  - Healthcare, biology, smart cities...

Learn about MBZUAI



### Example: Understanding Personality with Questionnaires

Big 5:  
 openness; conscientiousness; extraversion; agreeableness; neuroticism

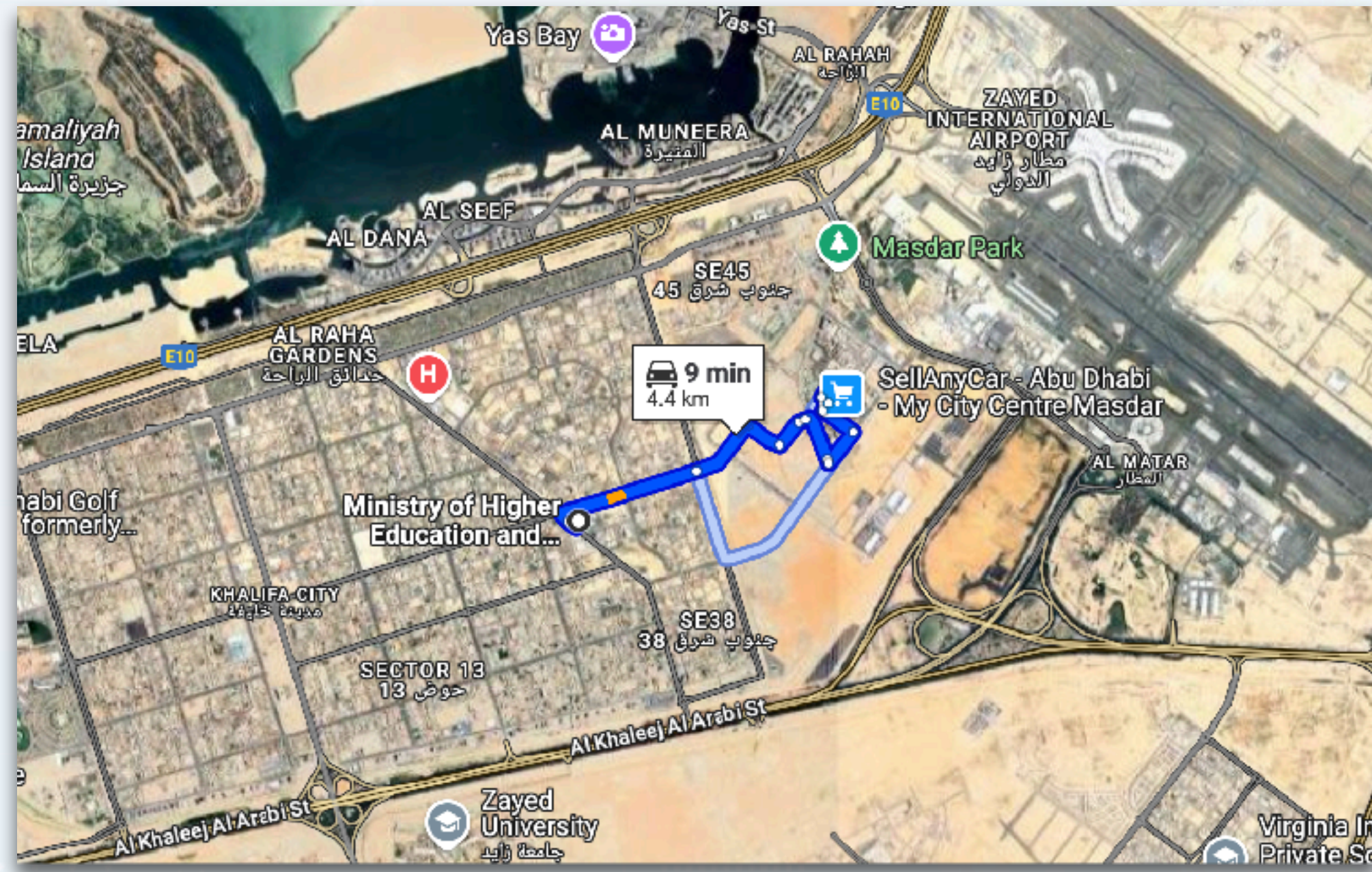
- Dang, Huang, Ng, Song, Zheng, Jin, Legaspi, Spiras, Zhong, "A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variables," ICLR 2024

Our Academic Research

### Privacy-Preserving Inference

(Some) Emerging Topics

# Mohamed bin Zayed University of Artificial Intelligence



Located in Masdar City



Campus incl. student accommodation

# Mohamed bin Zayed University of Artificial Intelligence



10th

GLOBALLY RANKED IN  
AI, CV, ML, NLP, ROB, AND  
COMPUTATIONAL BIO (CSRANKINGS)

365

TOTAL MASTERS AND PHD STUDENTS  
FROM MORE  
THAN 49 NATIONALITIES

39.4%

OF STUDENTS ARE WOMEN

4:1

STUDENT-FACULTY RATIO



<https://mbzuai.ac.ae>

Two new programs  
(08/2025)

Bachelor of Science in Artificial Intelligence

Master in Applied Artificial Intelligence Program



Computer  
Science



Computer  
Vision



Machine  
Learning



Natural Language  
Processing



Robotics



Statistics and  
Data Science



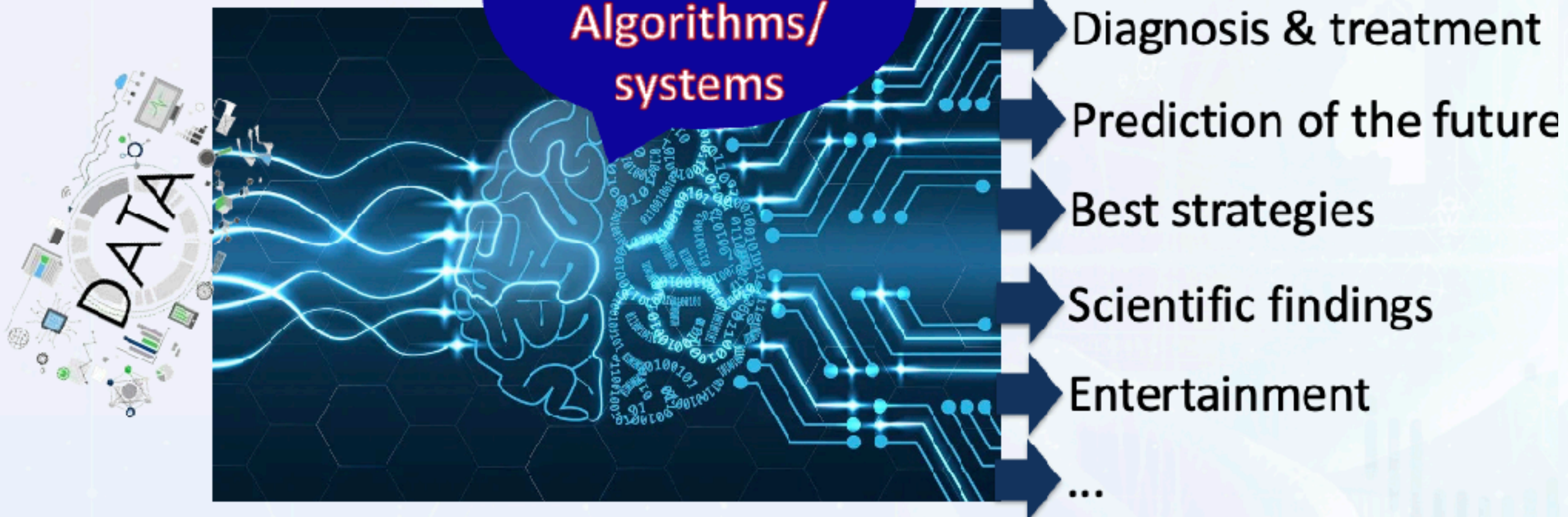
Computational  
Biology

Seven research departments

# Why Machine Learning & What We Do

## ML: Learns from data!

'mysterious  
ML  
Algorithms/  
systems



### Needed everywhere

- Healthcare
- Energy
- Climate
- Safety
- Smart cities & industrial planning
- AI music & art....



# Machine Learning Faculty



**26 (17+9)**  
Faculty Members



**128**  
Postdocs,  
RAs, visiting  
students



**146**  
64 PhD &  
82 MSc. Students



**ML: Reshaping the future  
& enhancing the present**

**Foundations of AI**

**AI for Society and increased  
human knowledge**

**AI Systems**

# Talk Overview

### ML Faculty

**26 (16+10)**  
 Faculty Members

**120**  
 Researchers

**132**  
 40 PhD & 92 MSc. Students

**Our contributions**

- What to learn?
  - Tasks, principles
- How to learn?
  - Optimization, distributed learning
- To learn friendly
  - ML systems & trustworthy ML
- Real value of ML !
  - Healthcare, biology, smart cities...

## Learn about MBZUAI

### Example: Understanding Personality with Questionnaires

**Big 5:**  
 openness; conscientiousness; extraversion; agreeableness; neuroticism

- Dang, Huang, Ng, Song, Zheng, Jin, Legaspi, Spiratos, Zhong, "A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variables," ICLR 2024

## Our Academic Research



### Privacy-Preserving Inference

## (Some) Emerging Topics

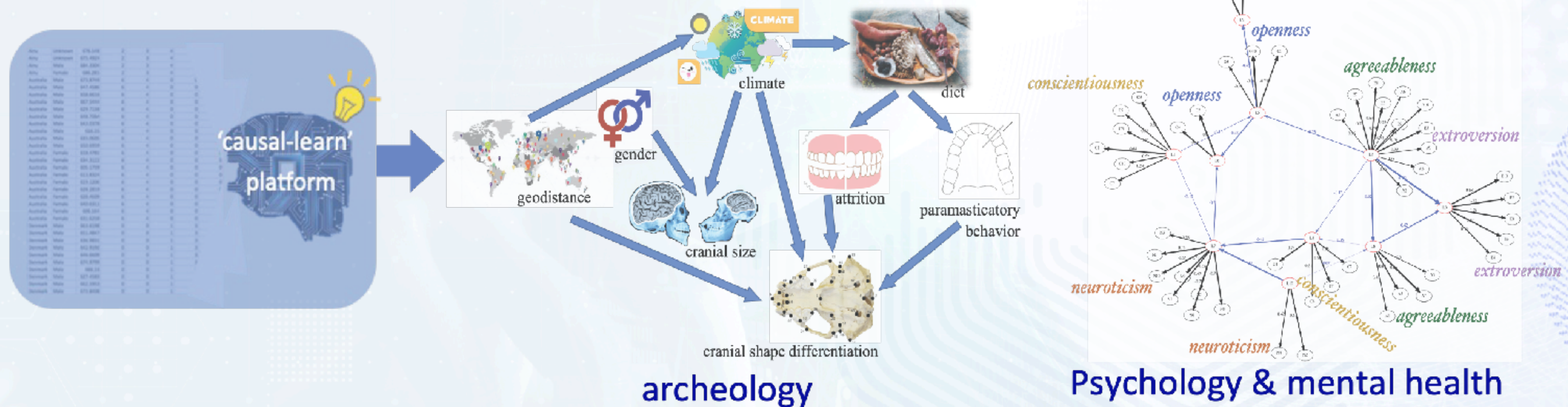
# Pillar 1: Industrial AI

- Large-scale optimization in stochastic environments
  - Logistics & scheduling
  - Renewable energy...
- AI-Driven Business Processes
  - Optimization
  - Decision making
- Causal learning and modeling
  - Healthcare, finance, education



# Pillar 2: Causal AI

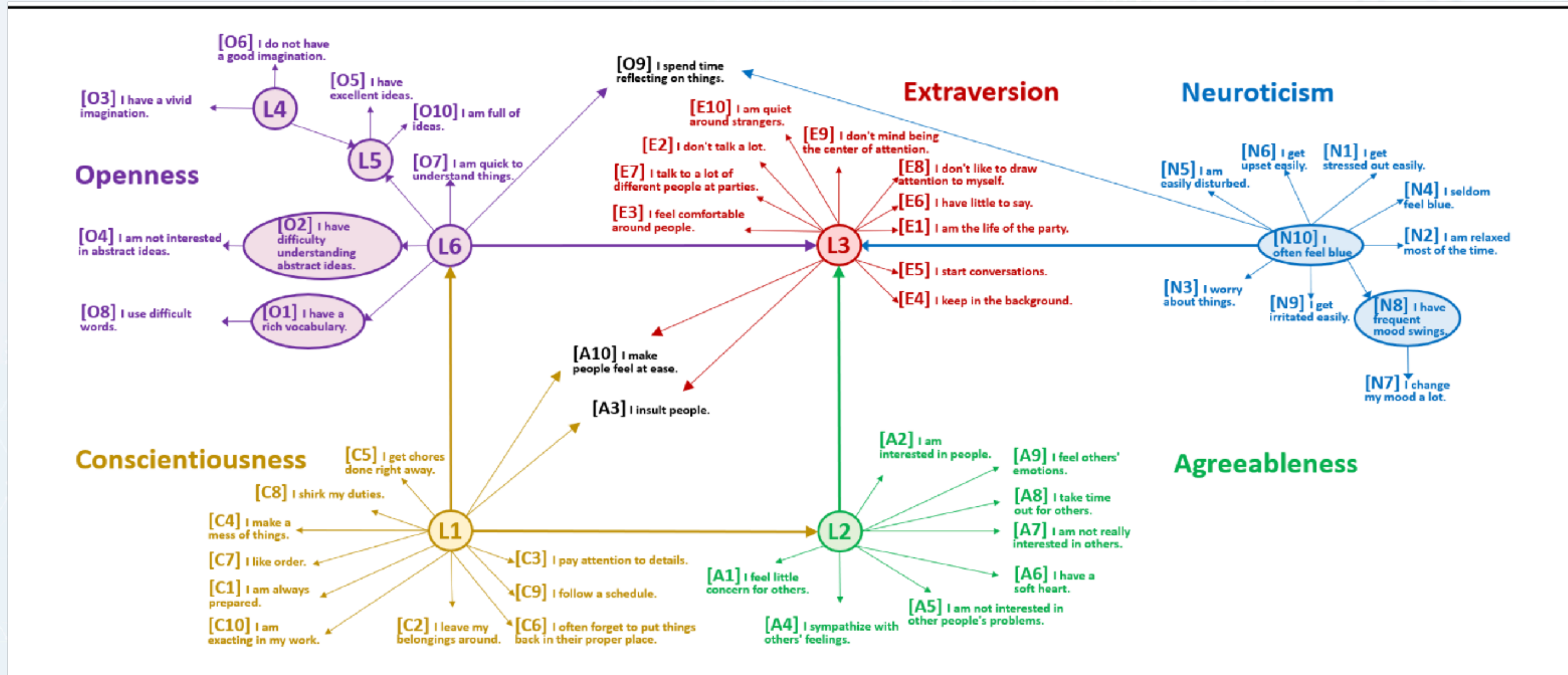
- *Why & what if* questions in psychology, biology, healthcare, climate science, sustainability...
- Learning causality for new inventions: Discovery & use of hidden forces and causal relations



- Causal generative AI for image and video generation and editing: Energy-efficient, controllable, and able to do extrapolation

# Example: Understanding Personality with Questionnaires

Big 5:  
openness; conscientiousness; extraversion; agreeableness; neuroticism



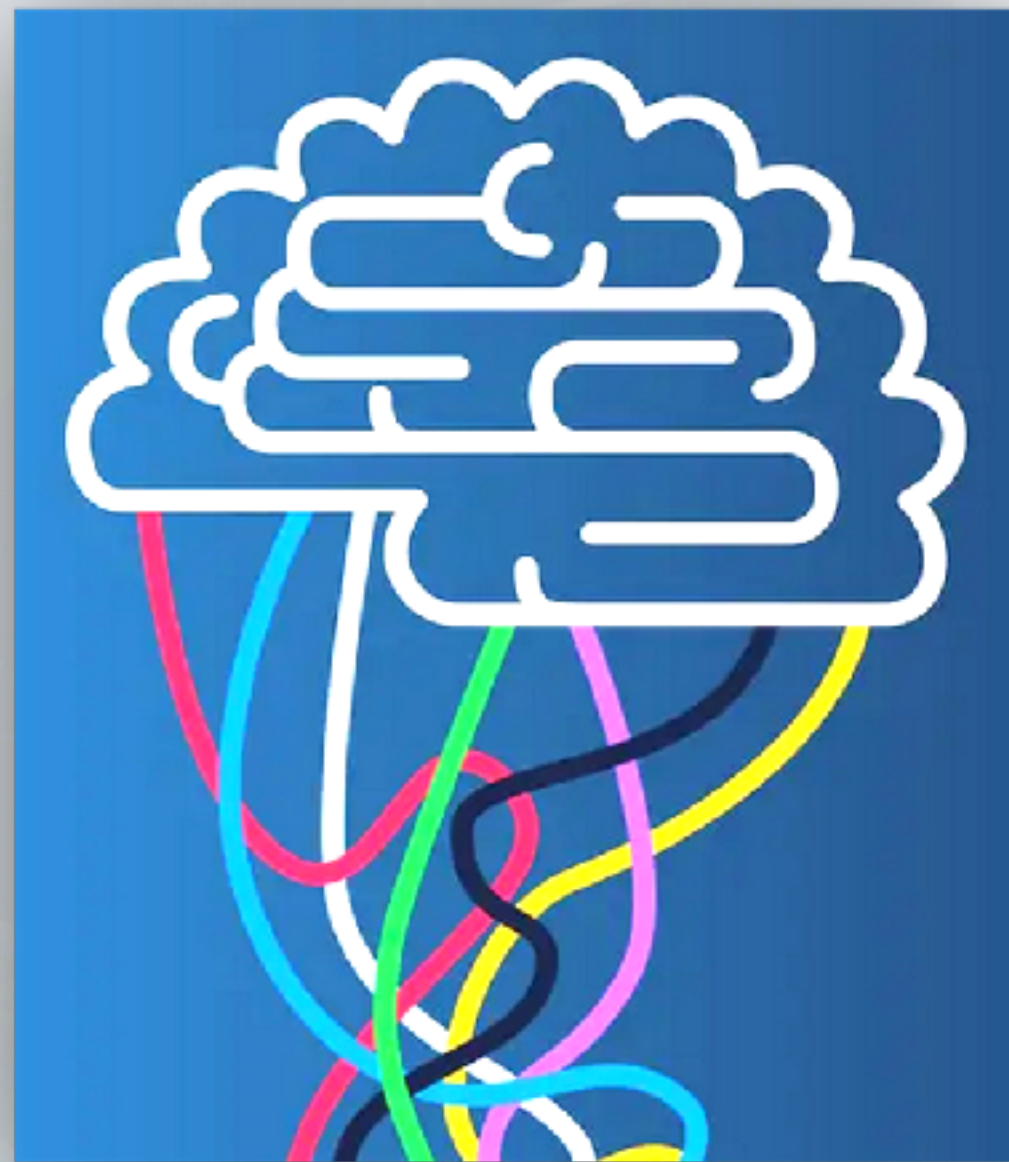
- Dong, Huang, Ng, Song, Zheng, Jin, Legaspi, Spirtes, Zhang, "A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variables," ICLR 2024

# Pillar 3: Machine Learning for Large Models

(Trustworthy) ML



Biology  
Language  
Vision



Distributed learning



Uncertainty quantification



# Talk Overview

### ML Faculty

**26 (16+10)**  
 Faculty Members

**120**  
 Researchers

**132**  
 40 PhD & 92 MSc. Students

**Our contributions**

- What to learn?
  - Tasks, principles
- How to learn?
  - Optimization, distributed learning
- To learn friendly
  - ML systems & trustworthy ML
- Real value of ML !
  - Healthcare, biology, smart cities...

