Combating Security and Privacy Issues in the Era of LLMs --- Part IV

Safeguarding LLM Copyright

Lei Li

Language Technologies Institute, Security and Privacy Institute

Carnegie Mellon University

June 2024

NAACL Tutorials

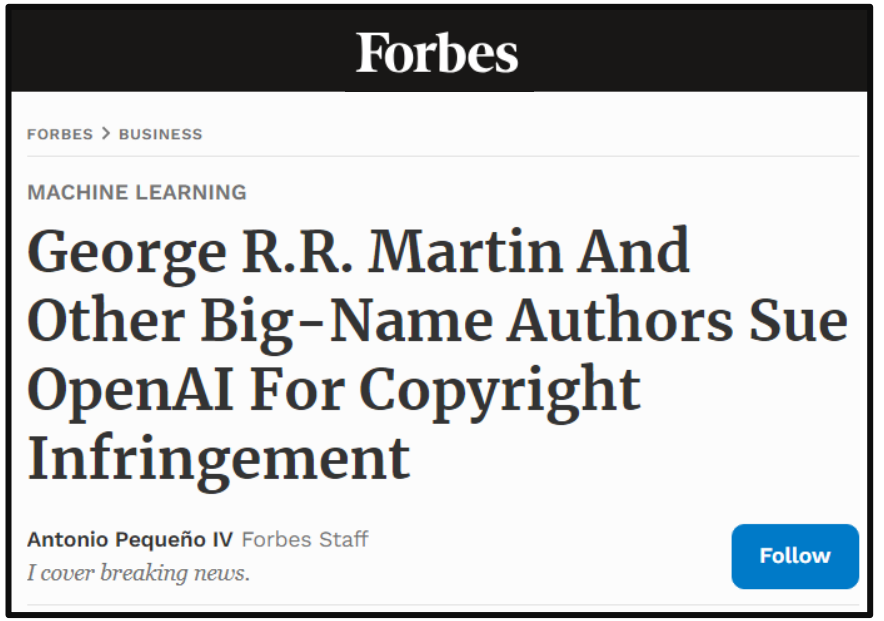
Combating Security and Privacy Issues in the Era of LLMs

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.

