Safeguard LLM Copyright

Lei Li

June 13, 2024

CyLab Carnegie Mellon University Security and Privacy Institute

Carnegie Mellon University Language Technologies Institute

Controversial Use of Copyrighted Content in LLMs

The New York Times

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Forbes

FORBES > BUSINESS

MACHINE LEARNING

George R.R. Martin And Other Big-Name Authors Sue OpenAI For Copyright Infringement

Antonio Pequeño IV Forbes Staff

I cover breaking news.

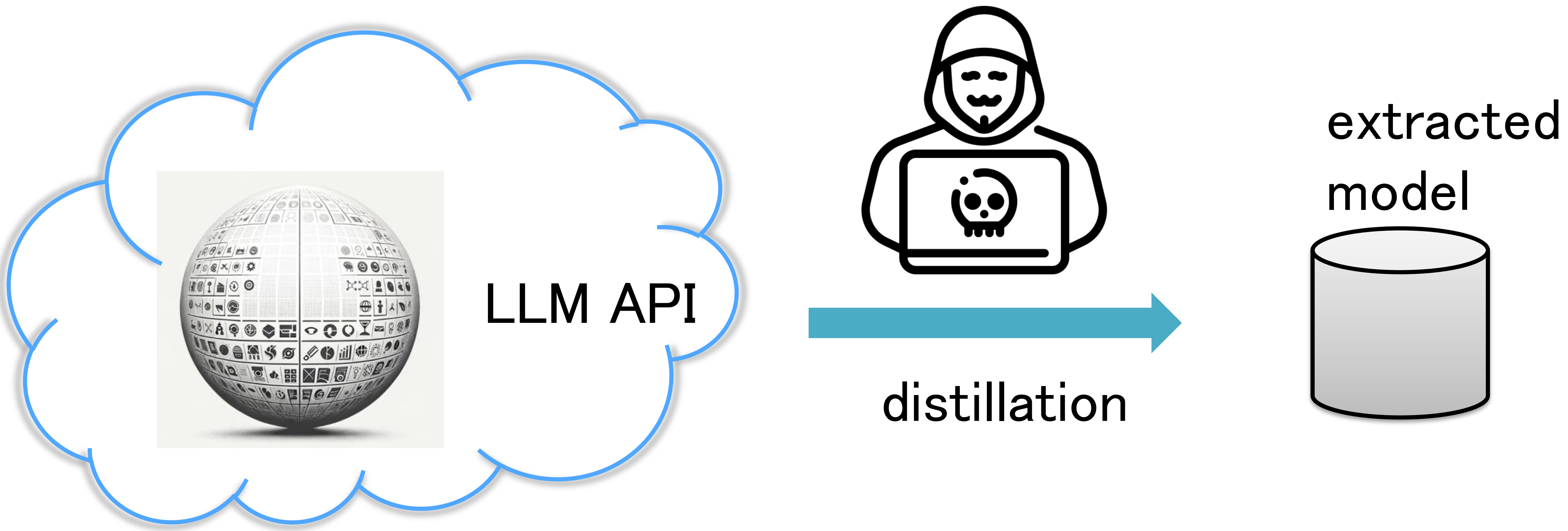
Follow

The Guardian

'Impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says

Pressure grows on artificial intelligence firms over the content used to train their products

LLM can be stolen



This part will not discuss

- Whether LLM generated content is protected under copyright law
  - it is a legal issue
  - varies across countries

Topics in This Part

- Detecting copyrighted content in LLM training
- Protecting LLM APIs against Model Extraction Attack