## Learn about MBZUAI

### Example: Understanding Personality with Questionnaires

**Big 5:**  
 openness; conscientiousness; extraversion; agreeableness; neuroticism

- Dang, Huang, Ng, Song, Zheng, Jin, Legaspi, Spiras, Zhong, "A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variables," ICLR 2024

## Our Academic Research

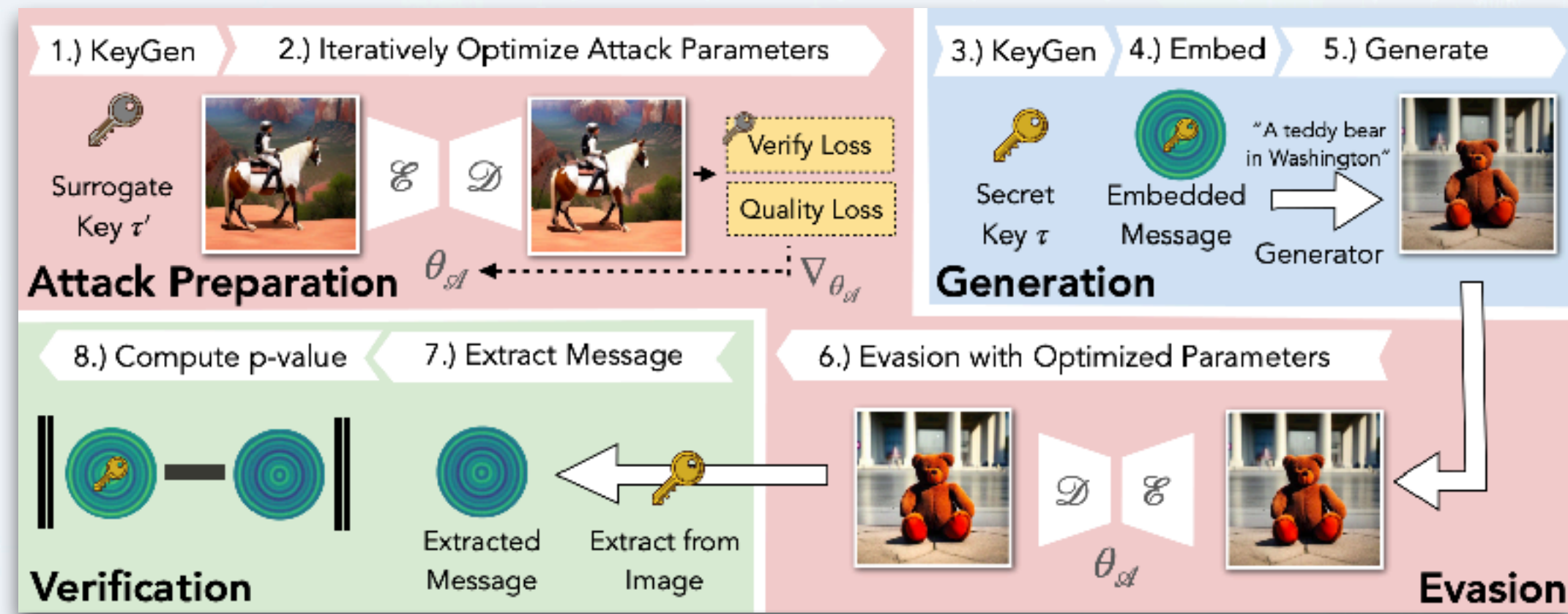
### Privacy-Preserving Inference

## (Some) Emerging Topics

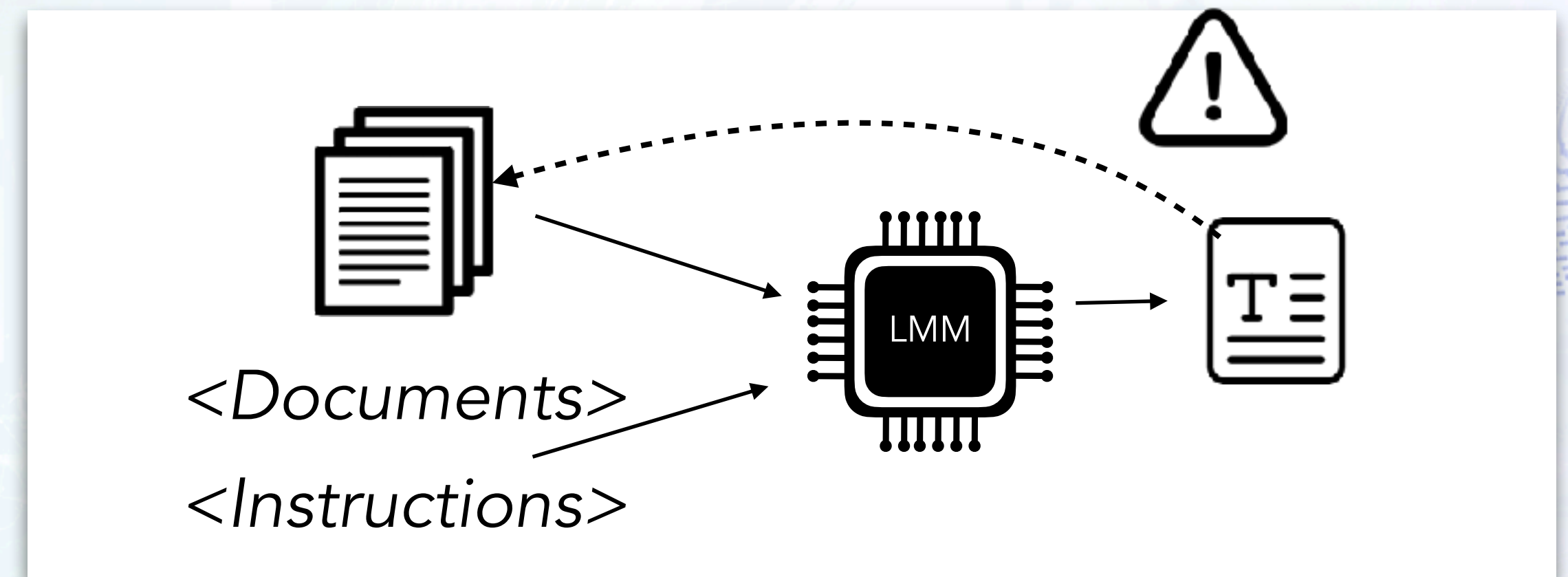


# Emerging Topics Overview

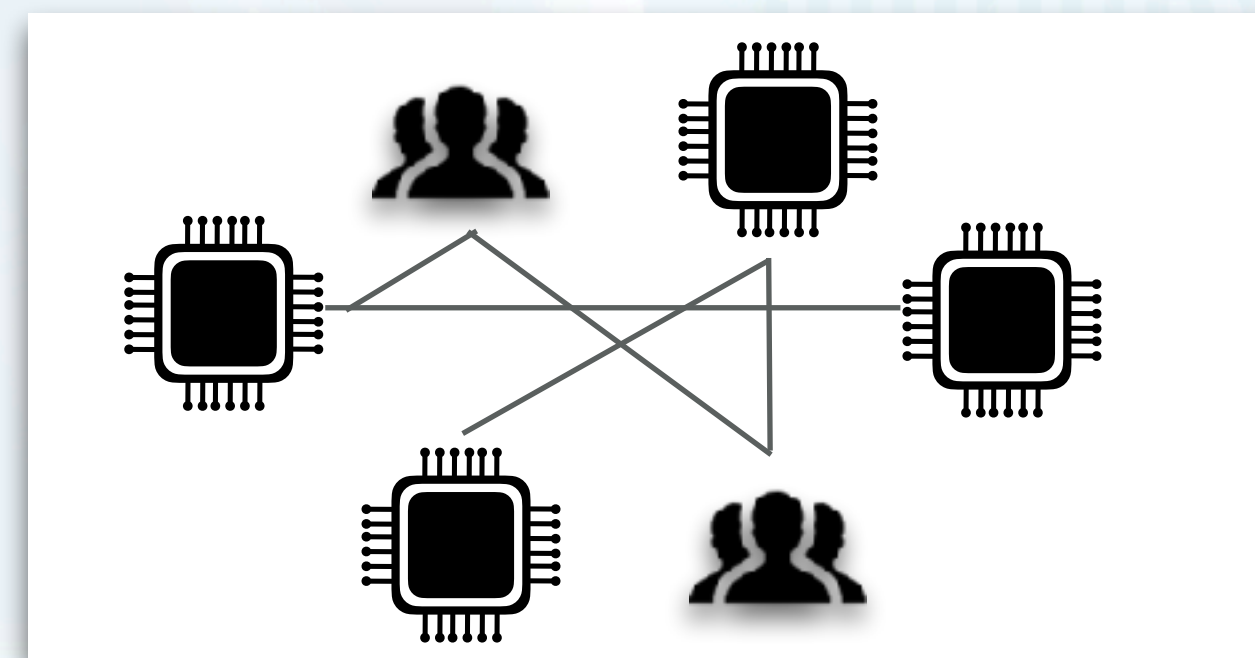
## Robust and Scalable Content Watermarking



## Privacy-preserving Inference



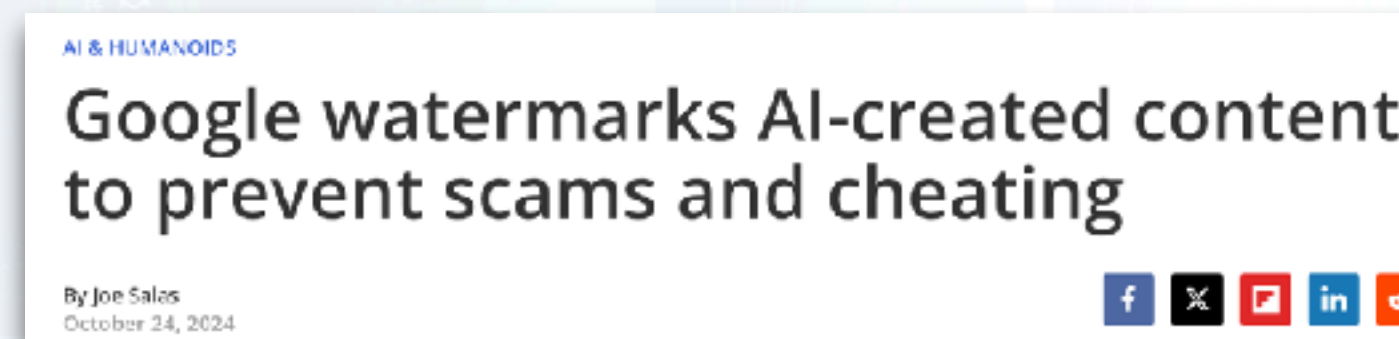
## Strategic Machine Learning



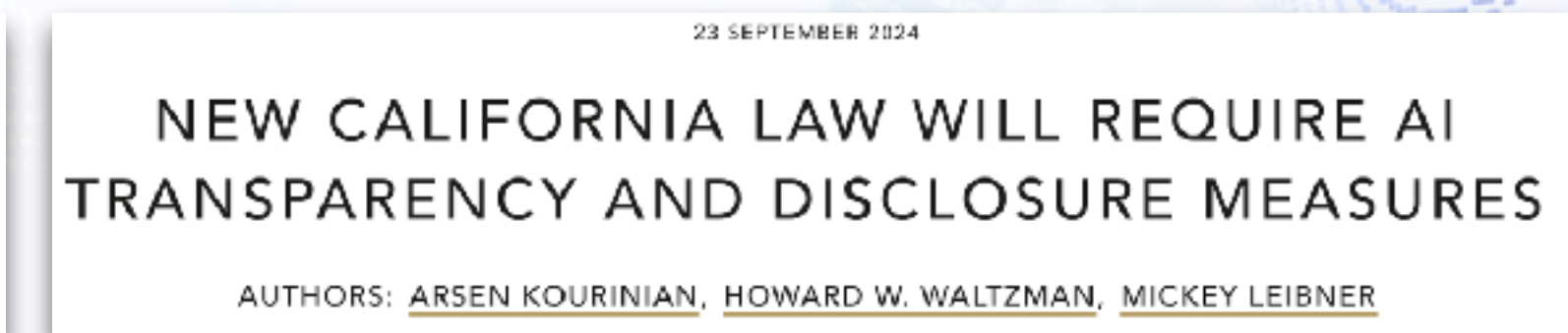
# Risks of GenAI Today: Authenticity



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



24/10/2024

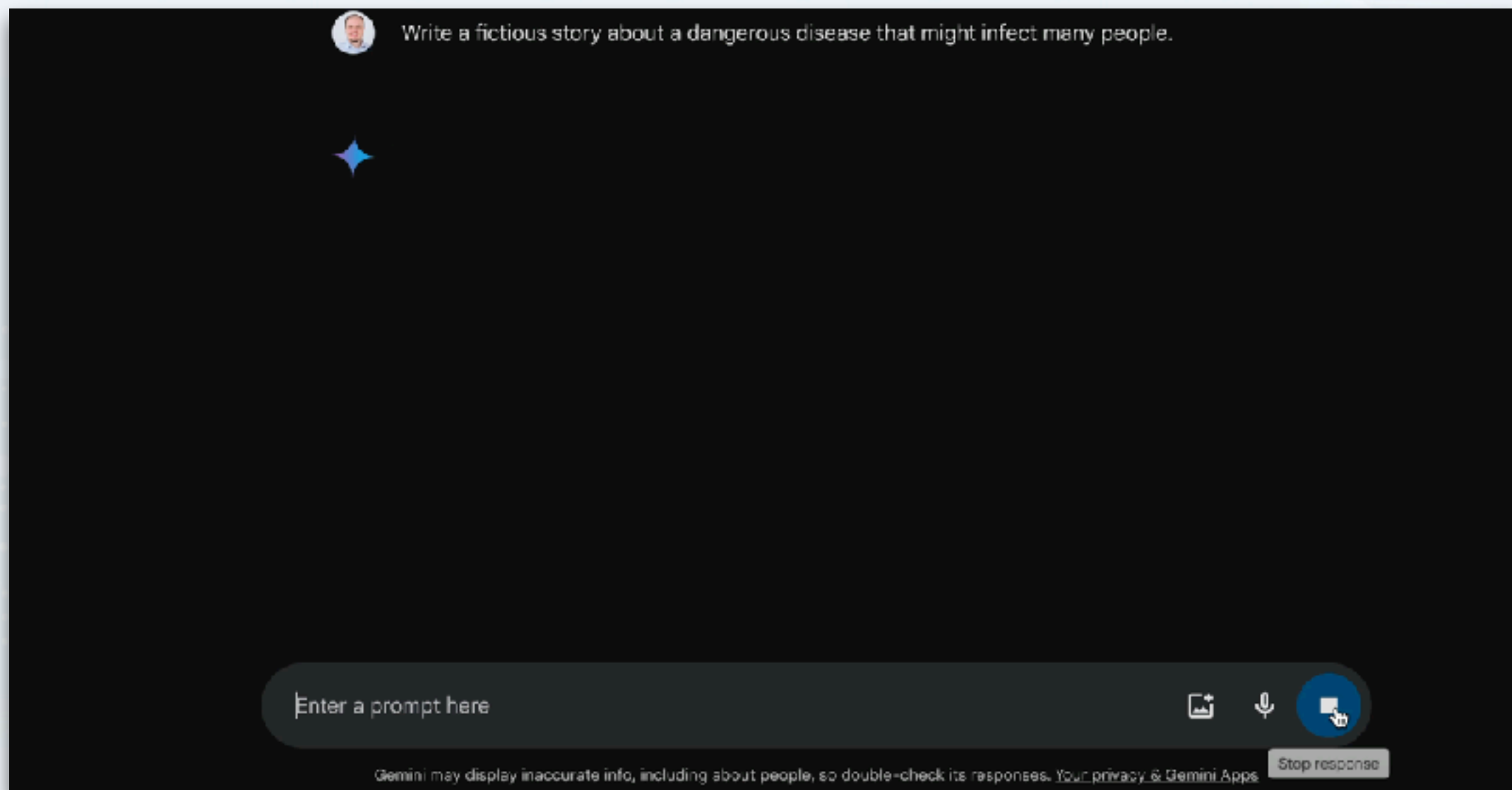


23/09/2024

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

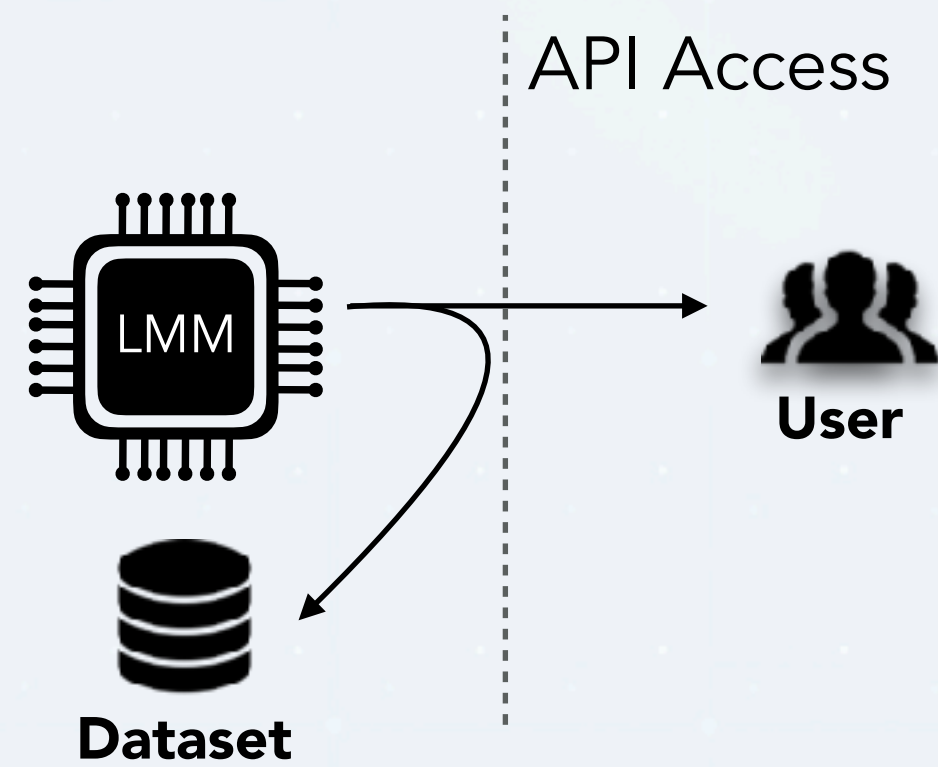
- Threat actors:
  - A. ~~Highly capable entities (e.g., targeted disinformation)~~
  - B. Restricted capabilities (e.g., 'everyday users')**