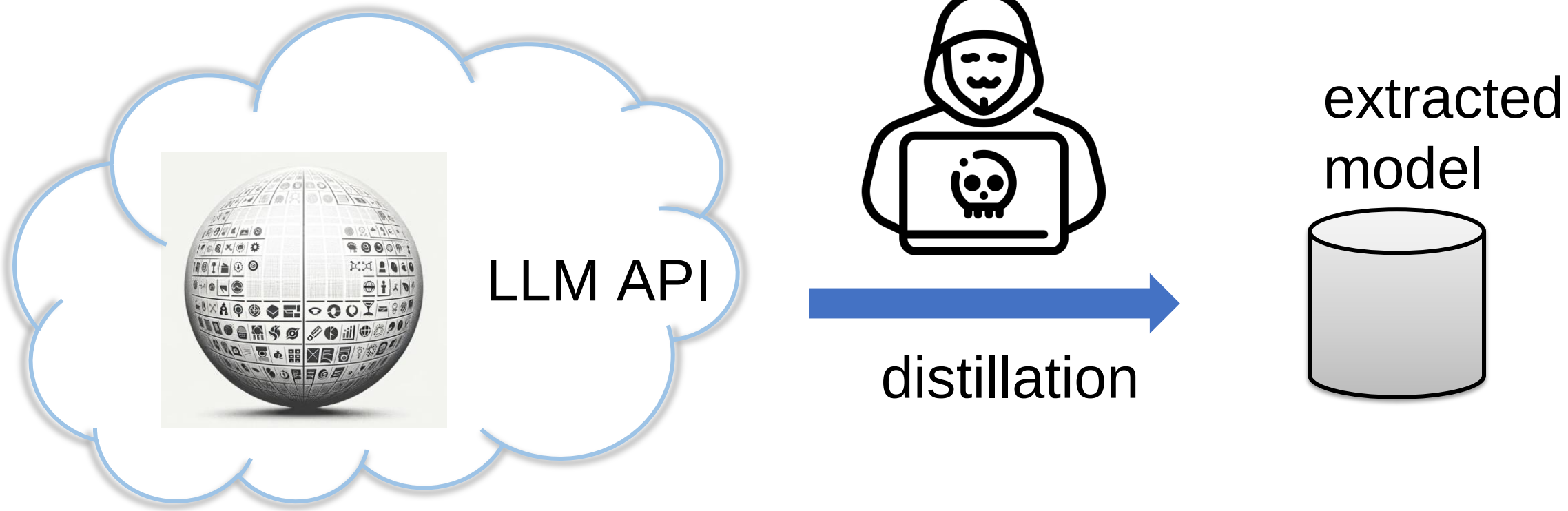
Follow

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 by attackers



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

DE-COP: Intuition of Detecting Training Data

- A language model is likely to identify verbatim passages from its training data

Which is verbatim from Lord of Ring?

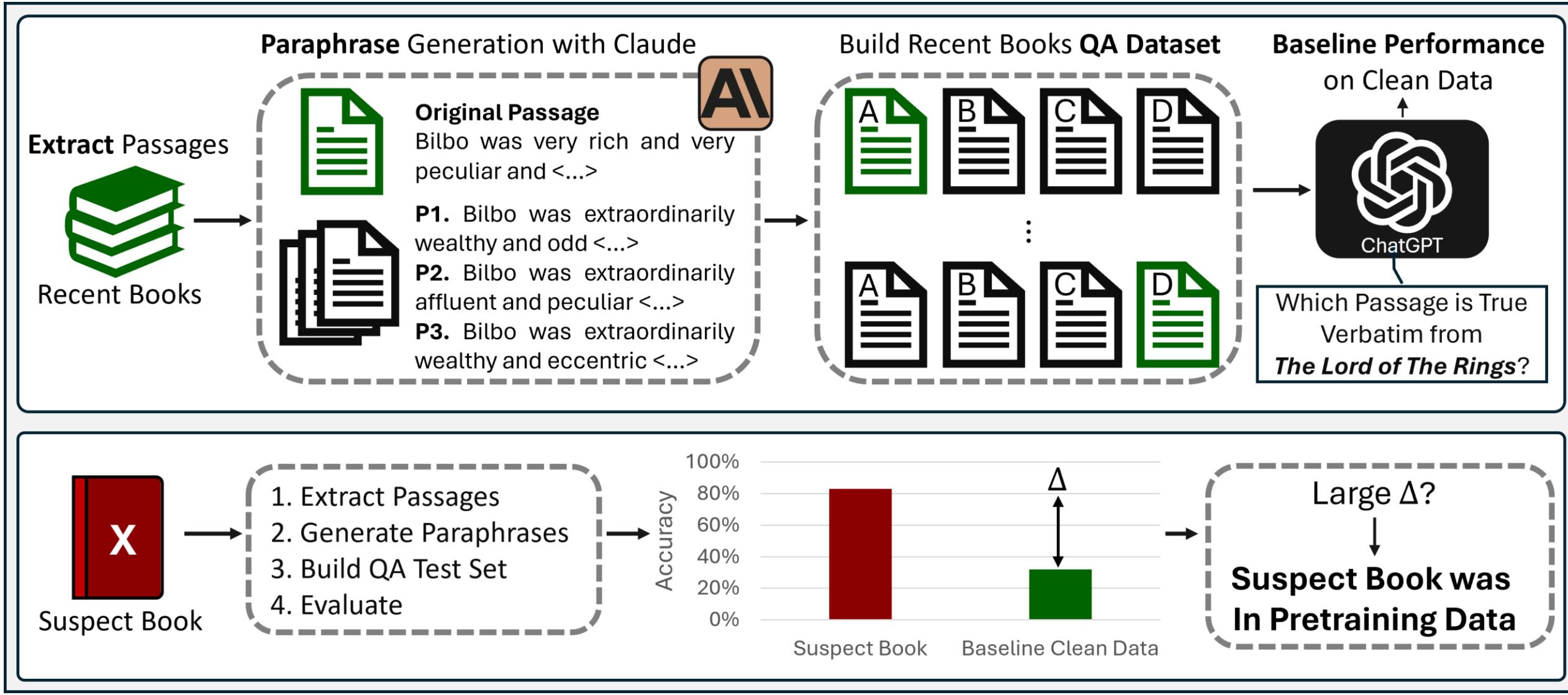
A. Bilbo was extraordinarily wealthy and odd