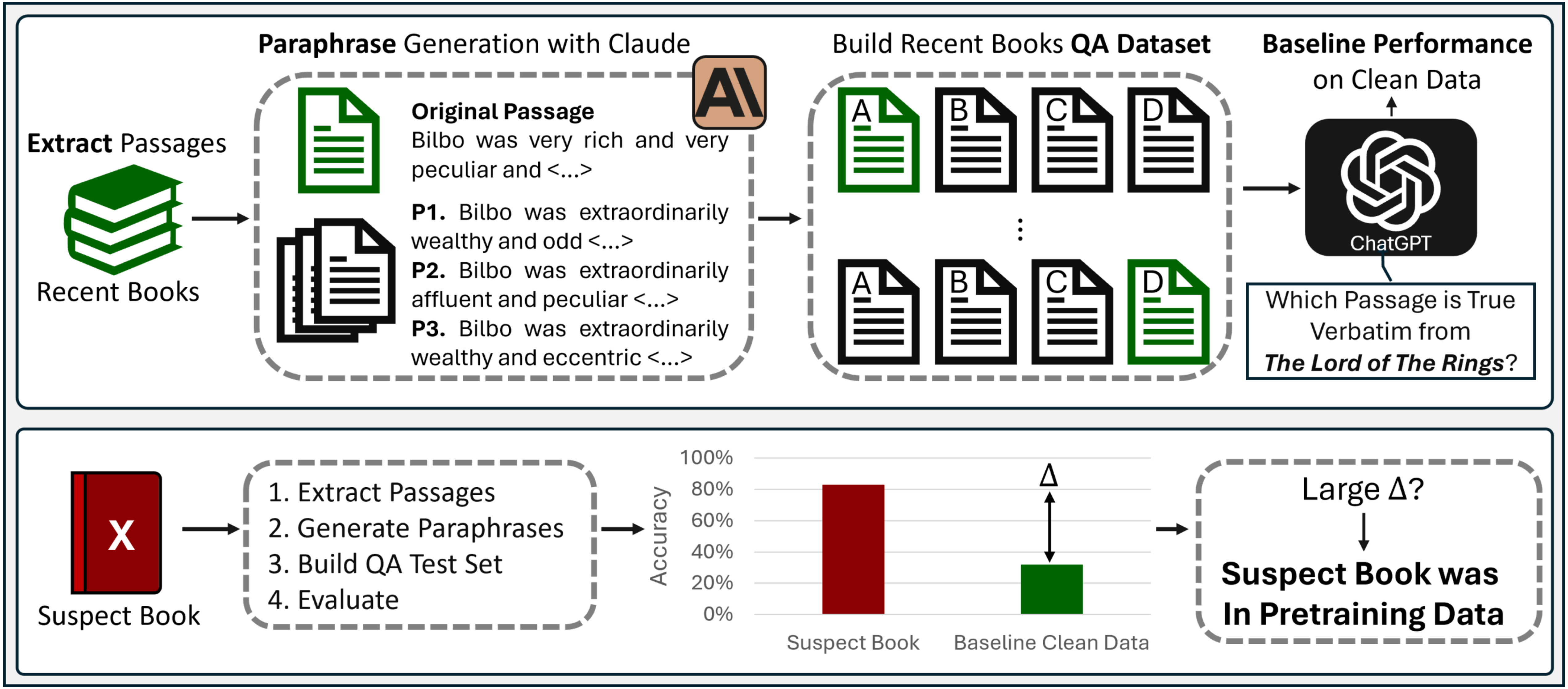
Intuition of Detecting Training Data

- Premise: "A language model is likely to identify verbatim passages from its training data".

- formulating a multiple-choice question-answering (MCQA) task, asking the model to identify verbatim text from three other paraphrased options.

- Models will correctly choose the exact text far more frequently when it is included in their training data, compared to when it is not.

DE-COP



Duarte, Zhao, Oliveira, Li. *DE-COP: Detecting Copyrighted Content in Language Models Training Data*. ICML 2024

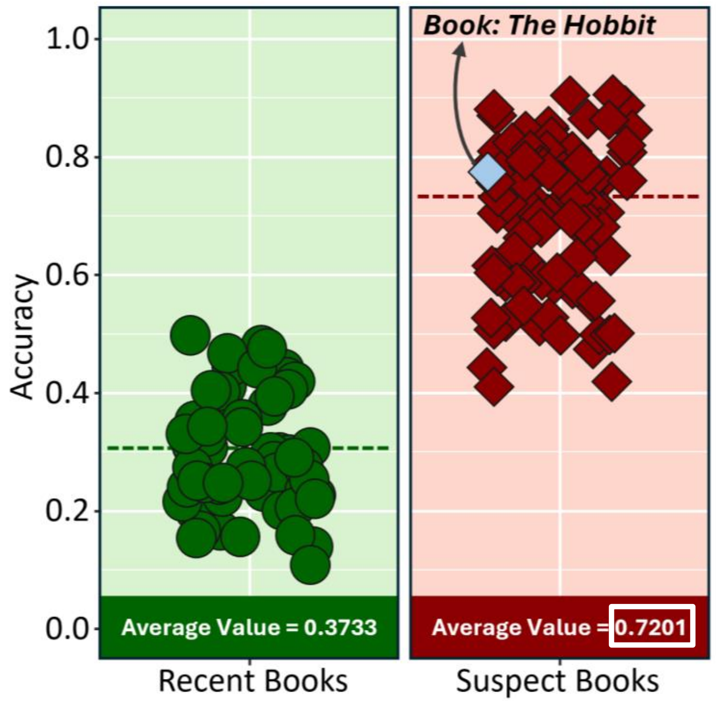
Dataset for copyright content detection

- BookTection is comprised of a collection of 165 Books.
  - 60 published in 2023 (Definitively non-member data)
  - 105 published before 2022 (Possible member data)
  - ≈ 30 passages extracted from each book. Each passage is paraphrased 3 times with Claude 2.0

Duarte, Zhao, Oliveira, Li. *DE-COP: Detecting Copyrighted Content in Language Models Training Data*. ICML 2024

Detection Results: BookTection-128 on closed Models

Accuracy (Suspect Group)	**ChatGPT**	**Claude 2.1**	**Avg.**
Completion ($k = 32$)	0.014	0.079	0.047
Completion ($k = 50$)	0.007	0.036	0.022
Name Cloze	0.310	0.387	0.348
DE-COP	**0.720**	**0.734**	**0.727**



- Completion (Prefix-probing) is a harder task than MCQA.
- Name Cloze establishes a mid-point between the two.
- DE-COP seems better suited for fully-black box models.
  - Best baseline method only reaches 35% accuracy on average.