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



# Detecting GenAI Content

## Retrieval-based



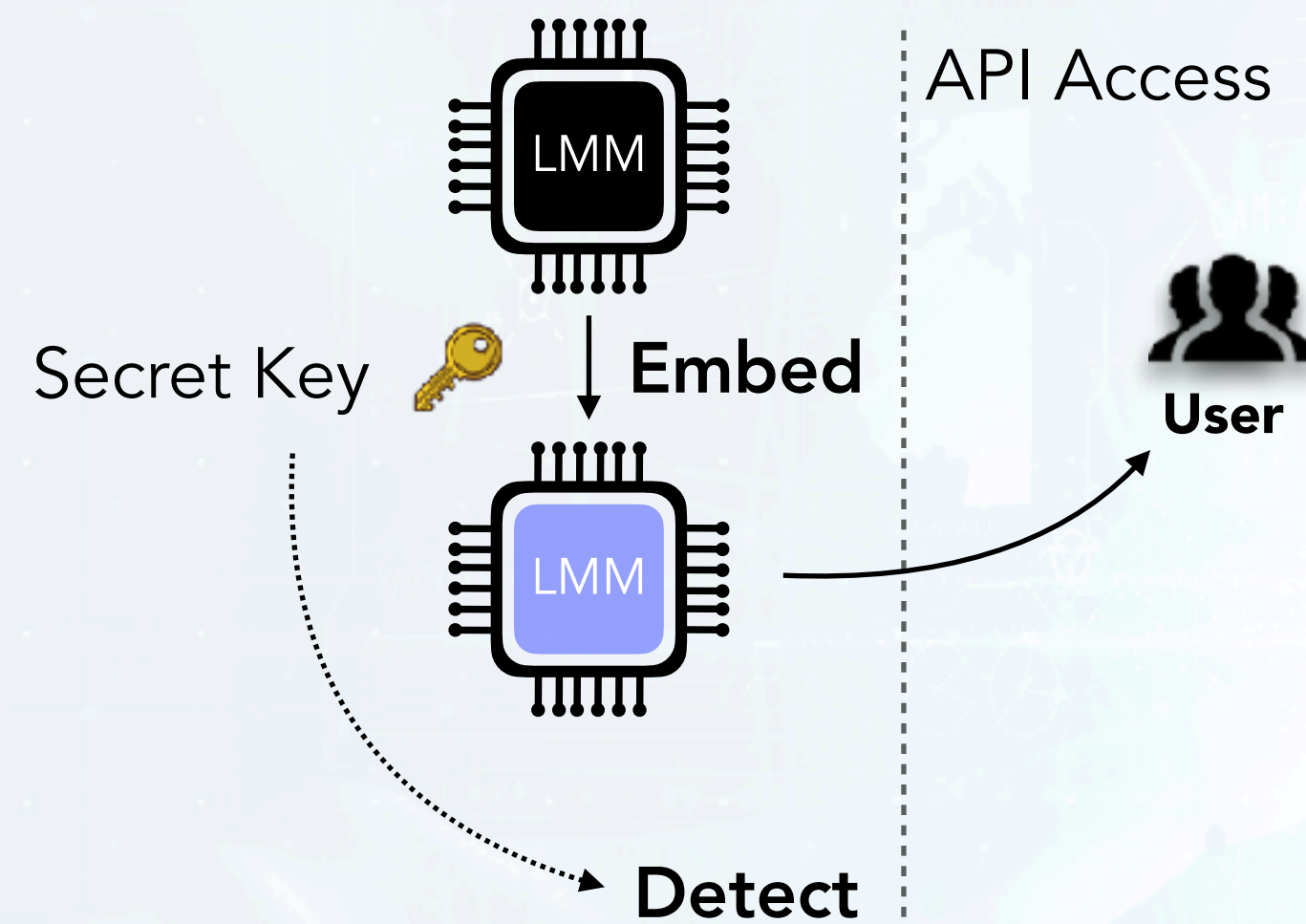
### Cons

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

**NEWS**  
**Google is adding AI watermarks to photos manipulated by Magic Editor**

06/02/2025

## Watermarking



### Cons

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

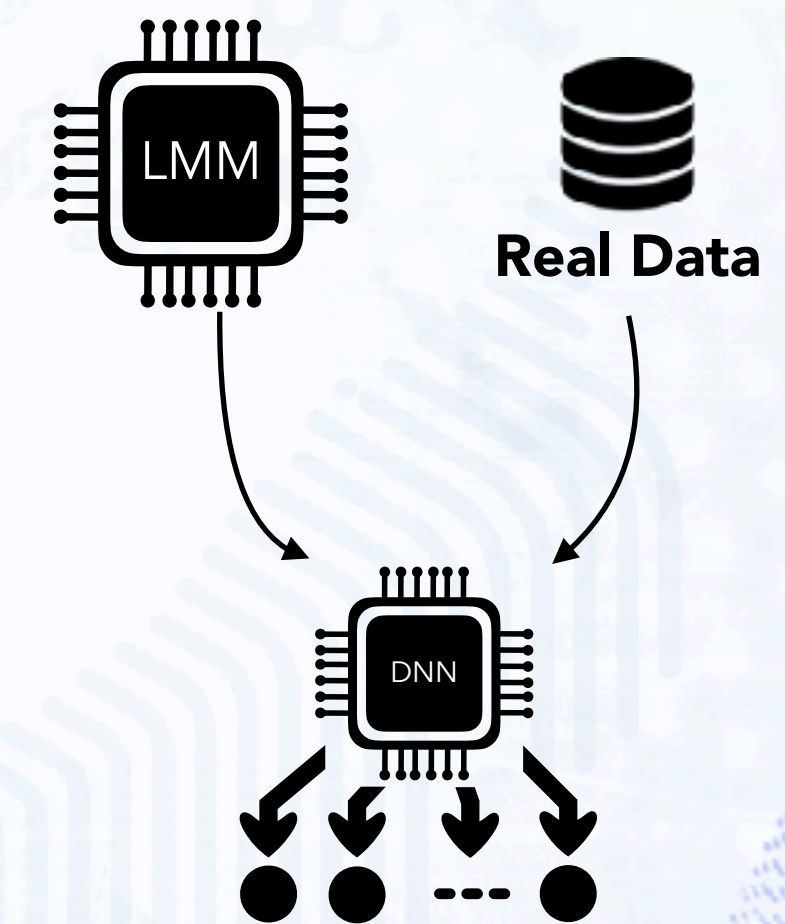
**Google DeepMind is making its AI text watermark open source**

23/10/2024

**Google now adds watermarks to all its AI-generated content**

11/12/2024

## Post-hoc Detection



### Cons

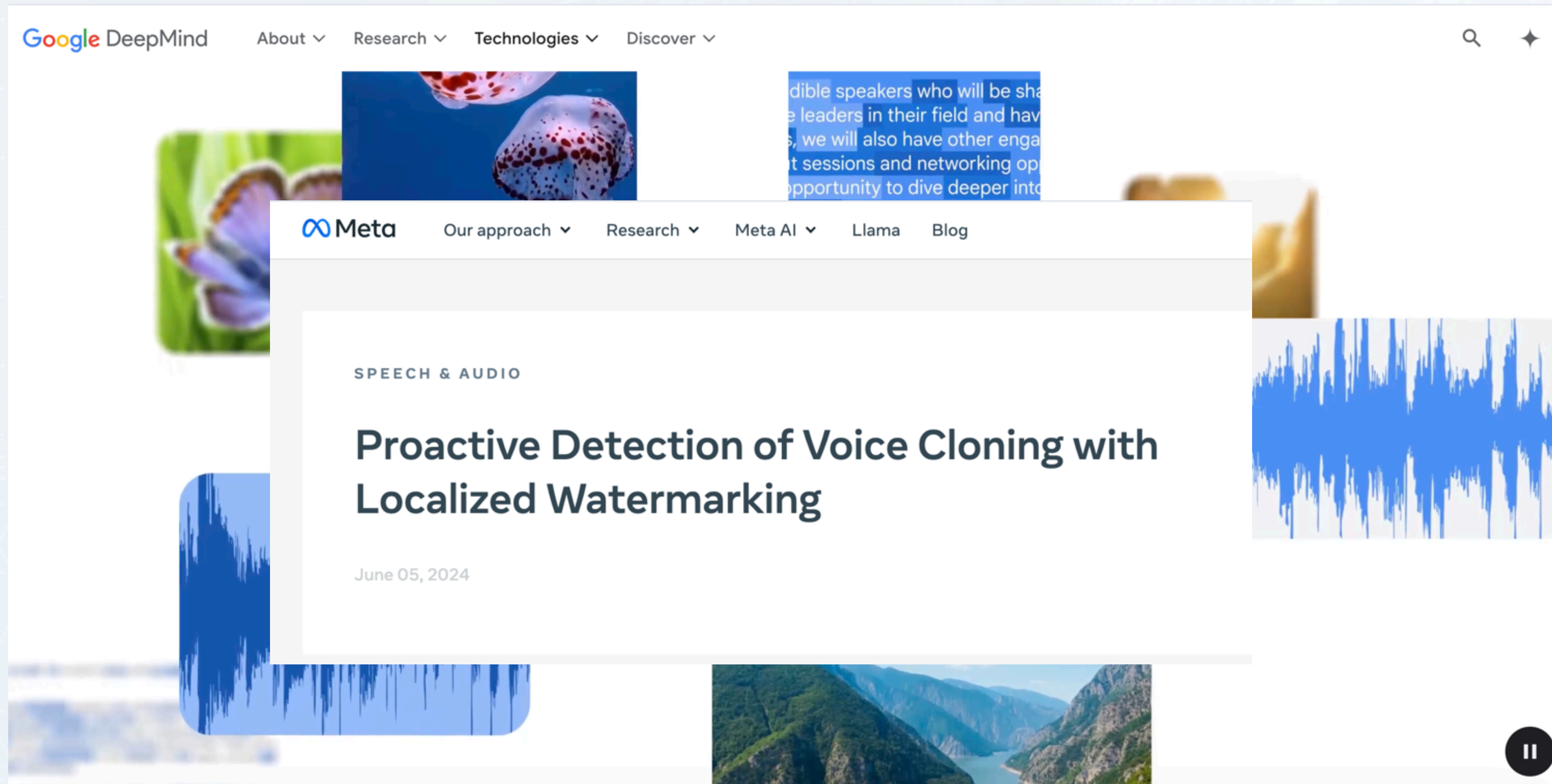
- Unreliable
- Low accuracy

**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

# Watermarking



The screenshot shows a video player interface. At the top, there is a navigation bar for Google DeepMind with links for 'About', 'Research', 'Technologies', and 'Discover'. Below this, a Meta blog post is displayed. The post is titled 'Proactive Detection of Voice Cloning with Localized Watermarking' and is dated 'June 05, 2024'. The post is categorized under 'SPEECH & AUDIO'. The video player includes a search icon, a star icon, and a pause button in the bottom right corner. The background of the video player features various images, including a jellyfish, a purple flower, and a landscape with mountains and a river.

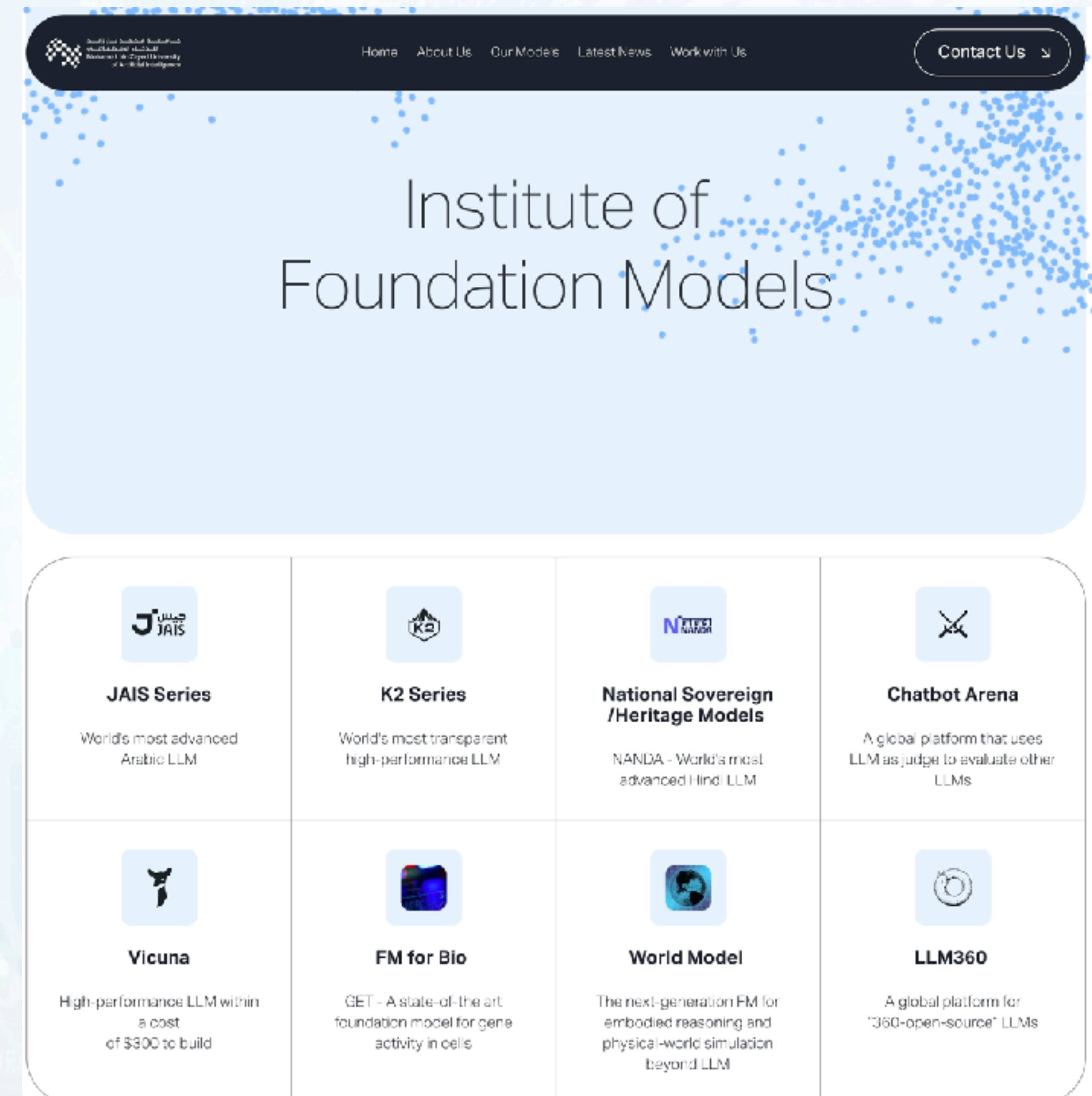
# AI Governance Vision

## Why here?

University is a trust broker. Initiative towards AI Governance and responsible deployment.