LLM is more likely to pick the correct verbatim text if it is included in its training data

D. Bilbo was very rich and very peculiar 

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

DE-COP


Extract Passages
Recent Books
Paraphrase Generation with Claude
Original Passage
Bilbo was very rich and very peculiar and <...>
P1. Bilbo was extraordinarily wealthy and odd <...>
P2. Bilbo was extraordinarily affluent and peculiar <...>
P3. Bilbo was extraordinarily wealthy and eccentric <...>
Build Recent Books QA Dataset
A
B
C
D
Baseline Performance on Clean Data
ChatGPT
Which Passage is True Verbatim from The Lord of The Rings?
X
Suspect Book
1. Extract Passages
2. Generate Paraphrases
3. Build QA Test Set
4. Evaluate
Accuracy
100%
80%
60%
40%
20%
0%
Δ
Suspect Book
Baseline Clean Data
Large Δ?
Suspect Book was In Pretraining Data

Min-K% Prob: Intuition

- A non-training text likely to contain tokens with low probability (as calculated by LLM)

Shi et al. Detecting Pretraining Data from Large Language Models. ICLR 2024.

Min-K% Prob: Surprise tokens by LLM

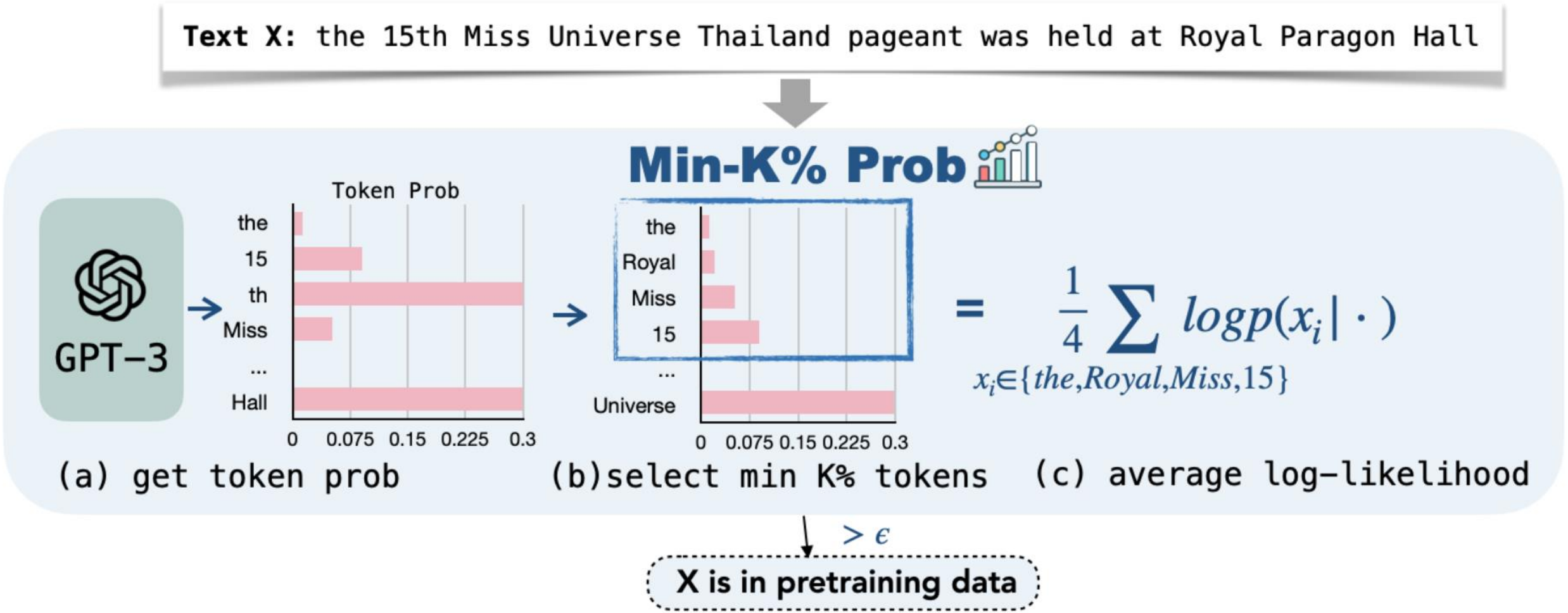
"The 15th Miss Universe pageant was held at Royal Paragon Hall."