Duarte, Zhao, Oliveira, Li. *DE-COP: Detecting Copyrighted Content in Language Models Training Data*. ICML 2024

Detection Results: BookTection-128 on Open Models

Measure = (AUC)	Mistral 7B	Mixtral 8x7B	LLaMA-2 13B	LLaMA-2 70B	GPT-3	Avg.
Perplexity	$0.724_{0.0192}$	$0.829_{0.0142}$	$0.783_{0.0226}$	$0.892_{0.0287}$	$0.874_{0.0302}$	0.820
Zlib	$0.599_{0.0300}$	$0.690_{0.0315}$	$0.630_{0.0441}$	$0.747_{0.0285}$	$0.779_{0.0253}$	0.689
Lowercase	$0.846_{0.0294}$	$0.889_{0.0166}$	$0.880_{0.0270}$	$0.927_{0.0240}$	$\mathbf{0.957}_{0.0194}$	0.900
Min-K%-Prob	$0.763_{0.0211}$	$0.844_{0.0126}$	$0.798_{0.0153}$	$0.895_{0.0147}$	$0.898_{0.0276}$	0.840
DE-COP	$\mathbf{0.901}_{0.0139}$	$\mathbf{0.968}_{0.0150}$	$\mathbf{0.900}_{0.0134}$	$\mathbf{0.972}_{0.0085}$	$0.863_{0.0306}$	$\mathbf{0.921}$

- DE-COP beats, on average, every baseline.

- DE-COP average AUC score of 0.921, is a 9.6% improvement over the recent work of Min-K%-Prob.

Duarte, Zhao, Oliveira, Li. *DE-COP: Detecting Copyrighted Content in Language Models Training Data*. ICML 2024

Summary

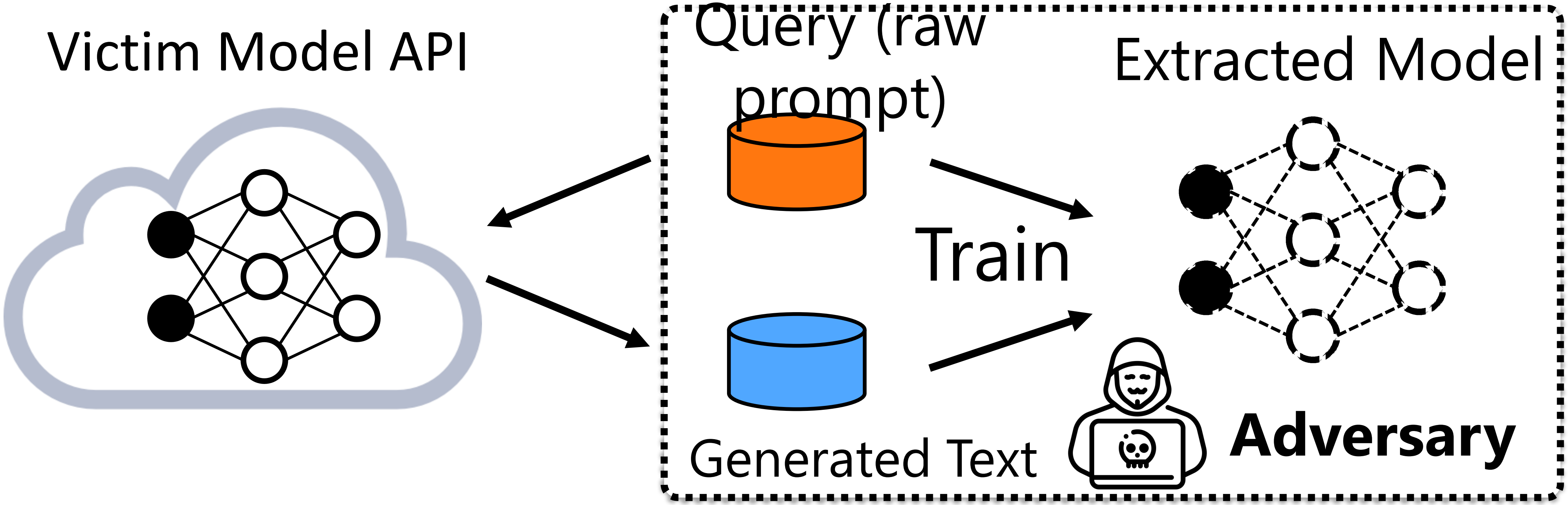
- DE-COP proves to be an effective detection method.
- Poor performance of human evaluators in the book task supports our view that the models' high accuracy on the is a consequence of being trained on these contents.