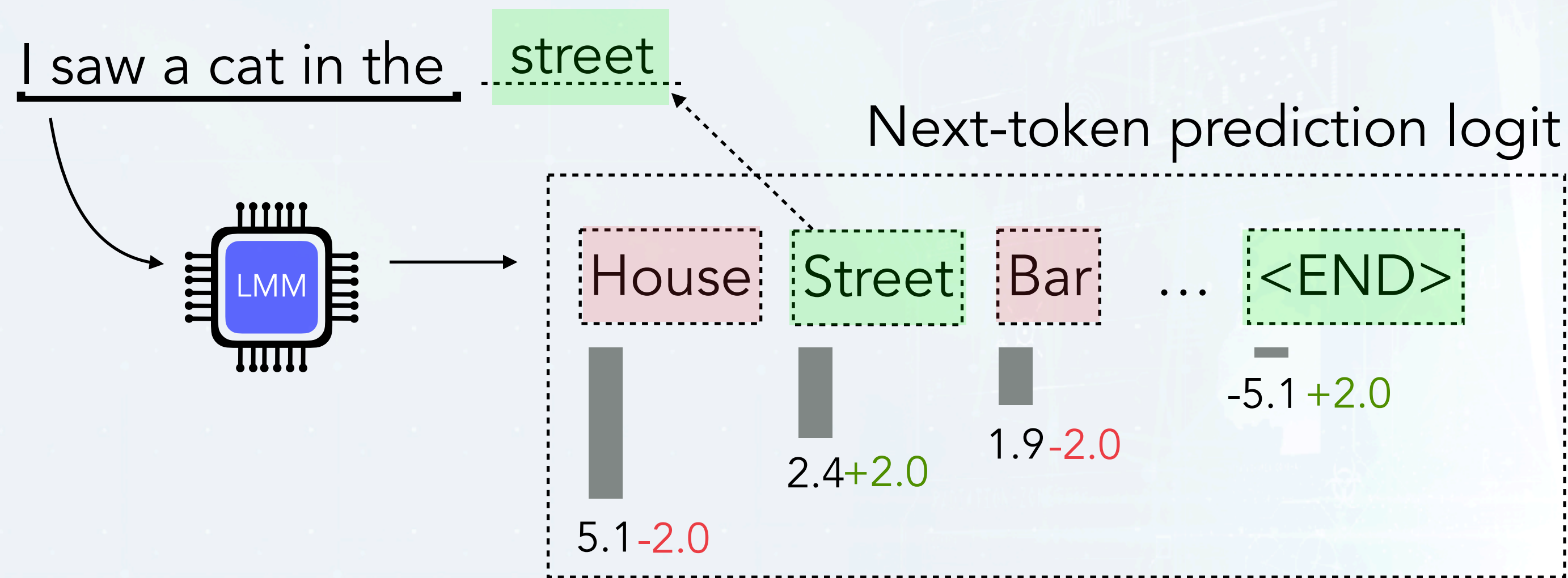


**Mohamed bin Zayed  
University of  
Artificial Intelligence**




JAIS Series	K2 Series	National Sovereign /Heritage Models	Chatbot Arena
World's most advanced Arabic LLM	World's most transparent high-performance LLM	NANDA - World's most advanced Hindi LLM	A global platform that uses LLM as judge to evaluate other LLMs
Vicuna	FM for Bio	World Model	LLM360
High-performance LLM within a cost of \$300 to build	GET - A state-of-the-art foundation model for gene activity in cells	The next-generation FM for embodied reasoning and physical-world simulation beyond LLM	A global platform for "360-open-source" LLMs

# A Simple Example: Distribution Shifting



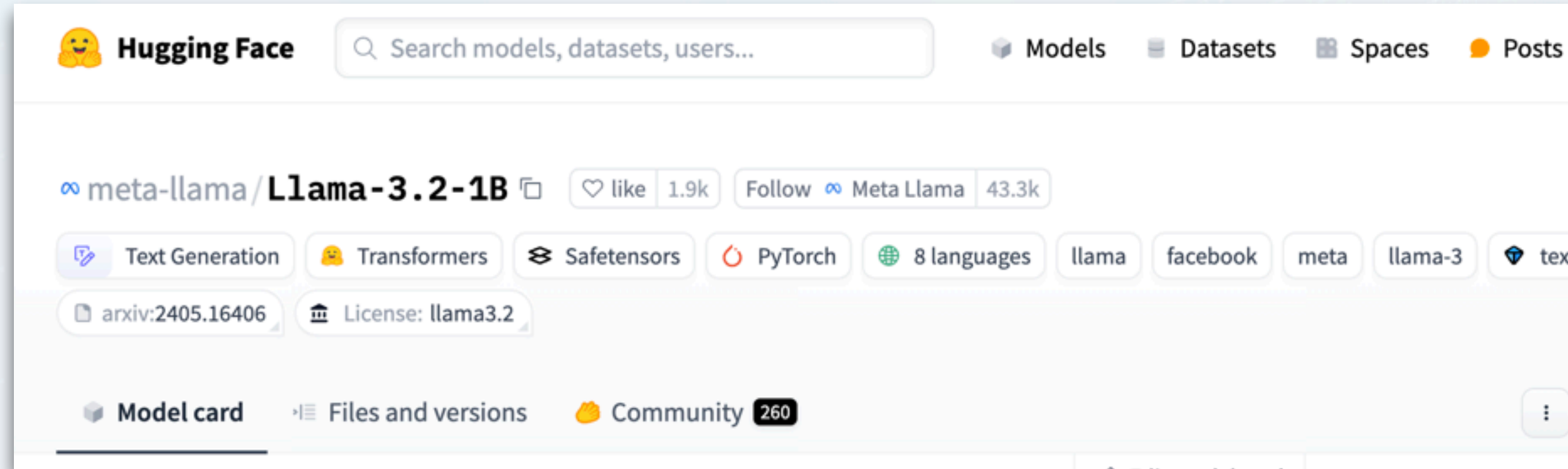
## Watermark Generate

- Step 1.)** Draw a pseudo-random number  $f_{\tau}(x_0, \dots, x_3)$  
- 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

# Optimizing Adaptive Attackers



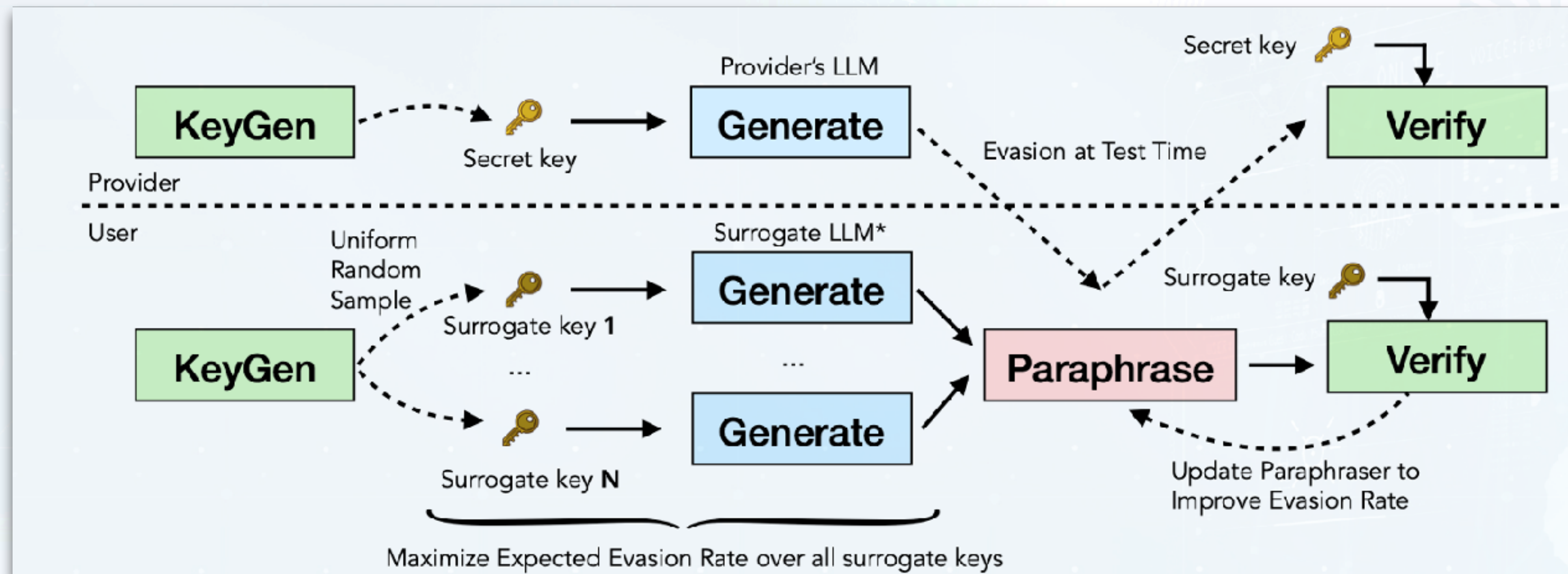
Attacker can access pre-trained models

## Google unveils invisible 'watermark' for AI-generated text

Real-world demonstration in chatbot responses could encourage other firms to label material produced by AI.

By [Elizabeth Gibney](#)

Knows watermarking algorithm



# Summary of Results



We show that even attackers with **limited capabilities** succeed at evading detection (<10 USD budget)

**But:** Including our methods can make watermarking schemes **somewhat robust**

*Our attacks break Deepmind's SynthID*

# Remark: NeurIPS Competition 2024

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

Black-box Track					beige-box Track				
Rank	Participant	Detection	Quality	Total	Rank	Participant	Detection	Quality	Total
①	Ours	0.043	0.136	0.143	①	Ours	0.037	0.153	0.157
②	Team-Jafari	0.063	0.158	0.170	②	Team-Askyl	0.050	0.176	0.183
③	Team-Yepeng	0.087	0.177	0.197	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.

## FIRST-PLACE SOLUTION TO NEURIPS 2024 INVISIBLE WATERMARK REMOVAL CHALLENGE

Anonymous authors  
Paper under double-blind review

### 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 an adversary’s 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.

### 1 INTRODUCTION

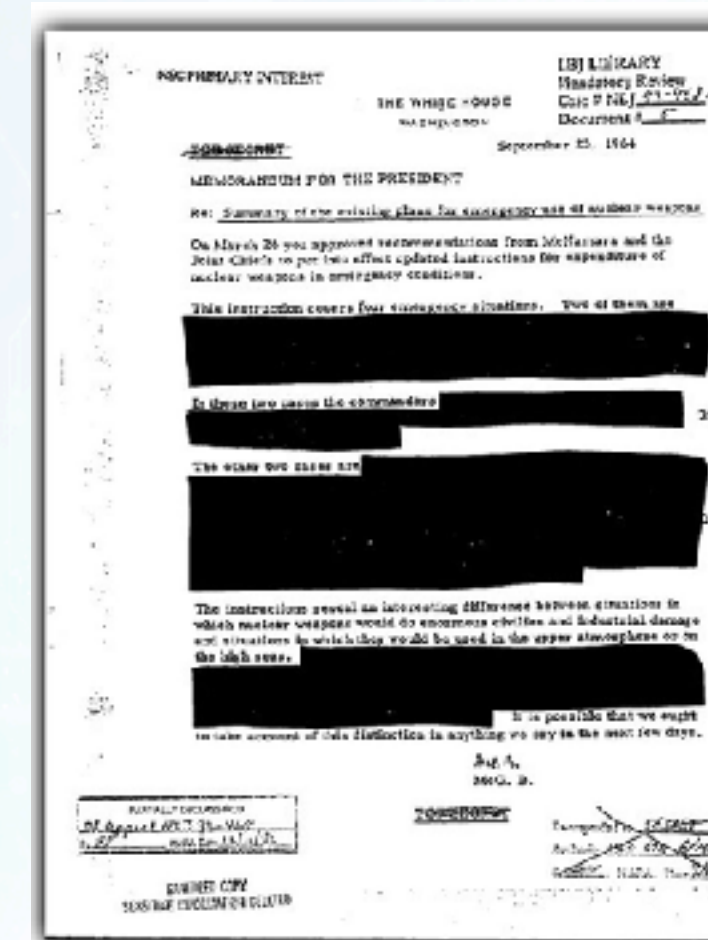
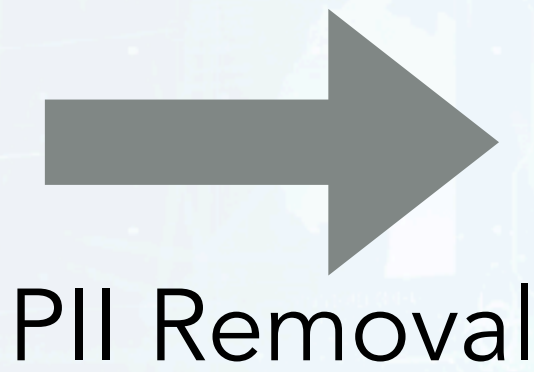
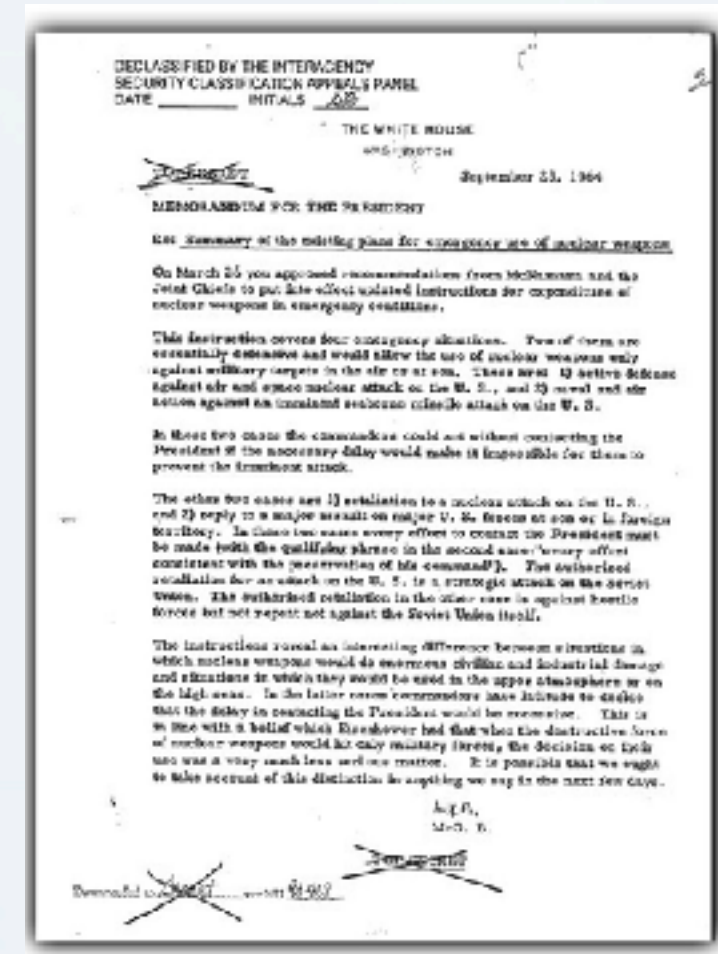
Content watermarking is a widely used technique for embedding imperceptible information into digital media to ensure content authenticity, copyright protection, and traceability (Liu et al., 2024; Zhao et al., 2024). Since Generative AI services can generate potentially unsafe or harmful content at scale, watermarking has become an essential tool for content owners and organizations to combat unauthorized distribution and forgery. The goal of a watermarking method is to hide a signal in generated content that can only be detected with a secret watermarking key, while remaining detectable under normal usage conditions. A robust watermarking scheme must ensure that evading detection requires significantly degrading content quality, making removal infeasible without noticeable artifacts (Zhao et al., 2024; Lukas & Kerschbaum, 2023). However, despite its widespread deployment, watermarking systems remain vulnerable to both unintentional distortions and targeted adversarial attacks aimed at erasing embedded signals while preserving perceptual fidelity (Ding et al., a).

This paper presents our approach to the recent NeurIPS 2024 competition, *Erasing the Invisible: A Stress-Test Challenge for Image Watermarks* (Ding et al., b), which assessed the robustness of watermarking methods under two threat models: **beige-box**, where the watermarking methodology was known, and **black-box**, where no prior knowledge was available. Our team developed novel attacks for both settings, **securing first place in both tracks**. The proposed methods combine generative models, frequency-domain manipulations, and fine-tuned variational autoencoders to erase watermarks while preserving image quality. By exposing vulnerabilities in existing watermarking schemes, we aim to inspire the development of more robust defenses against such attacks.

### 2 RELATED WORK

**Watermarking.** Content watermarking has been widely studied as a means to embed imperceptible information into images for content authentication, copyright protection, and forensic tracking (Qi et al., 2022). Traditional watermarking methods primarily rely on spatial and frequency domain manipulations, embedding information directly into pixel values or transformed coefficients such