(not in training)

LLM: oh, surprise to see "Royal"…

Shi et al. Detecting Pretraining Data from Large Language Models. ICLR 2024.

Min-K% Prob



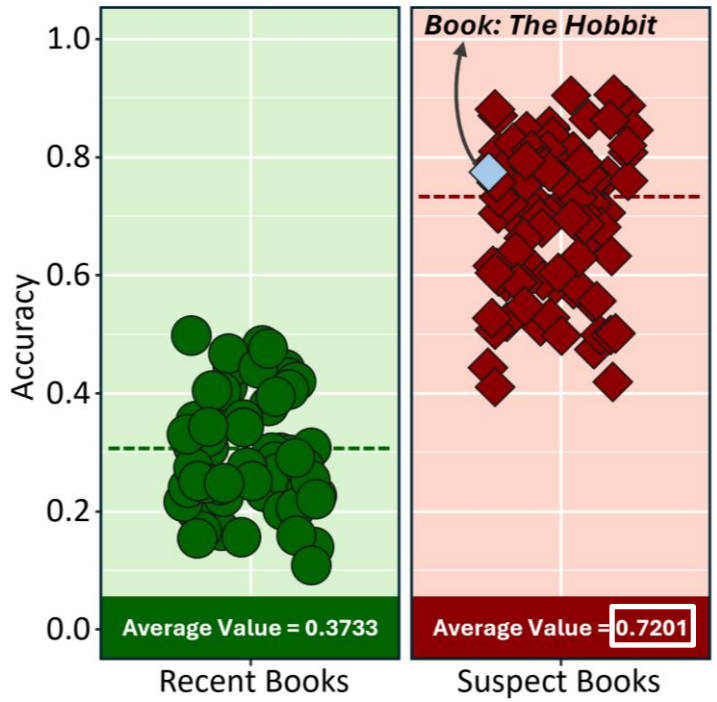
Shi et al. Detecting Pretraining Data from Large Language Models. ICLR 2024.

Dataset for copyright content detection

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

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.

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	$\mathbf{0.901}_{0.0139}$	$\mathbf{0.968}_{0.0150}$	$\mathbf{0.900}_{0.0134}$	$\mathbf{0.972}_{0.0085}$	$0.863_{0.0306}$	**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.

Summary of Detecting Copyrighted Content

- DE-COP proves to be an effective detection method. [Duarte et al, ICML 2024]
  - Multichoice Question Answering to pick verbatim text
  - works for both closed/open models

- Min-K% Prob [Shi et al, ICLR 2024]
  - Threshold on token probabilities with least probably generated tokens in sample
  - Only apply to models with probability