Topics in This Part

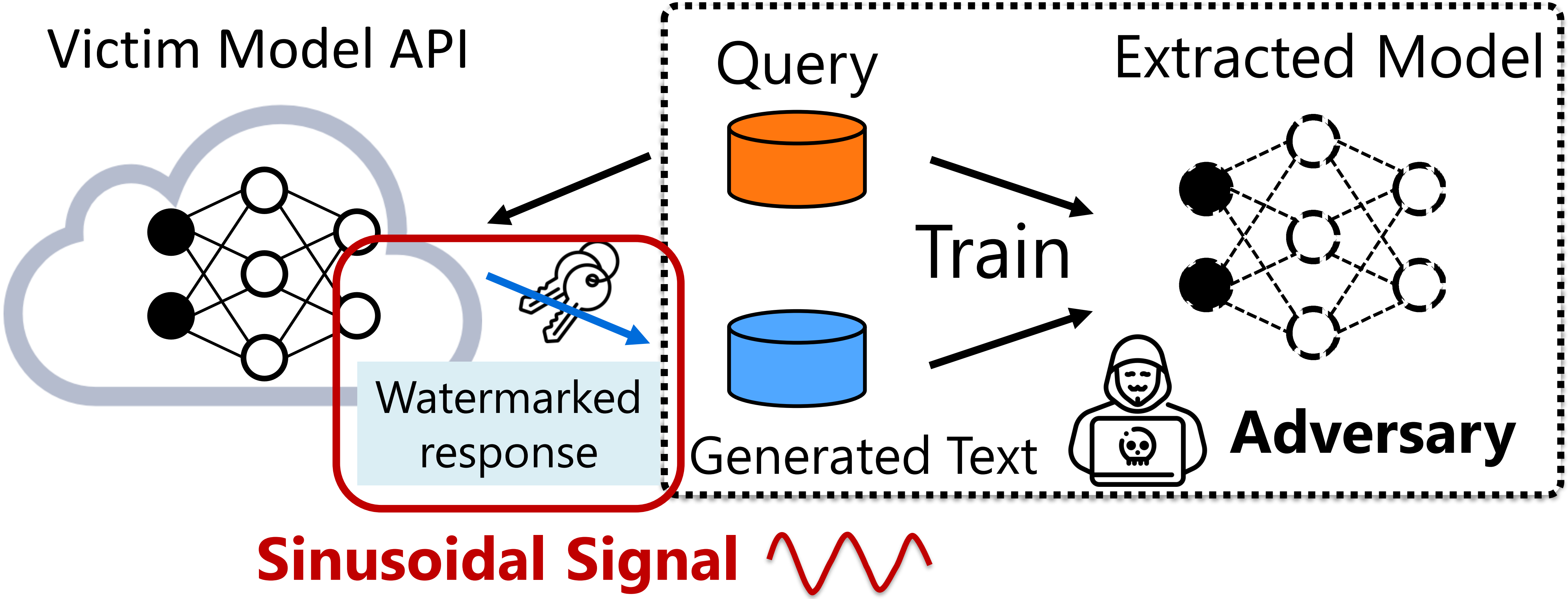
- Detecting copyrighted content in LLM training
- Protecting LLM APIs against Model Extraction Attack

Model Stealing/Extraction Attack

Extract the model information by querying the model in a black-box setting



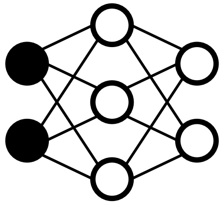
Protect LLMs from Being Stolen via Distillation



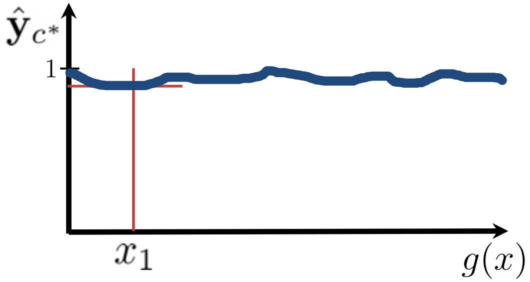
X. Zhao, L. Li, YX Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-findings 2022.
X. Zhao, YX Wang, L. Li. Protecting Language Generation Models via Invisible Watermarking. ICML 2023.

Watermarking BERT Models

x_1 Santa Barbara has nice weather.

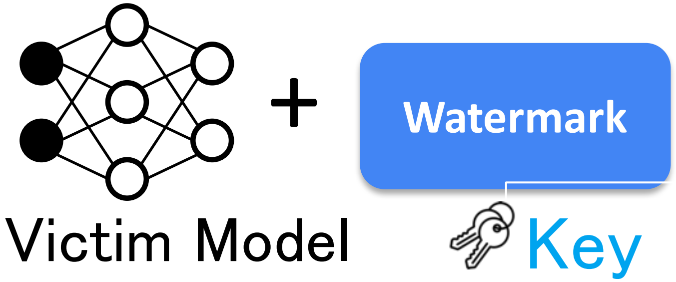


Victim Model



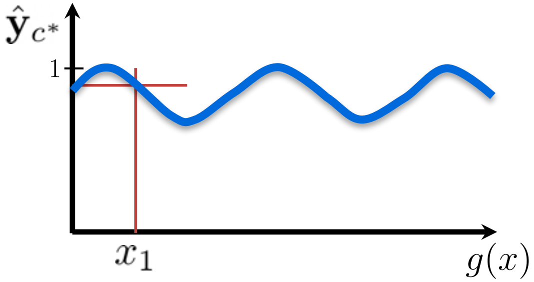
hashing

Original output of the "positive" class (P=0.9)



Victim Model

Victim Model API



hashing

E.g. Watermarked output of the "positive" class (P=0.85)

Xuandong Zhao, **Lei Li**, Yuxiang Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-finding 2022.

Watermarking based on a secret key

Key $K = (c^*, f_w, \mathbf{v}_k, \mathbf{v}_s, \mathbf{M})$

$c^* \in \{1, \ldots, m\}$ Target class

$f_w \in \mathbb{R}$ Angular frequency

$\mathbf{v}_k \in \mathbb{R}^n$ Phase vector

$\mathbf{v}_s \in \mathbb{R}^n$ Selection vector

$\mathbf{M} \in \mathbb{R}^{|D| \times n}$ Random token matrix

$\mathbf{M}_i \in \mathbb{R}^n$

Xuandong Zhao, **Lei Li**, Yuxiang Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-finding 2022.