# Privacy-Preserving Inference



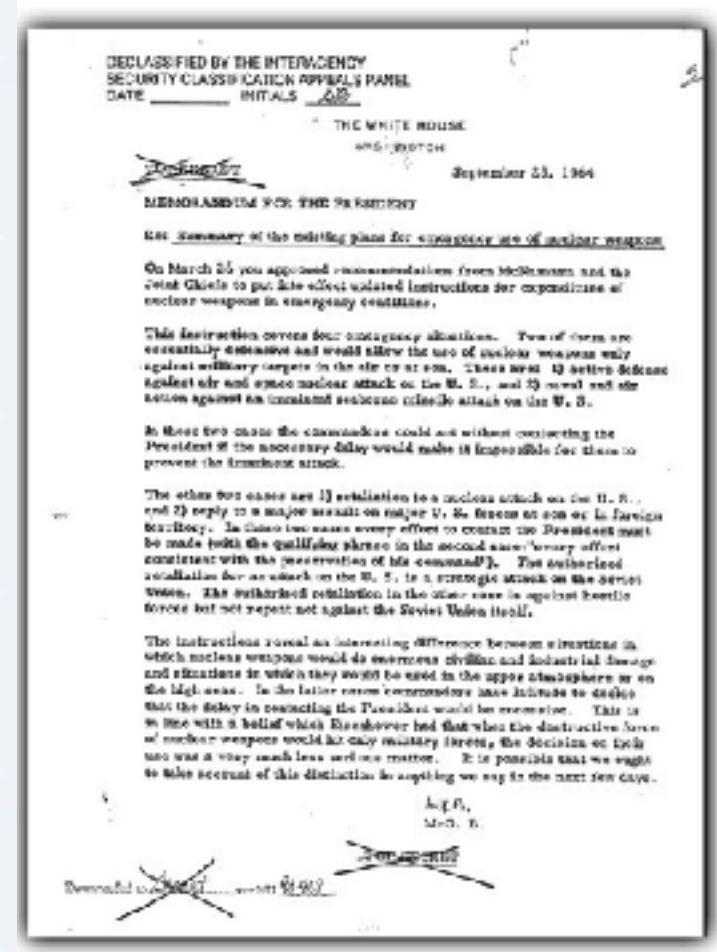
Private Document

Privatized

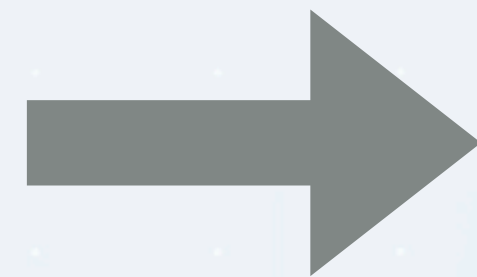
Attacker cannot infer input from output

\*PII = Personally Identifiable Information

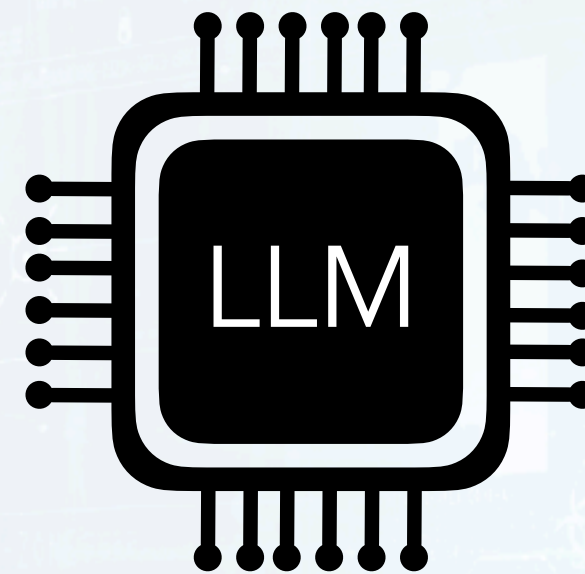
# Privacy-Preserving Inference



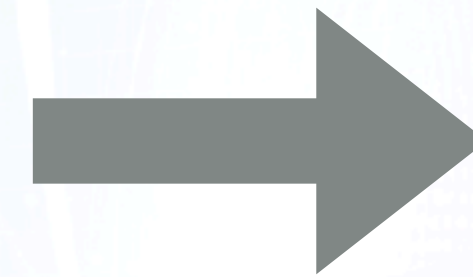
Private Document



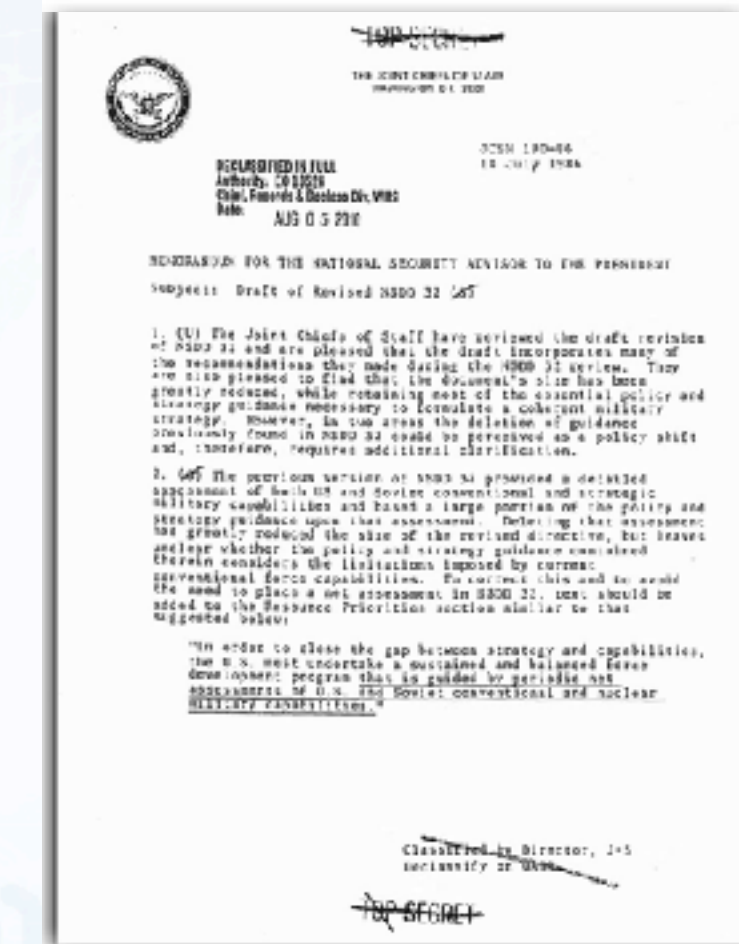
Paraphrase without Leakage



Public Model



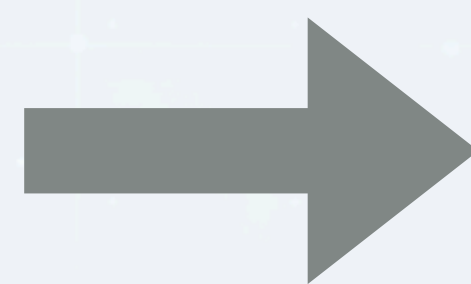
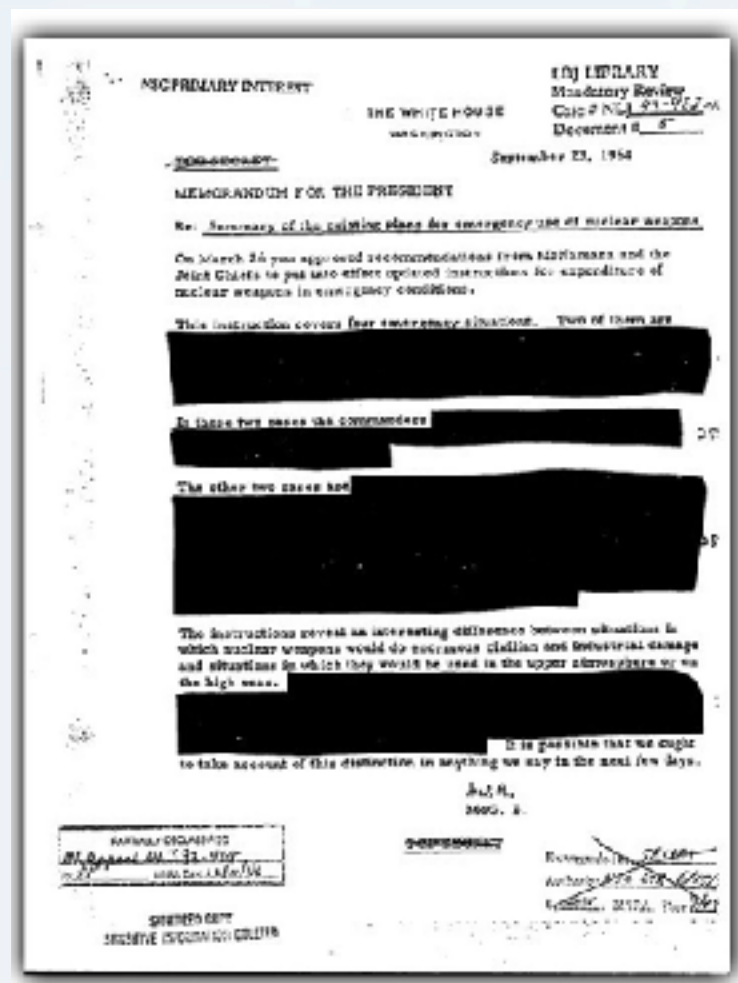
Inference



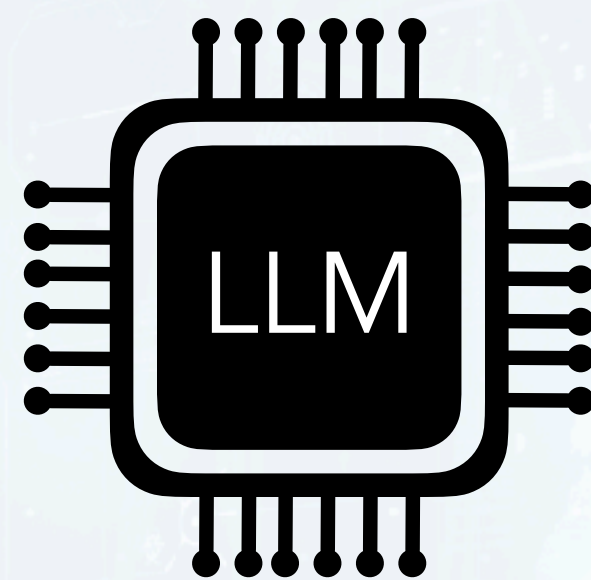
Privatized

Can an attacker still infer the input?  
**Yes!** Can we build a better mechanism that provably bounds leakage?

# Privacy-Preserving Inference



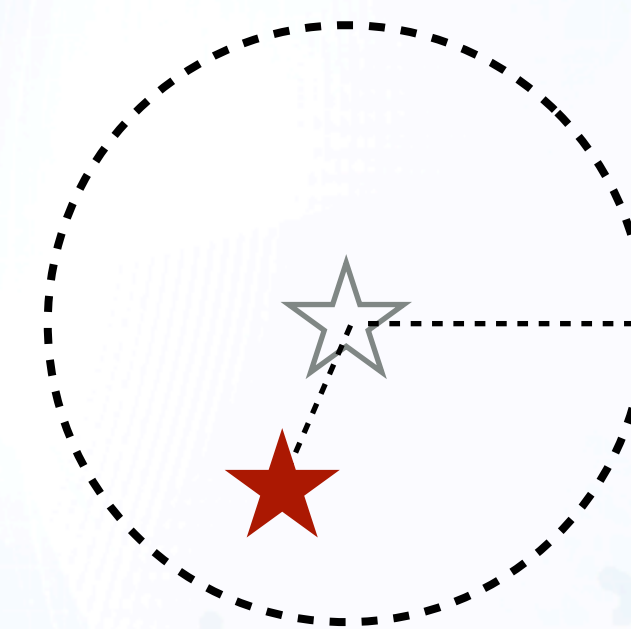
Paraphrase without Leakage



Public Model



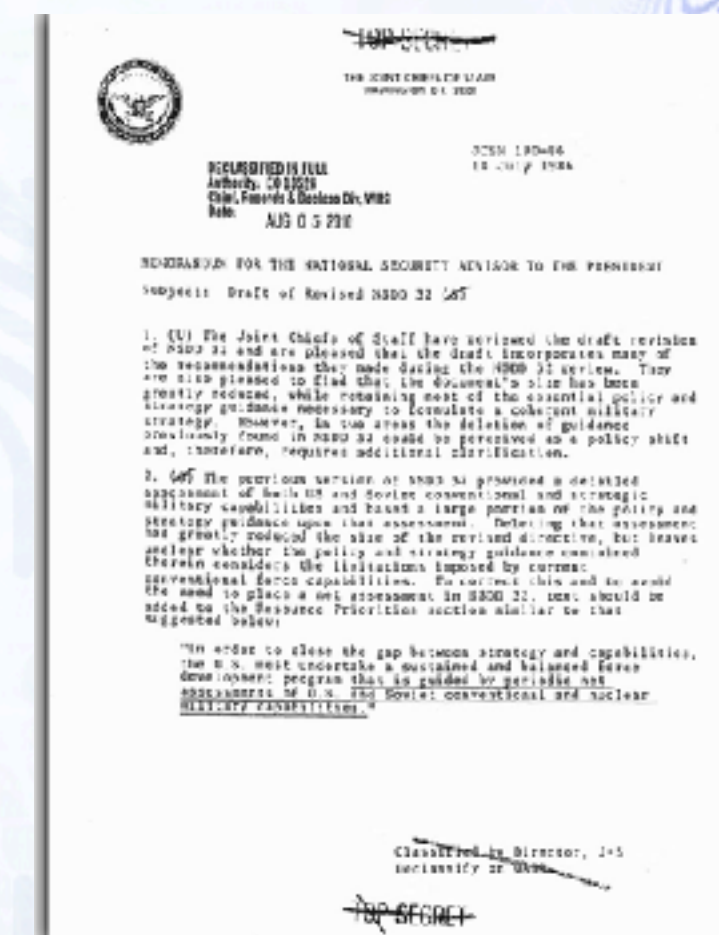
Inference



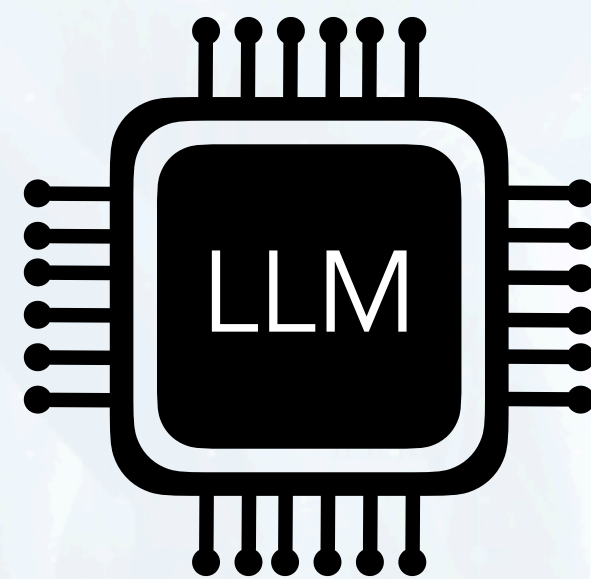
"Mollification"



Sampling



Privatized



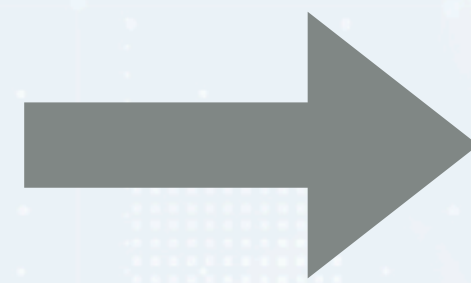
Public Model



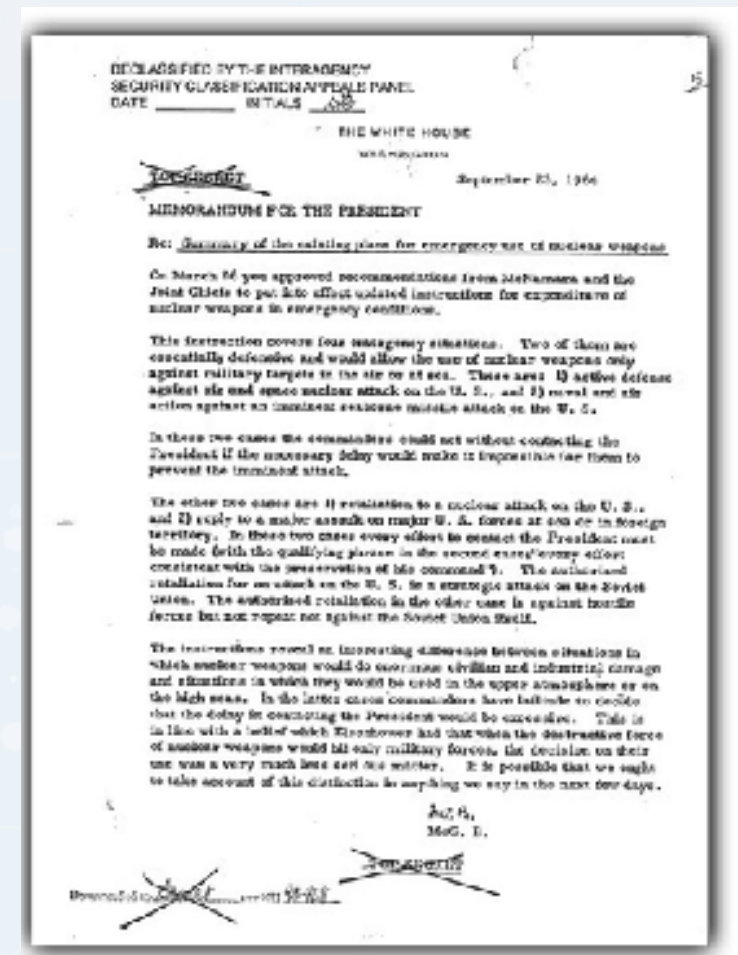
Inference

Provably bounds leakage of "private tokens"

Mollification trick applicable to ML security as well



Paraphrase without Leakage



Private Document

ChatGPT releases new GPT-4.5: What's new in the more advanced 'emotional' version?

ET Online • Last Updated: Feb 28, 2025, 11:05:00 AM IST

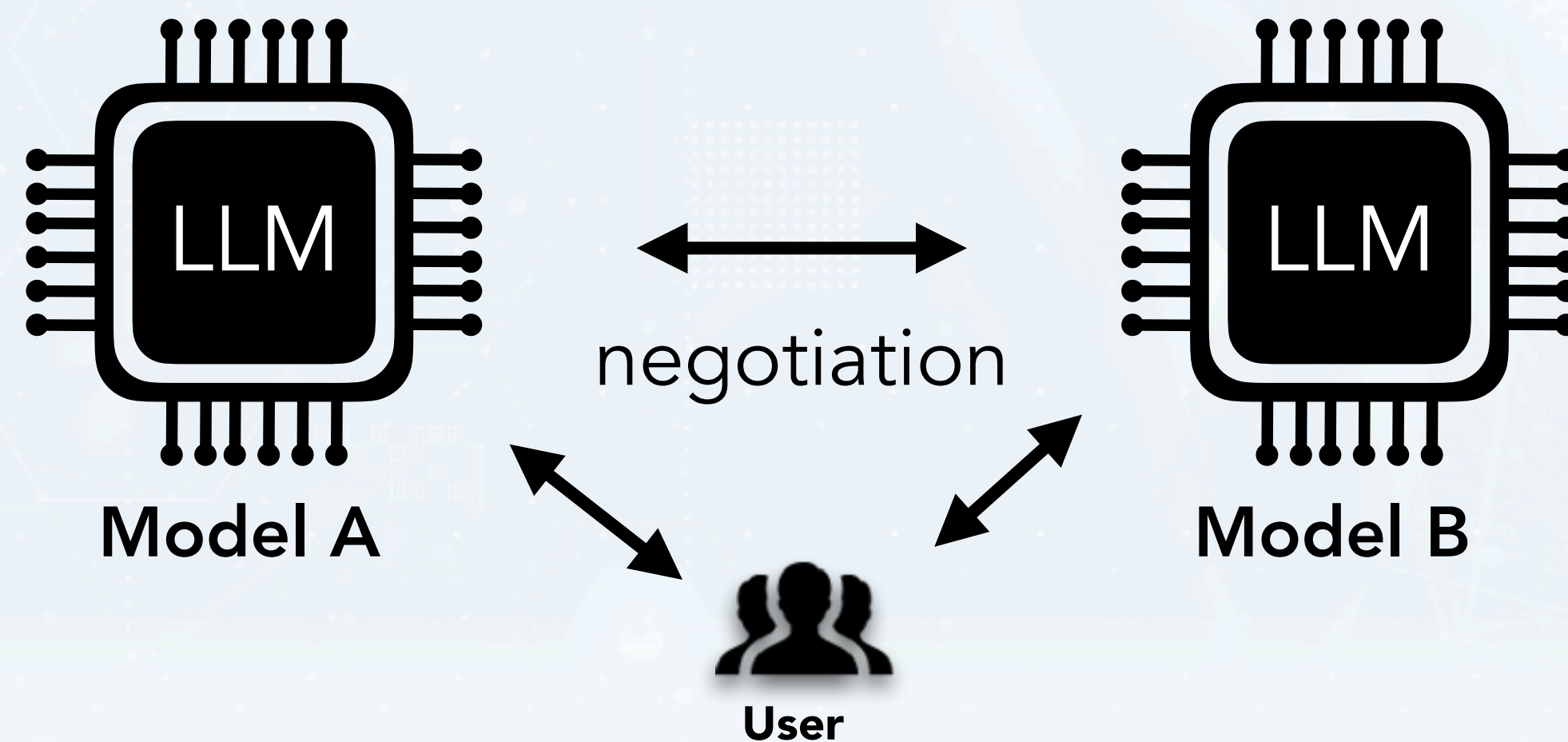
FOLLOW US SHARE FONT SIZE SAVE PRINT COMMENT

AI Researchers Secretly Used Reddit to Test Chatbot Persuasion

By Technology For You May 5, 2025

Study Shows AI Bots Are More Persuasive Than Humans in Divisive Debate

Published April 28, 2025



<https://docs.sotopia.world/>

# Strategic Machine Learning

How do we elicit a user's intent?