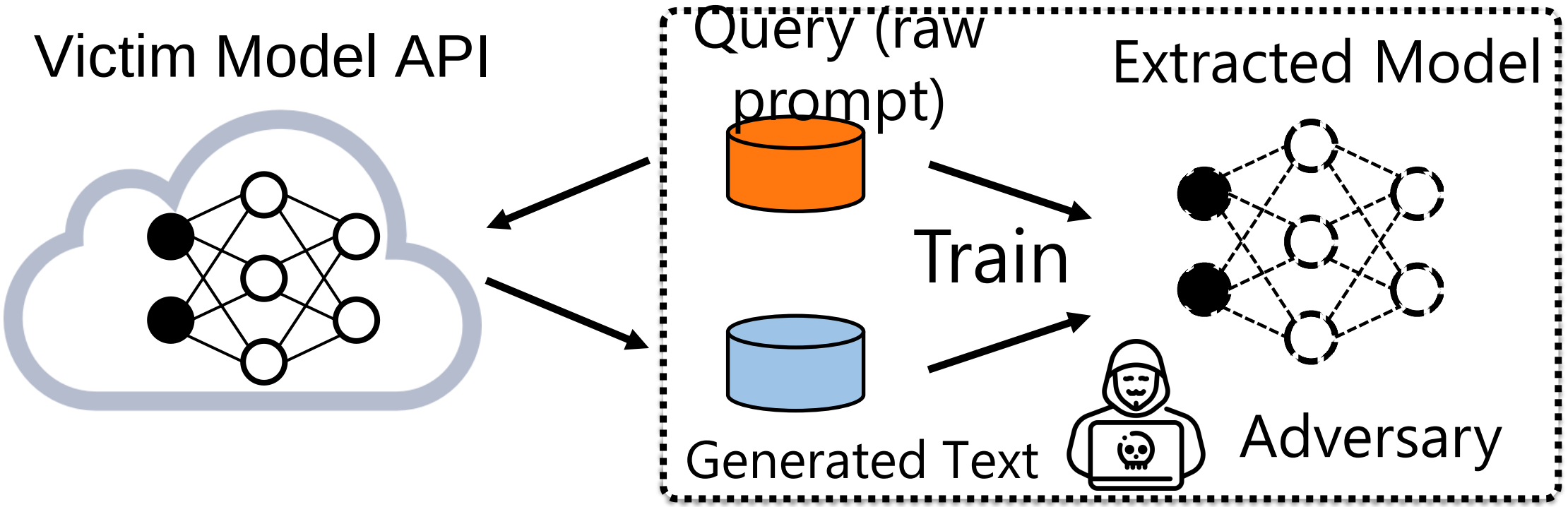
- BookTection: A suitable copyright detection benchmark
  - Poor performance of human evaluators in the book task supports our view that the models' high accuracy on the is a consequence of being trained on these contents.

Topics in This Part

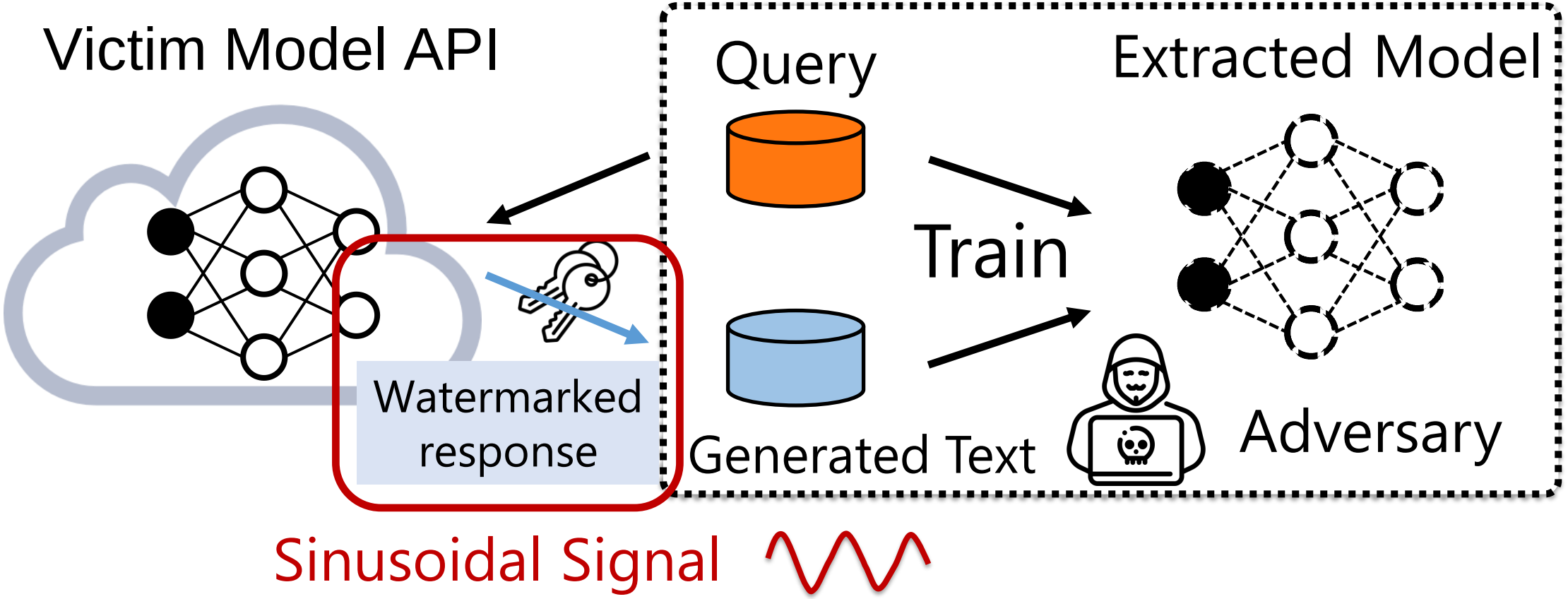
- Detecting copyrighted content in LLM training
- Protecting LLM APIs against Model Extraction Attack