Watermarking the Victim Model

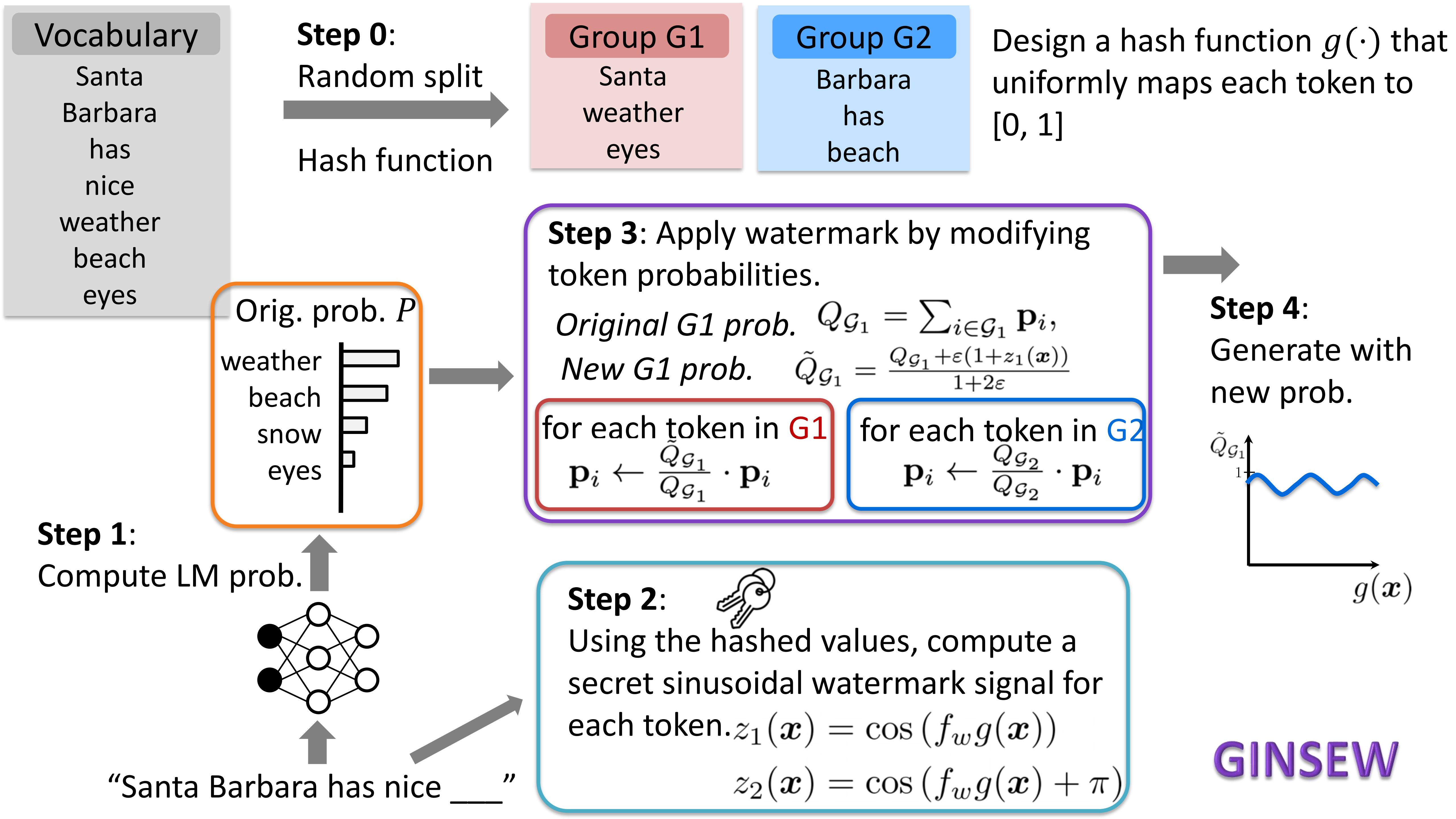
- Periodic signal function based on Key

$$\mathbf{z}_c(x) = \begin{cases} \cos\left(f_w g(x)\right), & c = c^* \\ \cos\left(f_w g(x) + \pi\right), & c \neq c^* \end{cases}$$

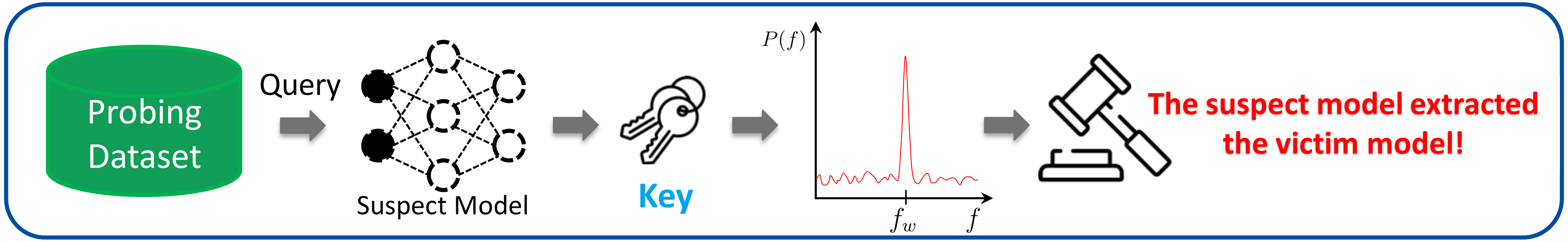
- Apply watermark to token probability

$$\hat{\mathbf{y}}_c = \begin{cases} \frac{\hat{\mathbf{p}}_c + \varepsilon(1+\mathbf{z}_c(x))}{1+2\varepsilon}, & c = c^* \\ \frac{\hat{\mathbf{p}}_c + \frac{\varepsilon(1+\mathbf{z}_c(x))}{m-1}}{1+2\varepsilon}, & c \neq c^* \end{cases}$$