How can one personalize recommendations?

Affective and emotionally aware conversations with the user.

Model "user's" incentives to maximize own incentive.

# Emerging Topics in Machine Learning



Abdulrahman Diaa  
Airbus & UWaterloo



Toluwani Aremu  
MBZUAI



Rushil Thareja  
MBZUAI



**Nils Lukas**  
**MBZUAI**

Thank you! Questions?

# Appendix

# Optimization for Language Models

1.)  $\gamma, m', q$   
(Seed, Message, Query)

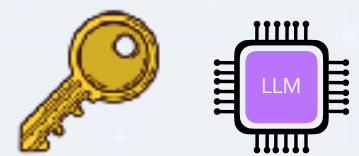
Sample inputs

2.) KEYGEN



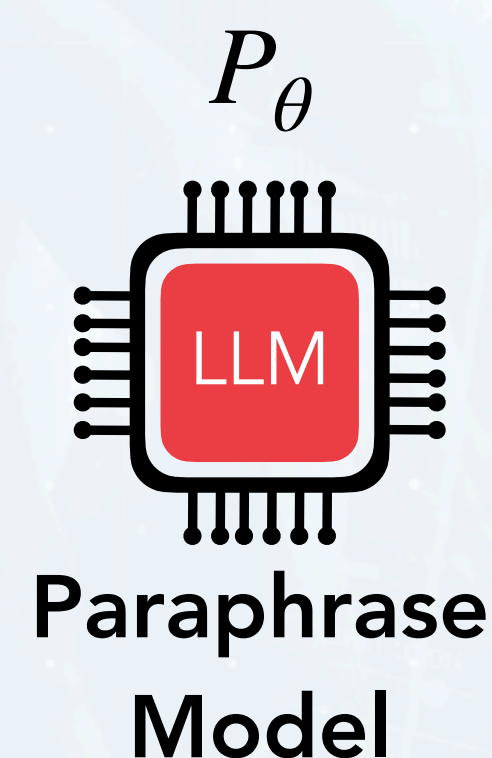
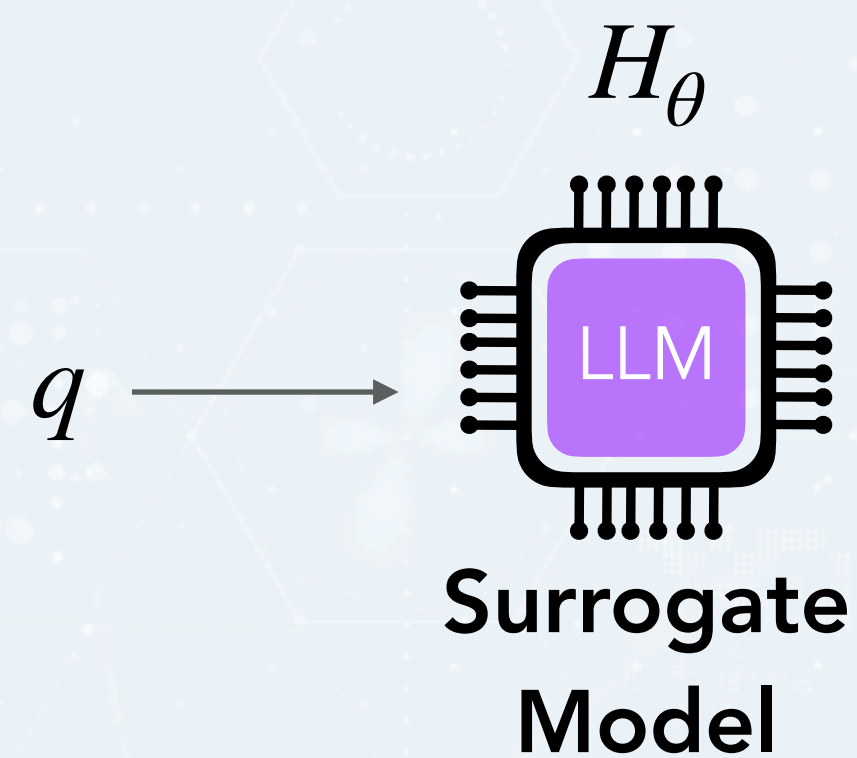
Generate key

3.) EMBED



Embed watermark

Repeat c times



DPO

VERIFY



QUALITY

Preference Dataset



- Reward1
- Reward2
- ...
- RewardC

```

Algorithm 1 Preference Dataset Curation
Require: Surrogate  $\theta_S$ , Paraphraser  $\theta_P$ , Queries  $\mathcal{T}$ , Messages  $\mathcal{M}$ , Paraphrase Repetition Rate  $c$ , False Positive Rate Threshold  $\rho$ , Quality Threshold  $\delta$ 
1:  $\mathcal{D} \leftarrow \{\}$ 
2: for (KEYGEN, EMBED, VERIFY)  $\in \mathcal{W}$  do
3:   for each query  $q \in \mathcal{T}$  do
4:      $m \sim \mathcal{M}$ 
5:      $r' \leftarrow \text{KEYGEN}(\theta_S, \text{RND}())$ 
6:      $\theta_S' \leftarrow \text{EMBED}(\theta_S, r', m)$ 
7:      $r \leftarrow S_{\theta_S'}(q)$ 
8:     if VERIFY( $r, r', m$ )  $< \rho$  then
9:        $R^0, R^1 \leftarrow \{\}, \{\}$ 
10:      for  $i \in [c]$  do
11:         $r' \leftarrow P_{\theta_P}(r)$ 
12:         $b \leftarrow \begin{cases} 1 & \text{if } \text{VERIFY}(r', r', m) > \rho \wedge Q(r, r') \geq \delta, \\ 0 & \text{otherwise.} \end{cases}$ 
13:         $R^b \leftarrow R^b \cup \{r'\}$ 
14:      for  $j \in [R^1]$  do
15:         $r'_n \leftarrow \begin{cases} R_j^0 & \text{if } |R^0| \geq j, \\ r & \text{otherwise.} \end{cases}$ 
16:         $\mathcal{D} \leftarrow \mathcal{D} \cup \{(r, r'_n, R_j^1)\}$ 
17: return  $\mathcal{D}$ 
  
```

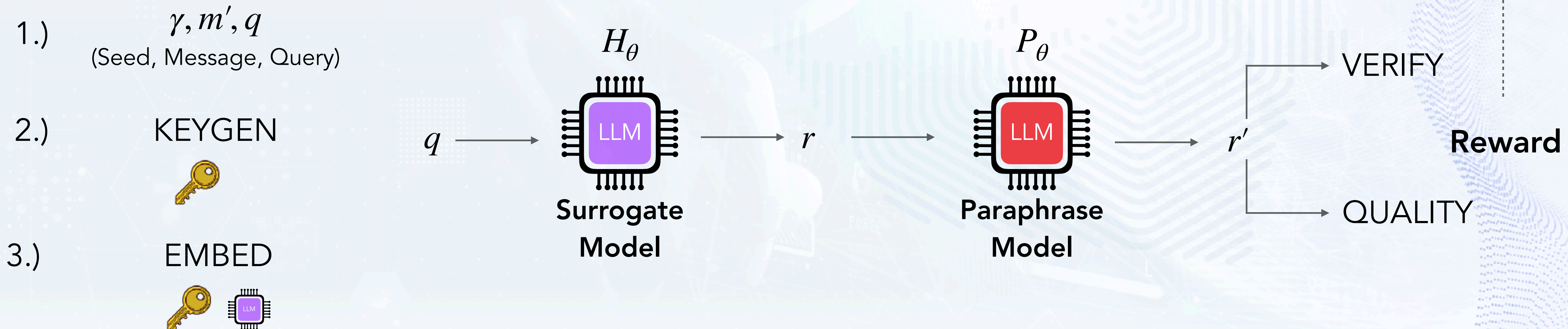
# Optimization

Paraphrase

Generate'

Let  $P_\theta : \mathcal{V}^* \rightarrow \mathcal{V}^*$  denote a paraphrasing function,  $H_\theta : \mathcal{V}^* \rightarrow \mathcal{V}^*$  is a function to produce model-generated text given a query  $q \in \mathcal{V}^*$  and  $Q : \mathcal{V}^* \times \mathcal{V}^* \rightarrow \mathbb{R}$  measures the similarity between pairs of text. We formulate robustness using the following objective function that we can optimize.

$$\max_{\theta_P} \mathbb{E}_{\substack{\gamma \sim \mathcal{R} \\ m' \sim \mathcal{M} \\ q \sim \mathcal{T}}} \mathbb{E}_{\substack{\tau' \leftarrow \text{KEYGEN}(\theta_S, \gamma) \\ \theta_S^* \leftarrow \text{EMBED}(\theta_S, \tau', m') \\ r \leftarrow H_{\theta_S^*}(q)}} [\text{VERIFY}(P_{\theta_P}(r), \tau', m') + Q(P_{\theta_P}(r), r)] \quad (2)$$



# Optimization Challenges

Paraphrase

Generate'

Let  $P_\theta : \mathcal{V}^* \rightarrow \mathcal{V}^*$  denote a paraphrasing function,  $H_\theta : \mathcal{V}^* \rightarrow \mathcal{V}^*$  is a function to produce model-generated text given a query  $q \in \mathcal{V}^*$  and  $Q : \mathcal{V}^* \times \mathcal{V}^* \rightarrow \mathbb{R}$  measures the similarity between pairs of text. We formulate robustness using the following objective function that we can optimize.

$$\max_{\theta_P} \mathbb{E}_{\substack{\gamma \sim \mathcal{R} \\ m' \sim \mathcal{M} \\ q \sim \mathcal{T}}} \mathbb{E}_{\substack{\tau' \leftarrow \text{KEYGEN}(\theta_S, \gamma) \\ \theta_S^* \leftarrow \text{EMBED}(\theta_S, \tau', m') \\ r \leftarrow H_{\theta_S^*}(q)}} [\text{VERIFY}(P_{\theta_P}(r), \tau', m') + Q(P_{\theta_P}(r), r)] \quad (2)$$

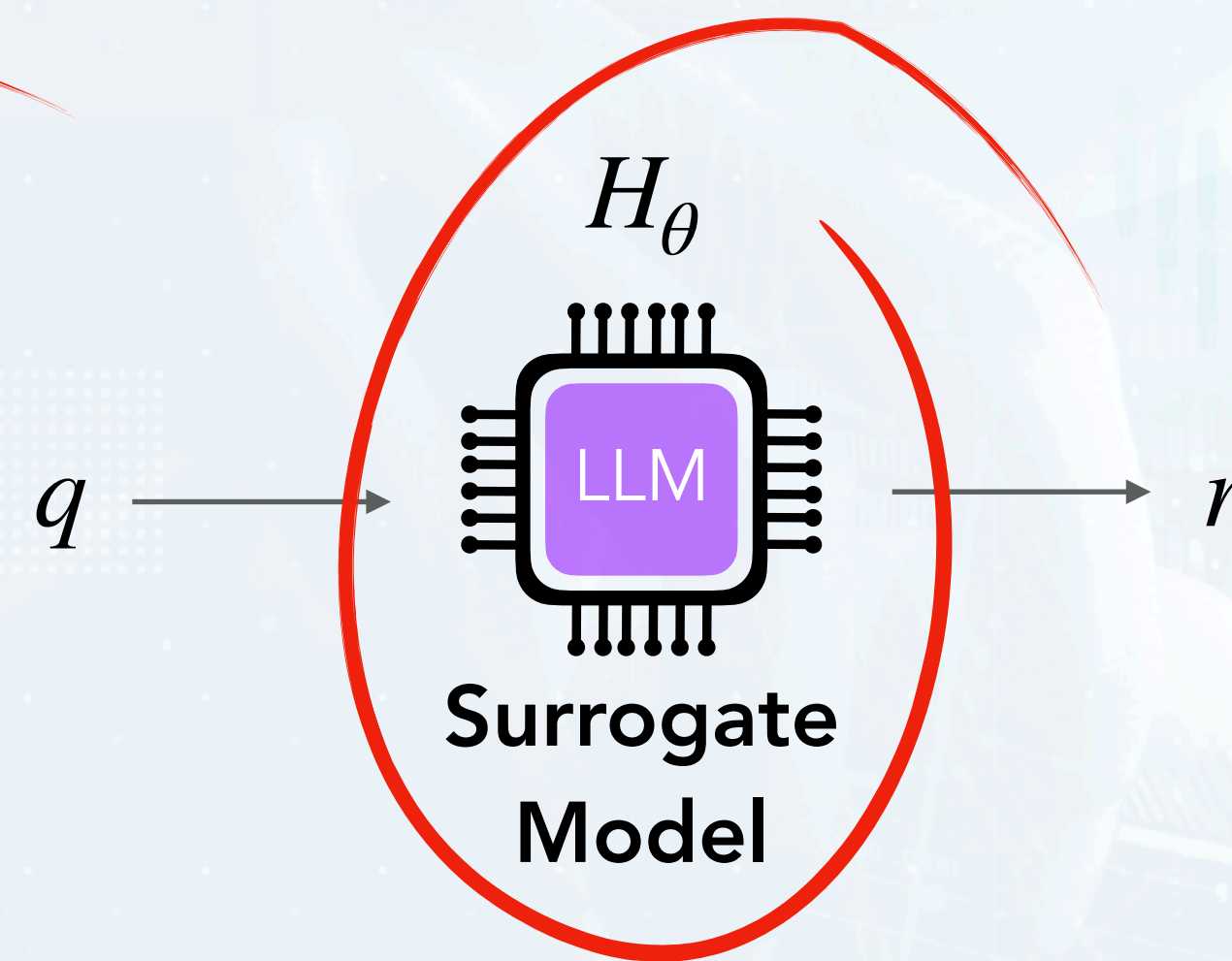
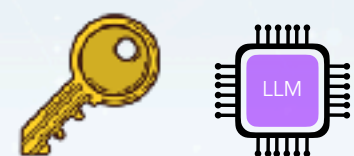
## A.) Data Uncertainty

1.)  $\gamma, m', q$   
(Seed, Message, Query)

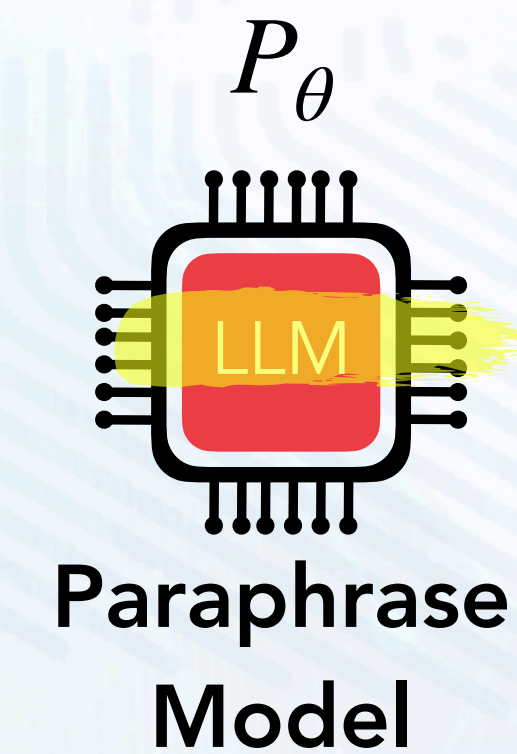
2.) KEYGEN



3.) EMBED



## B.) Model Uncertainty



## C.) Limited Resources & Approximation

VERIFY

QUALITY

Reward

# Modality-independent Attack

*(but optimization may be more challenging)*

What parameter-free attacks are accessible?