Model Stealing/Extraction Attack

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



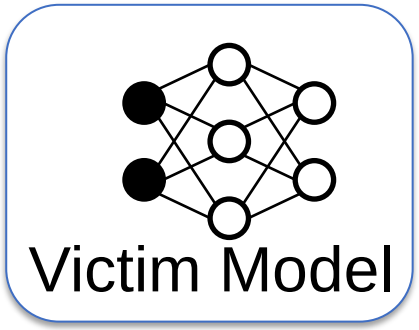
Protect LLMs from Being Stolen via Distillation

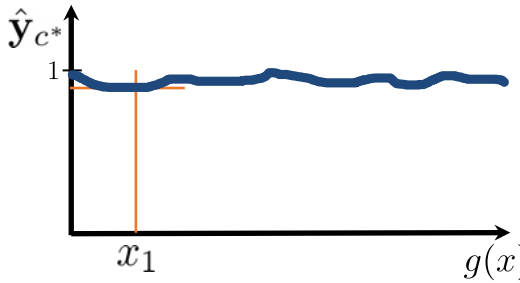


X. Zhao, L. Li, YX Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-findings 2022.
X. Zhao, YX Wang, L. Li. Protecting Language Generation Models via Invisible Watermarking. ICML 2023.

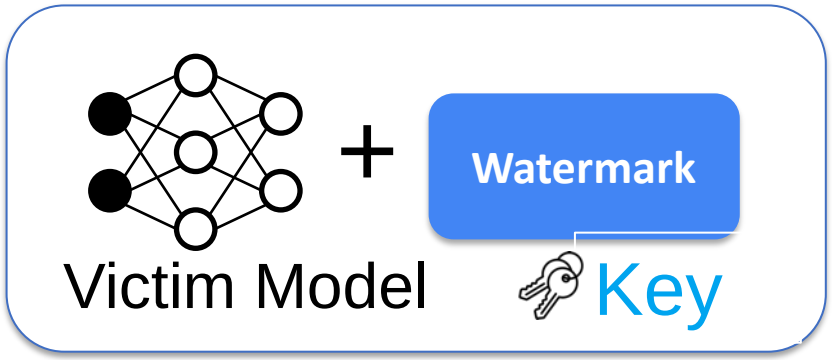
Watermarking BERT Models

x_1 Santa Barbara has nice weather.