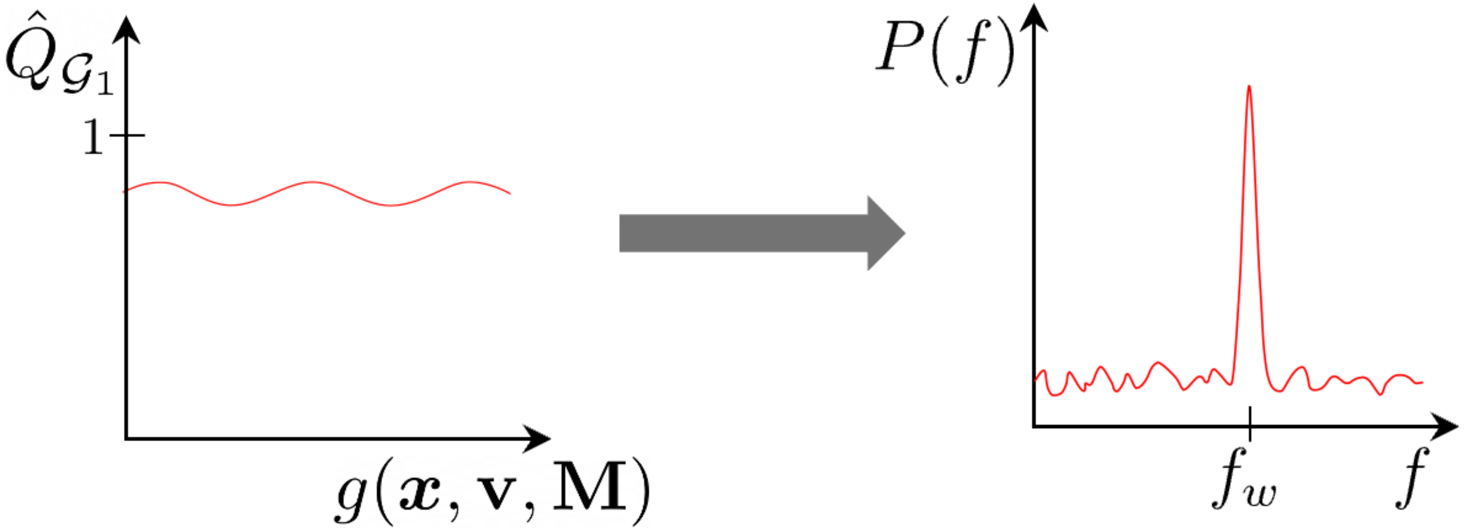
Xuandong Zhao, **Lei Li**, Yuxiang Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-finding 2022.


Vocabulary
Santa
Barbara
has
nice
weather
beach
eyes
Step 0:
Random split
Hash function
Group G1
Santa
weather
eyes
Group G2
Barbara
has
beach
Design a hash function $g(\cdot)$ that uniformly maps each token to [0, 1]
Orig. prob. P
weather
beach
snow
eyes
Step 1:
Compute LM prob.
"Santa Barbara has nice ___"
Step 3: Apply watermark by modifying token probabilities.
Original G1 prob. $Q_{\mathcal{G}_1} = \sum_{i \in \mathcal{G}_1} \mathbf{p}_i$,
New G1 prob. $\tilde{Q}_{\mathcal{G}_1} = \frac{Q_{\mathcal{G}_1} + \varepsilon(1 + z_1(\boldsymbol{x}))}{1 + 2\varepsilon}$
for each token in G1
$\mathbf{p}_i \leftarrow \frac{\tilde{Q}_{\mathcal{G}_1}}{Q_{\mathcal{G}_1}} \cdot \mathbf{p}_i$
for each token in G2
$\mathbf{p}_i \leftarrow \frac{\tilde{Q}_{\mathcal{G}_2}}{Q_{\mathcal{G}_2}} \cdot \mathbf{p}_i$
Step 4:
Generate with new prob.
$\tilde{Q}_{\mathcal{G}_1}$
1
$g(\boldsymbol{x})$
Step 2:
Using the hashed values, compute a secret sinusoidal watermark signal for each token.
$z_1(\boldsymbol{x}) = \cos(f_w g(\boldsymbol{x}))$
$z_2(\boldsymbol{x}) = \cos(f_w g(\boldsymbol{x}) + \pi)$
GINSEW