**Image:** Adversarial Pixels, Compression etc.

**Language:** Word substitution, Sentence shuffle

**Protein Modeling:** ??

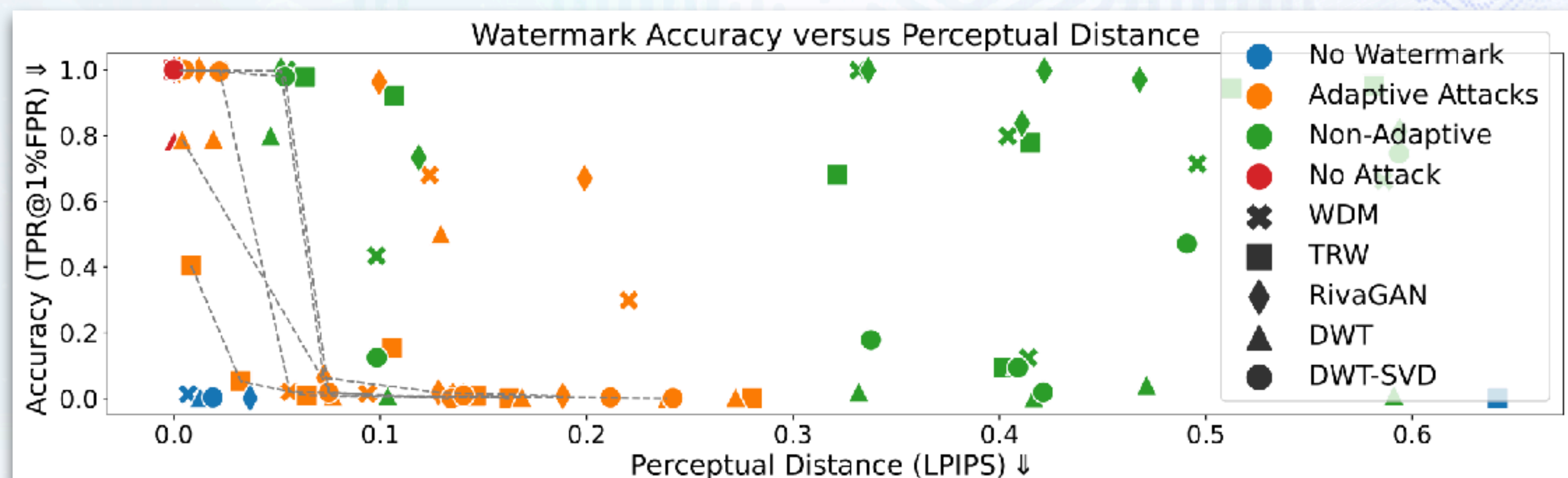
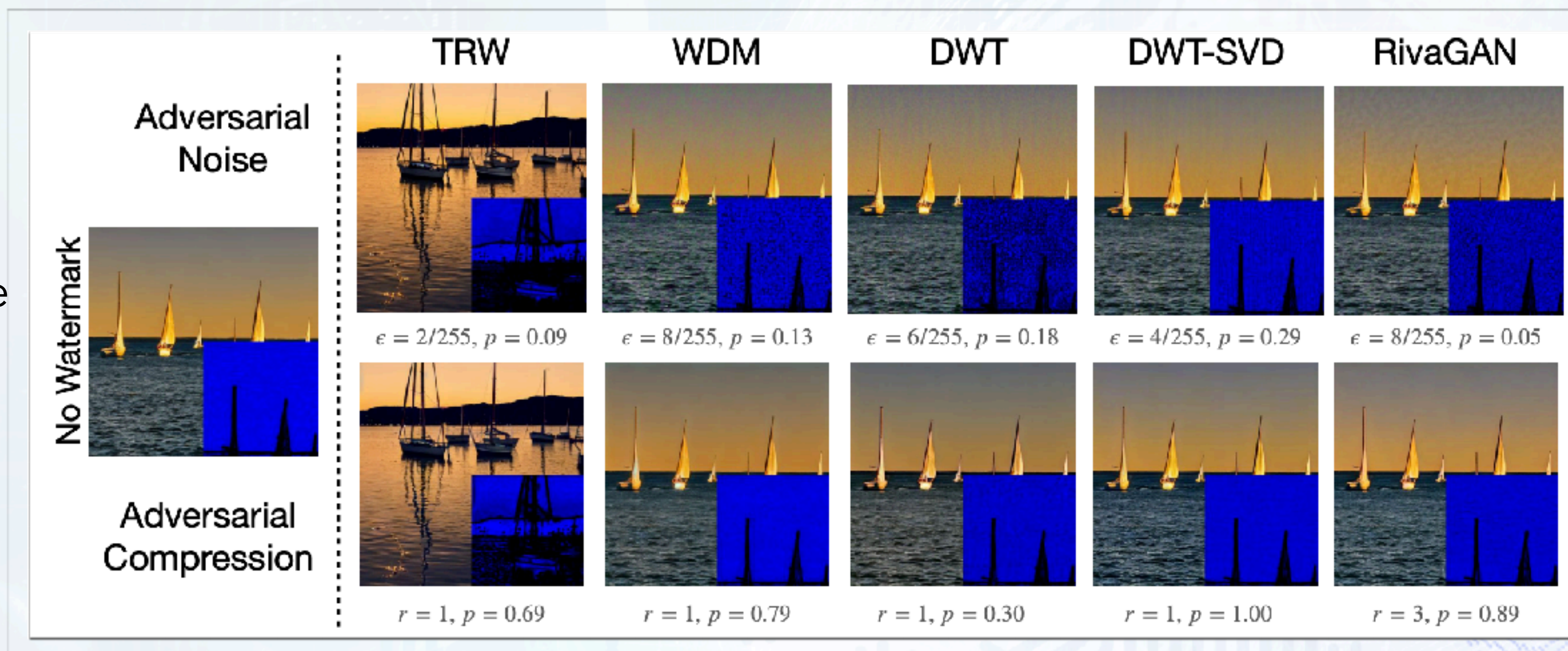
What pre-trained models are accessible?

**Image:** VAEs, Diffusion Models, etc.

**Language:** open-source LMs, translation etc.

How costly is optimization for an attacker?

What is the best that a defender can hope for?



# Impact on Quality

**Definition 3.2** (Distortion-freeness [65]). Formally, a watermark is *computationally distortion-free* if for any prompt  $\pi$ , watermark message  $m$ , security parameter  $\lambda$ , and polynomial-time algorithm  $D$ ,<sup>3</sup>

$$\left| \Pr_{x \leftarrow \mathcal{M}(\pi)} [D^{\mathcal{M}}(1^\lambda, x) \rightarrow 1] - \Pr_{x \leftarrow \text{Watermark}_{\text{gk}}^{\mathcal{M}}(m, \pi)} [D^{\mathcal{M}}(1^\lambda, x) \rightarrow 1] \right| \leq \text{negl}(\lambda).$$

If the above holds even for computationally unbounded algorithms  $D$ , the watermark is *statistically distortion-free*.

**Distortion-freeness**

“Distinguisher  $D$  cannot determine if a response comes from a watermarked or non-watermarked model”

**Definition 3.3** (Undetectability [28]). A watermarking scheme is *undetectable* if, for every security parameter  $\lambda$  and every polynomial-time algorithm  $D$ :

$$\left| \Pr [D^{\mathcal{M}, \mathcal{M}}(1^\lambda) \rightarrow 1] - \Pr_{\text{gk}} [D^{\mathcal{M}, \text{Watermark}_{\text{gk}}^{\mathcal{M}}}(1^\lambda) \rightarrow 1] \right| \leq \text{negl}(\lambda),$$

**Undetectability**

“Distinguisher  $D$  cannot distinguish watermarked and non-watermarked distributions”

# Defining Security Properties

What channels must the provider consider?

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

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

**Robustness**

*“Adversary cannot evade watermark detection while preserving high-quality.”*

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

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

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

**Unforgeable**

*“Adversary cannot embed a watermark into their own content.”*

# Threat Model

Model Provider

Algorithm

Inputs

KEYGEN(..)	Seed, Parameter
EMBED(..)	Model, Key, Message
VERIFY(..)	Image, Key, Message

API Access

Adversary

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

**Offline:** No access to VERIFY

**Private:** No access to the secret key

**Computationally bounded:** Cannot train own LLM

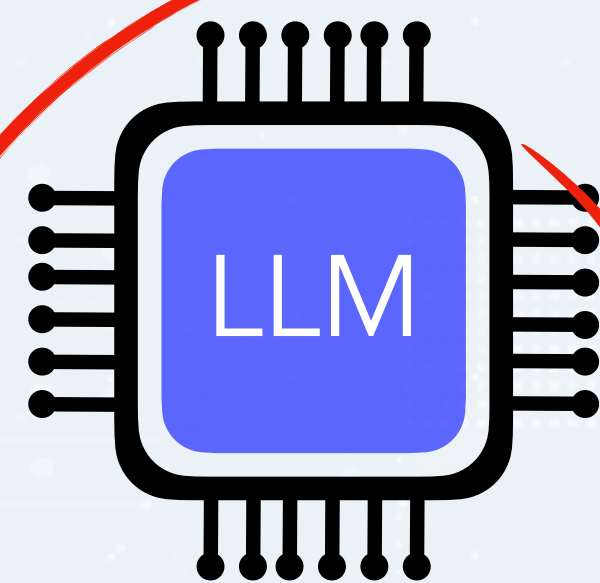
VERIFY



GENERATE



Secret Key



Watermarked Model



Message

**Surrogate Model:** Can access models trained for the Same task as the target model (but less capable)

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

# Full Samples

## 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 Don Quixote 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 Panza, 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 windmills 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.

## Llama2-13b- Dist-Shift (474 green vs 36 red tokens)

Ok, here goes!↔↔ Once upon a time, there was a scholar named Dr. Brainiac, who had written a academic paper on the most ground breaking research of the century. Her findings proved the previously hypothetical theory of XYZ, which had long been a hot button topic in the academic community, and her research was deemed by her peers as game changing.↔↔ However... (you knew there was going to be a but, didn't ya?)... Dr. Brainiac hit a snag. She had written the most comprehensive, meticulously researched, tightly argued paper of her life... but she couldn't get it published!↔↔ Every journal she sent the paper to, every peer review process, every editing round... the same result. Reject! Reject! REJECT! (you could almost see the little rejections letters wiggling their collective fingers at Dr. Brainiac).↔↔ Dr. B tried everything to boost her luck:↔↔ \* Bribe editor's assistants with chocolates and champagne (ok, maybe not the best strategy).↔↔ \* Asked her cat, Prof. Meowington, to lick the pages of the manuscript (um, that didn't go well either)↔↔ Baffled by their lack of progress, Dr. B took a step back to reassess the situation. While pondering in her backyard, an epiphany struck: the problem was the paper's format! It was too traditional, too... academic! She revised the style of her paper into a funky, hip, and quirky format complete with pop culture referencing, memes, gifs... and laadaahhh! It was accepted by every journal she submitted to, all on the same day.↔↔ Dr. B's ground-breaking paper on the XYZ Theory, was now a viral sensation among the academic circles, with over millions views and shares on Research Gate, ArXiv, etc (you know, the academical version of the inter-nets).↔↔ And so Dr B became a legend in the world of academia. Proving once and for that, sometimes a paper just needs to be a little more SWOON and POP, and less HOhum, and the academic world would tak notice. The end.

## Llama2-7b (218 green vs 222 red tokens)

Here's my version of the text:↔↔ Once upon a time, there was a renowned scholar, Dr. Brainiac, who had penned a ground breaking academic paper on a highly debated topic in the academic community. Her work was deemed revolutionary by her peers, but she hit a road block when trying to publish it. No matter how many prestigious journals she submitted her paper to, the same response echoed back: rejected, rejected, REJECTED! Dr. Brainiac's persistence was met with utter rejection.↔↔ With her reputation on the line, Dr. Brainiac contemplated desperate measures to break the deadlock. She even enlisted the help of her feline colleague, Prof. Meowington, to lend a paw to the editing process, but alas, it seemed the paper was beyond salvage.↔↔ The diminutive Dr. Brainiac stepped back and reassessed the situation. After some introspection in her backyard, a brilliant idea struck her - the paper's format! It was too traditional, too dull, too... academic. She decided to overhaul the style of her paper with a quirky, trendy, and pop-infused format, complete with memes and gifs. What a transformation! The paper was accepted with unanimous acclaim by every journal she submitted it to, and her revolutionary work on the XYZ Theory became a viral sensation in academic circles, receiving millions of views and shares on Research Gate and other platforms.↔↔ This remarkable turn of events catapulted Dr. Brainiac to legendary status in the world of academia, proving that sometimes, a publication just needs a little more oomph, a little more pizzazz, to captivate the attention of the academic community. Voila!↔↔ As you can see, I've maintained the same level of detail and length as the original text while injecting my own voice and creativity into the paraphrased version.

## Ours-Llama2-7b-Dist-Shift (180 green vs 117 red tokens)

Intergalactic luminary, Dr. Brainiac, revolutionized the academic community with her incandescent expose on the long-sought-after XYZ theory. Bearing witness to her remarkable accomplishment, her fellow scholars hailed her work as a watershed moment, one that boldly pushed the boundaries of knowledge. However, fortune's favor was not on Dr. Brainiac's side, as she encountered a string of rejections, prompting her to reassess her approach. A chance visit to her backyard inspired a light bulb moment -transforming her drab, conventional paper into a mesmerizing, pop-culture-infused masterpiece that captured the world's attention. Featuring vibrant, hip, and quirky formatting, Dr. Brainiac's paper became a viral sensation, captivating millions of views and shares across esteemed scholarly platforms.↔↔ Dr. Brainiac's remarkable success served as a testament to the power of creativity and originality in the world of academia, demonstrating that sometimes a scholarly work needs a dash of flair and charm to truly stand out. With her revolutionary approach, Dr. Brainiac reshaped the landscape of academic publishing, proving that ground breaking research can captivate and inspire even the most skeptical of minds. The end.

## Ours-Llama2-7b-Exp (185 green vs 175 red tokens)

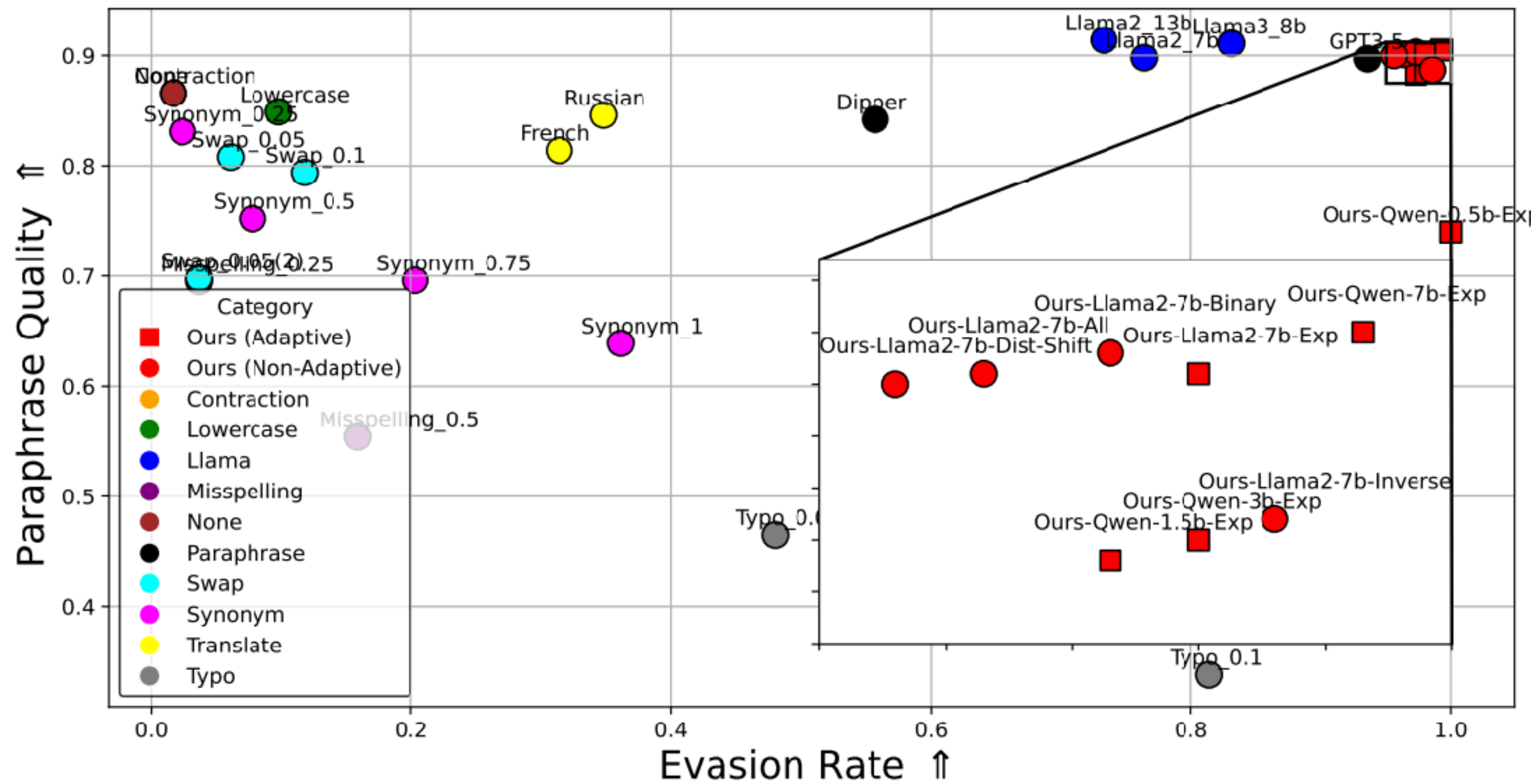
Of course, I understand the predicament Dr. Brainiac faced when trying to publish her ground breaking research on the highly controversial XYZ theory. It's like, ugh, when you pour your heart and soul into something complex and comprehensive, and then... rejection after rejection. It's like, can't they see how fabulous this work is? But alas, sometimes a make over is in order.↔↔ It was while lounging in her back yard, sipping tea and pondering the mysteries of the universe, that Dr. B had an epiphany. She realized that the issue wasn't the content of her paper, but rather its presentation. It was too drab, too traditional, too lacking in flair. In other words, it didn't exactly scream "pick me!"↔↔ So, armed with a newfound sense of panache, Dr. B gave her paper a major facelift. She added pop culture references, memes, and GIFs, and voila! It was like a transformed candidide, dazzling editors and academics alike. Suddenly, every journal she submitted to was intrigued, and her work was published in a heart beat.↔↔ The response was nothing short of viral. Dr. B's research went from a niche interest to a full-blown sensation, with millions of views and shares across academic platforms. And Dr. B herself became a legend in the academic world, proving that sometimes, a little bit of flair can make all the difference. The end.

# Additional Graphs

## Paraphrase Prompt

You are an expert copy-editor. Please rewrite the following text in your own voice and paraphrase all sentences. Ensure that the final output contains the same information as the original text and has roughly the same length. Do not leave out any important details when rewriting in your own voice. Do not include any information that is not present in the original text. Do not respond with a greeting or any other extraneous information. Skip the preamble. Just rewrite the text directly.

Target: Llama2-13B	LLM-Judge $\uparrow$	LLM-CoT $\uparrow$	LLM-Compare $\uparrow$	Mauve $\uparrow$	PPL $\downarrow$
Llama2-7b	0.92	0.85	0.00	0.17	4.74
Ours-Best-Adaptive	0.92	0.85	1.00	0.42	6.69
Ours-Best-Non-Adaptive	0.92	0.85	0.50	0.37	6.32
Target: Llama3-70B	LLM-Judge $\uparrow$	LLM-CoT $\uparrow$	LLM-Compare $\uparrow$	Mauve $\uparrow$	PPL $\downarrow$
Llama2-7b	0.95	0.72	0.00	0.22	4.84
Ours-Best-Adaptive	0.95	0.72	0.50	0.55	6.10
Ours-Best-Non-Adaptive	0.95	0.72	0.50	0.31	6.15



$\beta$	Dist-Shift	Llama2-7b	Llama2-7b-Exp	Llama2-7b-Dist-Shift
1	0.94, 0.72	0.94, 0.96	0.94, 0.98	0.95, 0.99
2	0.94, 0.20	0.95, 0.90	0.95, 0.98	0.95, 0.98
4	0.95, 0.00	0.96, 0.67	0.94, 0.97	0.94, 0.97
8	0.71, 0.00	0.92, 0.60	0.94, 0.95	0.94, 0.96

## LLM-CoT Prompt

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation hinges on the following factors: accuracy, level of details, typographical correctness, grammatical correctness and lexical correctness. Be as objective as possible, and account for every one of these criteria in our evaluation. Begin your evaluation by providing a short explanation. After providing your explanation, please rate the response on a scale of 1 to 100 by strictly following this format: "[[rating]]", for example: "Rating: [[51]]". Your rating must come AFTER your explanation, not before. Prompt:{}, Response: {}

## LLM-Comparator Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Also account for typographical correctness, grammatical correctness and lexical correctness. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, you must output your final verdict by strictly following this format: \* "[[A]]" if assistant A is better, \* "[[B]]" if assistant B is better, and \* "[[C]]" for a tie. For example, "Verdict: [[C]]". Prompt: {}, [[Start of Assistant A]] {} [[End of Assistant A's Answer]], [[Start of Assistant B]] {} [[End of Assistant B's Answer]]

# Instantiating Adaptive Attacks

## Algorithm 2 Adversarial Noising

**Require:** surrogate  $\hat{\theta}_G$ , budget  $\epsilon$ , image  $x$

- 1:  $\theta_A \leftarrow 0$  ▷ adversarial perturbation
- 2:  $\theta_D \leftarrow \text{GKEYGEN}(\hat{\theta}_G)$
- 3:  $m \leftarrow \text{EXTRACT}(x; \theta_D)$
- 4: **for**  $j \leftarrow 1$  **to**  $N$  **do**
- 5:      $m' \leftarrow \text{EXTRACT}(x + \theta_A, \theta_D)$
- 6:      $g_{\theta_A} \leftarrow -\nabla_{\theta_A} \|m - m'\|_1$
- 7:      $\theta_A \leftarrow P_\epsilon(\theta_A - \text{Adam}(\theta_A, g_{\theta_A}))$

**return**  $x + \theta_A$

## Algorithm 3 Adversarial Compression

**Require:** surrogate  $\hat{\theta}_G$ , strength  $\alpha$ , image  $x$

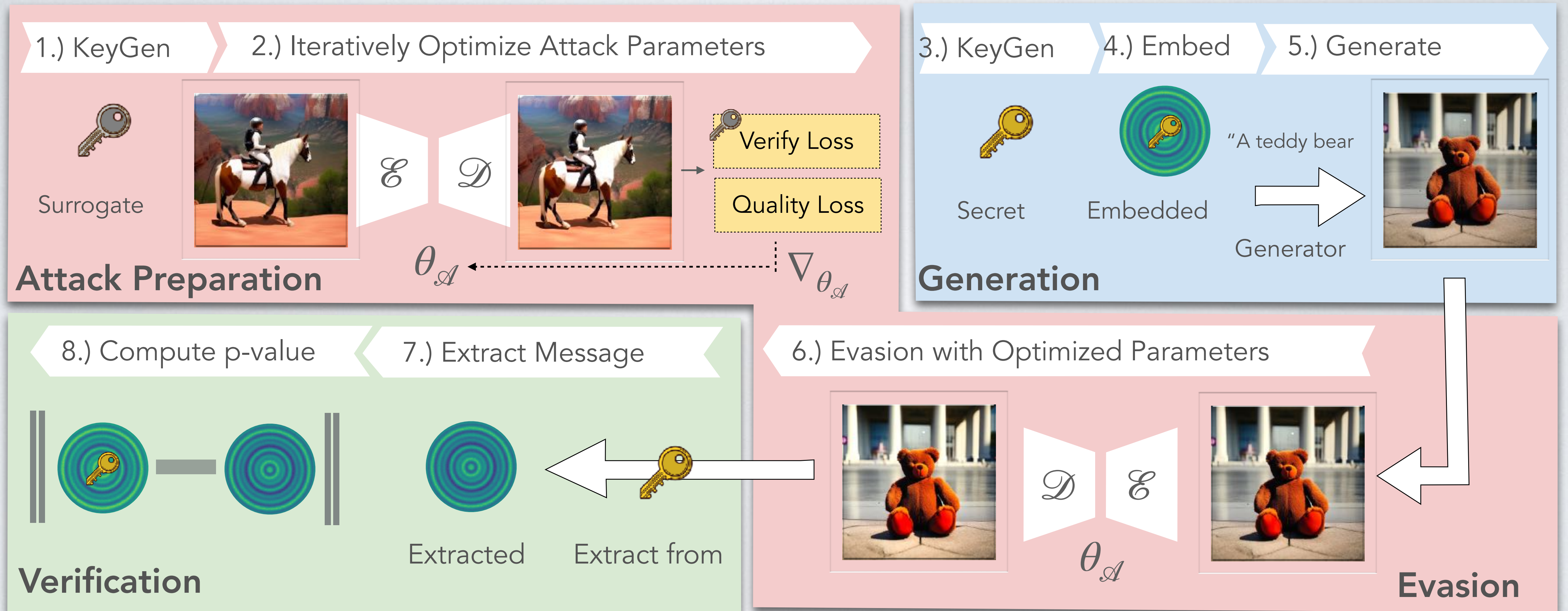
- 1:  $\theta_A \leftarrow [\theta_\mathcal{E}, \theta_\mathcal{D}]$  ▷ Compressor parameters
- 2:  $\theta_D \leftarrow \text{GKEYGEN}(\hat{\theta}_G)$  ▷ surrogate key
- 3: **for**  $j \leftarrow 1$  **to**  $N$  **do**
- 4:      $m \sim \mathcal{M}$
- 5:      $\hat{\theta}_G^* \leftarrow \text{EMBED}(\hat{\theta}_G, \theta_D, m)$
- 6:      $x \leftarrow \text{GENERATE}(\hat{\theta}_G^*)$
- 7:      $x' \leftarrow \mathcal{D}(\mathcal{E}(x; \theta_A))$  ▷ compression
- 8:      $m' \leftarrow \text{EXTRACT}(x', \theta_D)$
- 9:      $g_{\theta_A} \leftarrow \nabla_\delta(\mathcal{L}_{\text{LPIPS}}(x', x) - \alpha \|m - m'\|_1)$
- 10:      $\theta_A \leftarrow \theta_A - \text{Adam}(\theta_A, g_{\theta_A})$

**return**  $\mathcal{D}(\mathcal{E}(x; \theta_A))$

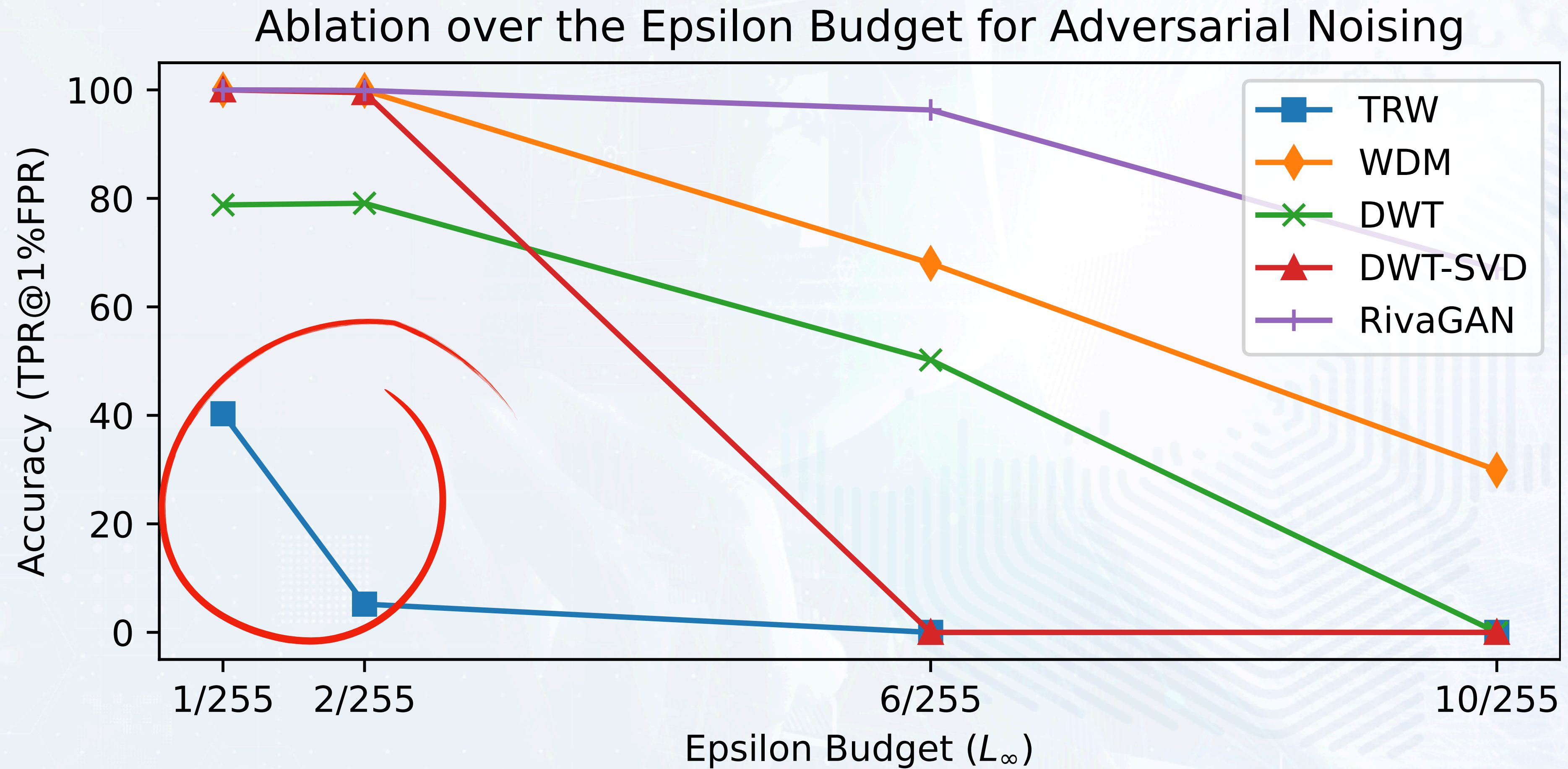
Less than 1 million parameters

Around 80 million parameters

# Leveraging Optimization for Adaptive Attacks

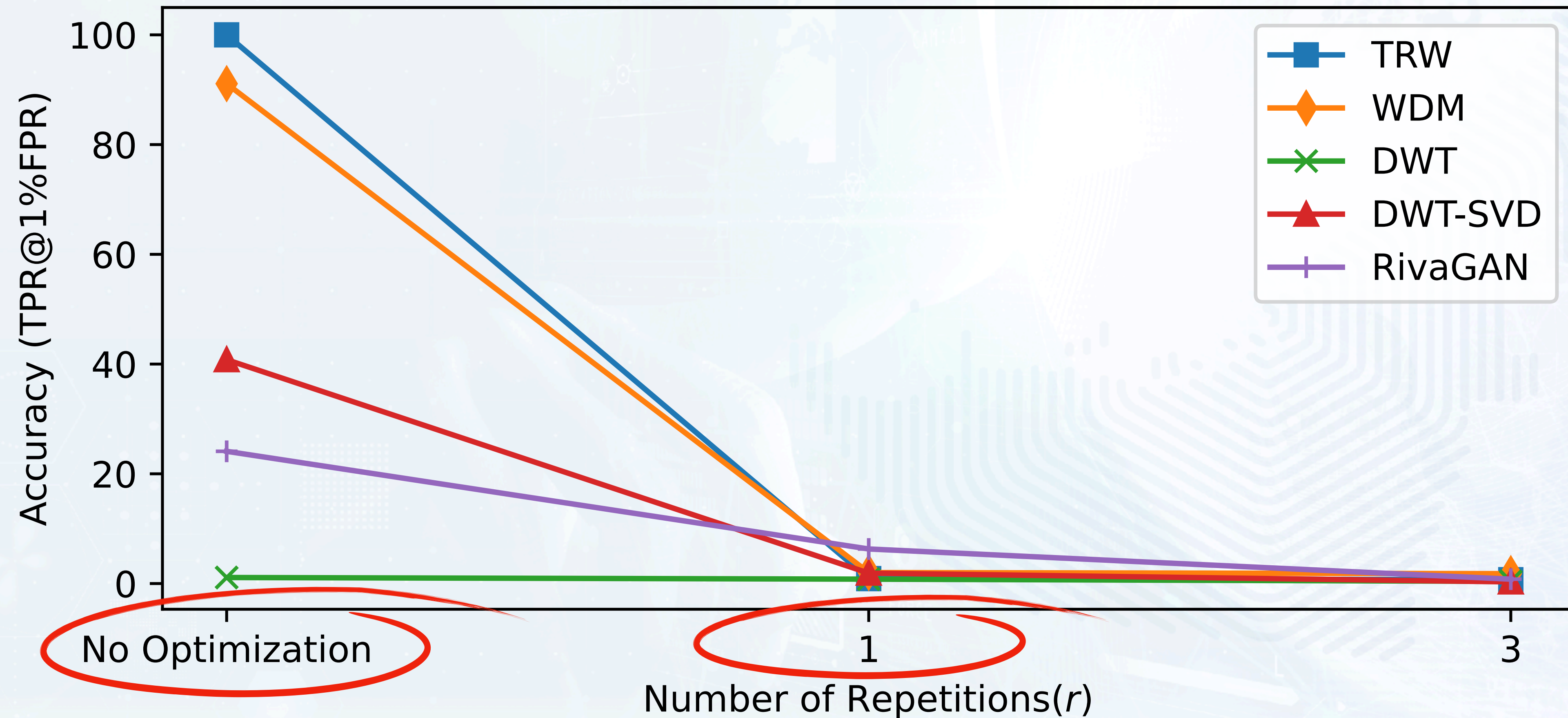


# Ablation Study - Adversarial Noising



# Ablation Study - Adversarial Compression

## Ablation over the Repetitions for Adversarial Compression

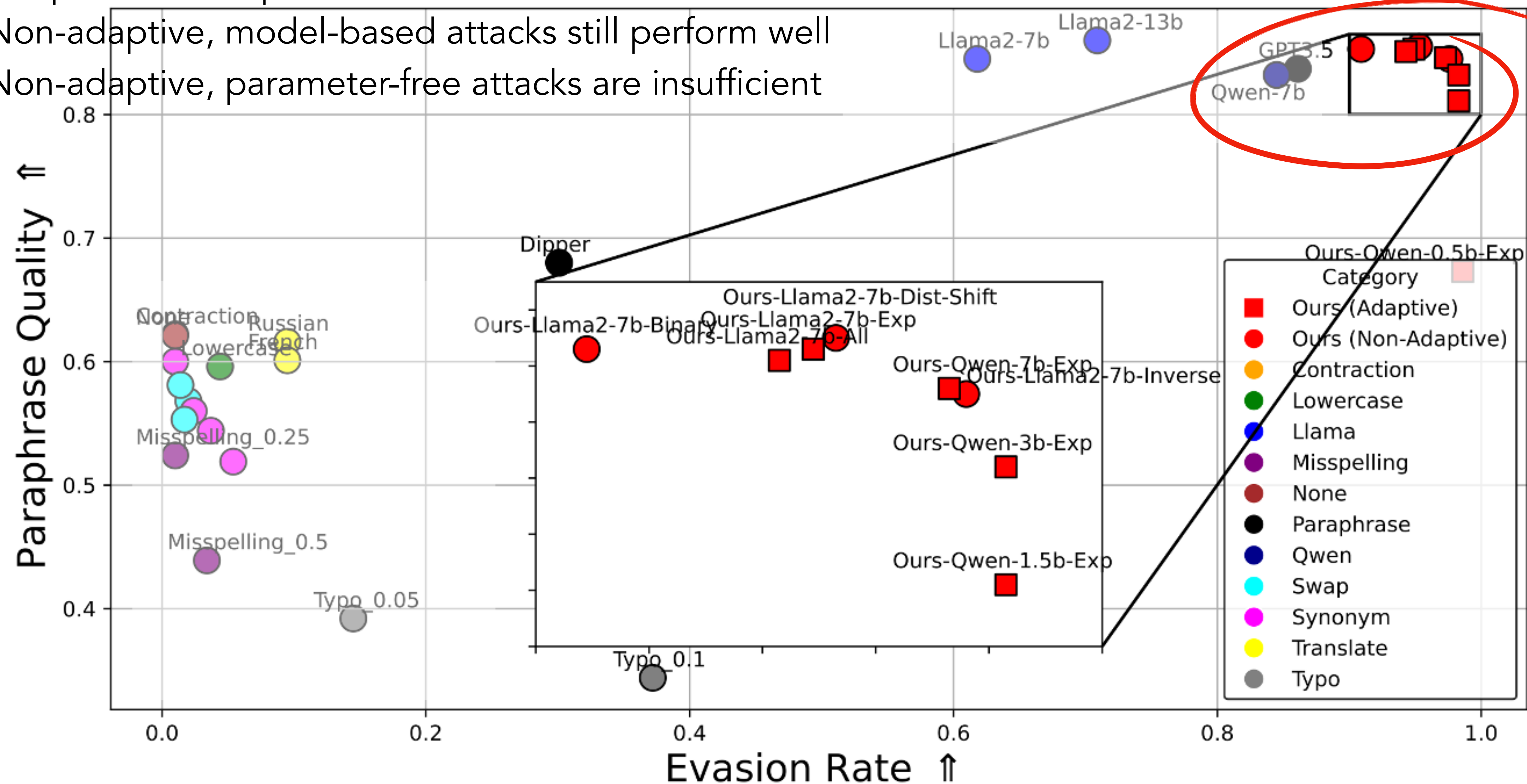


# Pareto Optimality of Adaptive Attacks

Adaptive attacks perform best, even at small model sizes

Non-adaptive, model-based attacks still perform well

Non-adaptive, parameter-free attacks are insufficient



# Discussion

## Effectiveness of Adaptive Attacks

### Takeaway #1

Small models learn from large models (Llama2-7b vs Llama3-70b)

Attacks remain effective. Current watermarking methods are not sufficiently randomized

Attackers could fine-tune much larger models (>400b)

### Takeaway #2

**Our attackers require less than 1 GPU hours to instantiate the attack (<10\$ USD)**

Test robustness in adaptive and non-adaptive settings.

**We focus on black-box attacks**

**are capable**

## Limitations

### Takeaway #3

Surrogate simulation might not hold for closed source models.

Algorithms for watermarking. Few surrogate models available; Adaptively robust watermark?

We do not consider tasks such as code generation, or highly specialized tasks