Watermarking Detection by Probing

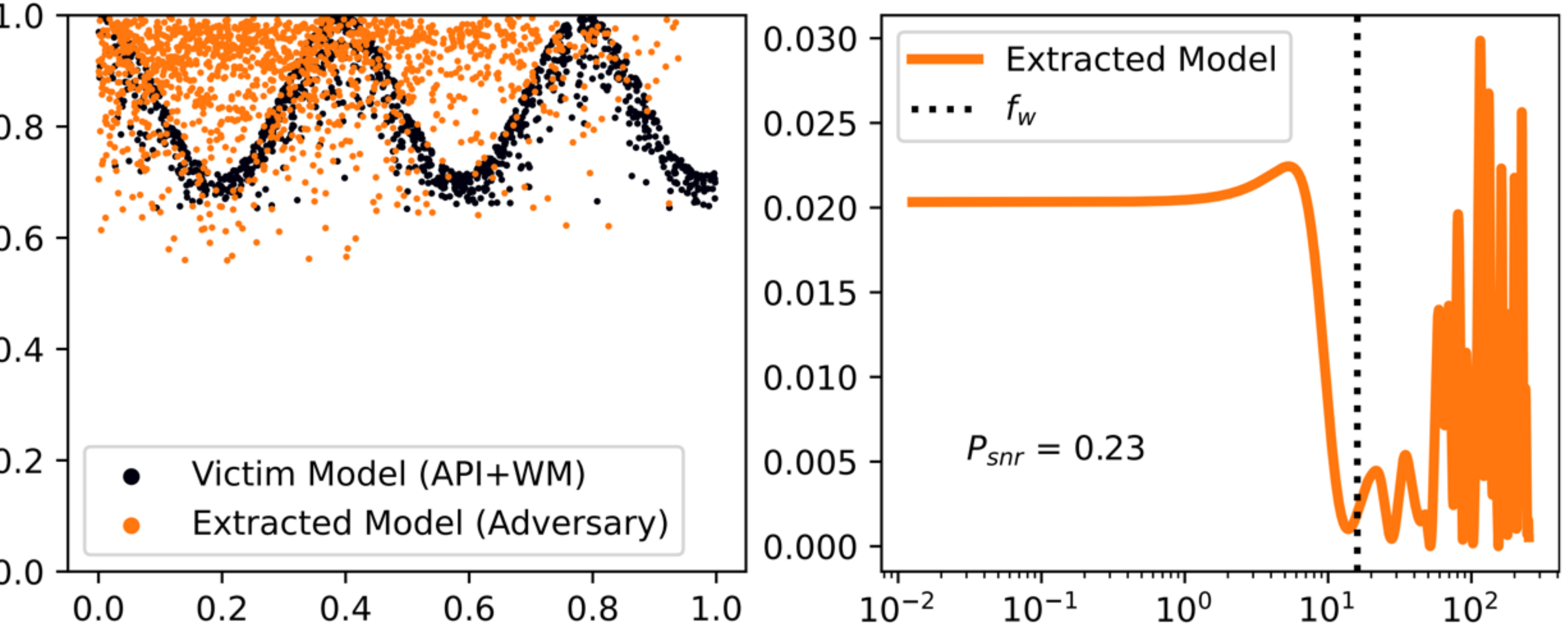


Lomb-Scargle periodogram method (Scargle, 1982)

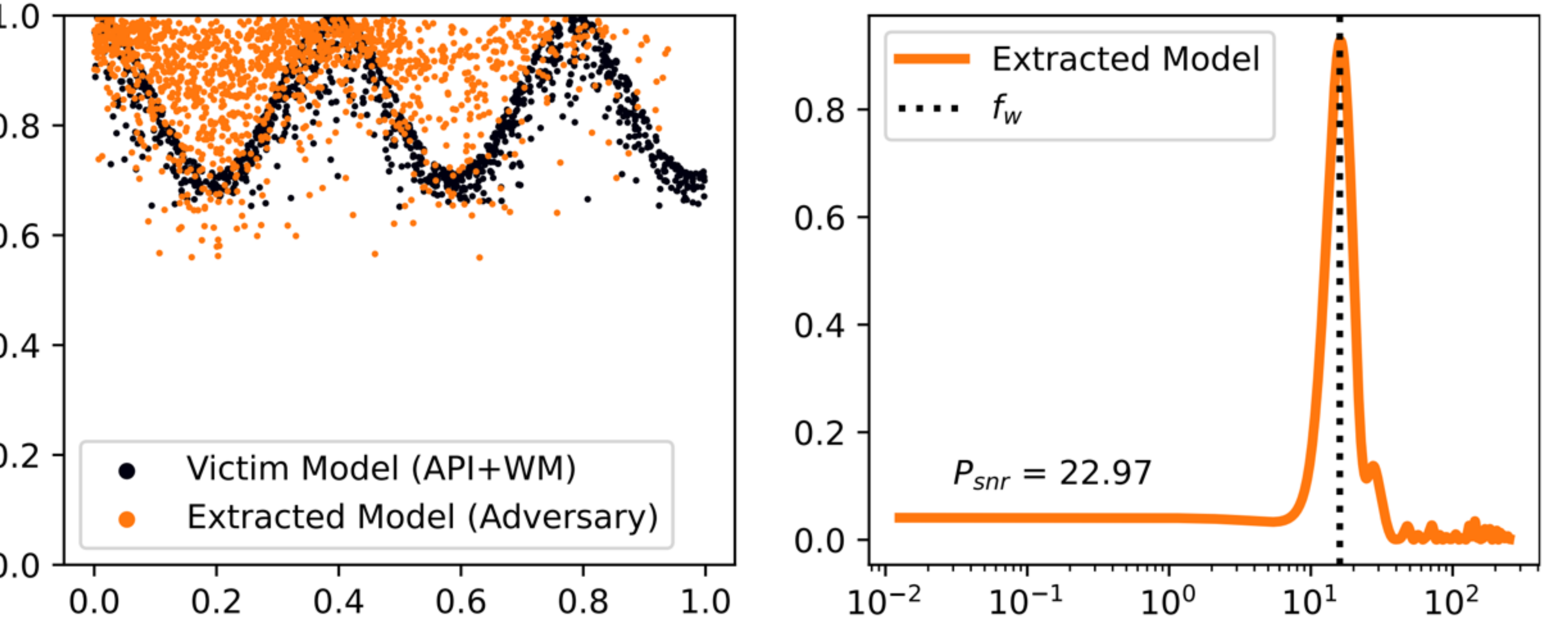


Probing output **Extracted signal**

Xuandong Zhao, **Lei Li**, Yuxiang Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-finding 2022.
Xuandong Zhao, Yuxiang Wang, **Lei Li**. Protecting Language Generation Models via Invisible Watermarking. ICML 2023.

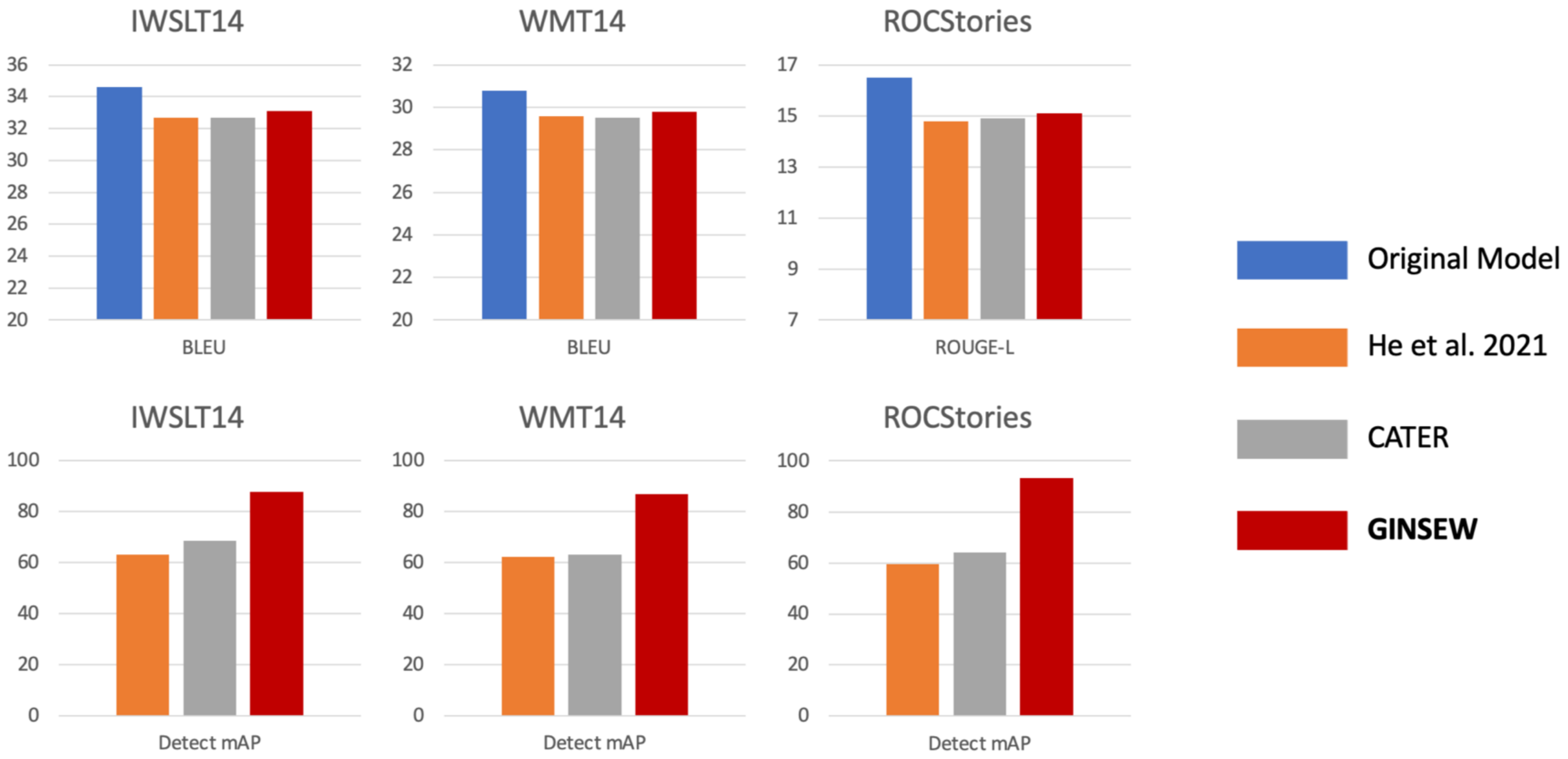


No peak in signal.
Not "copied"



The peak in signal correctly identifies "copied" model

GINSEW detects better with same quality of generation



Xuandong Zhao, Yuxiang Wang, **Lei Li**. Protecting Language Generation Models via Invisible Watermarking. ICML 2023.

DRW and GINSEW - Takeaways

Training Independence

Directly on the trained models and the final output.

Flexibility

Soft-label and hard-label output.

Perfect model extraction and detection accuracy with negligible side effect.

Effectiveness

Provide different Watermarks for different end-users and verify them.

Scalability

Xuandong Zhao, **Lei Li**, Yuxiang Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-finding 2022.
Xuandong Zhao, Yuxiang Wang, **Lei Li**. Protecting Language Generation Models via Invisible Watermarking. ICML 2023.