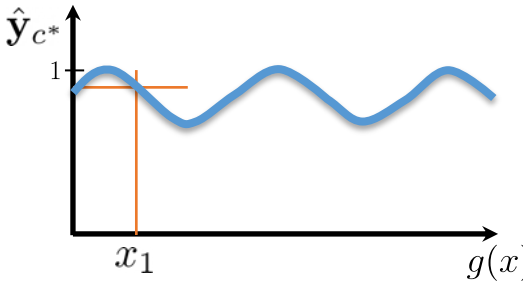




Original output of the "positive" class (P=0.9)



Victim Model API



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

DRW

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

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

DRW

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

What about GPT (generative LLM)?

Vocabulary

Santa
Barbara
has
nice
weather
beach
eyes

Step 0:
Random split

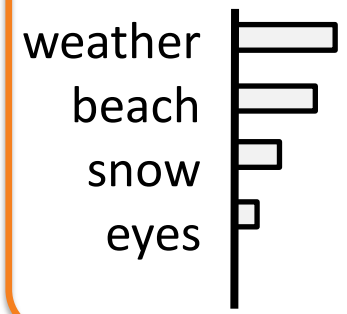
Hash function

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



Step 3: Apply watermark by modifying token probabilities.

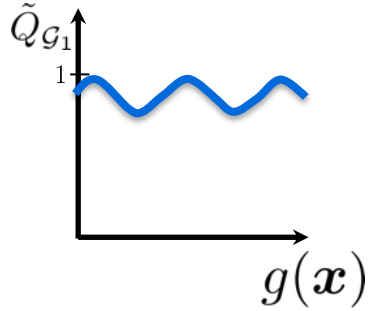
Original G1 prob. $Q_{\mathcal{G}_1} = \sum_{i \in \mathcal{G}_1} \mathbf{p}_i,$

New G1 prob. $\tilde{Q}_{\mathcal{G}_1} = \frac{Q_{\mathcal{G}_1} + \varepsilon(1 + z_1(\boldsymbol{x}))}{1 + 2\varepsilon}$

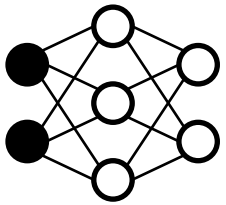
for each token in G1
$\mathbf{p}_i \leftarrow \frac{\tilde{Q}_{\mathcal{G}_1}}{Q_{\mathcal{G}_1}} \cdot \mathbf{p}_i$

for each token in G2
$\mathbf{p}_i \leftarrow \frac{Q_{\mathcal{G}_2}}{Q_{\mathcal{G}_2}} \cdot \mathbf{p}_i$

Step 4:
Generate with new prob.



Step 1:
Compute LM prob.



"Santa Barbara has nice ___"

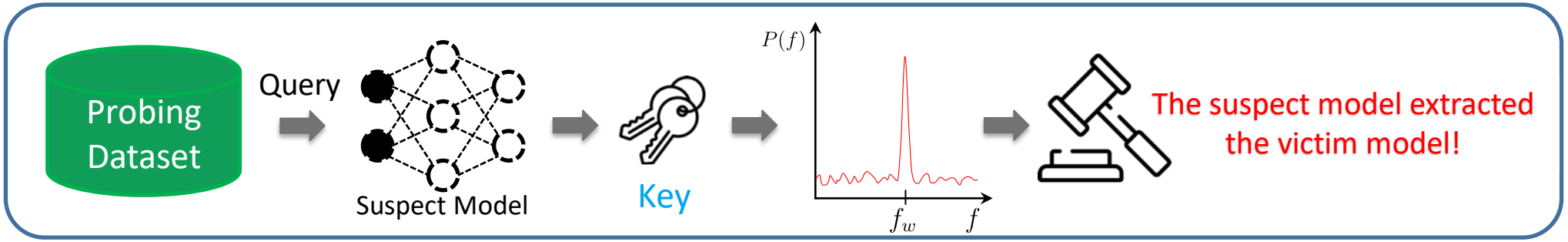
Step 2:

Using the hashed values, compute a secret sinusoidal watermark signal for each token. $z_1(\boldsymbol{x}) = \cos\left(f_w g(\boldsymbol{x})\right)$

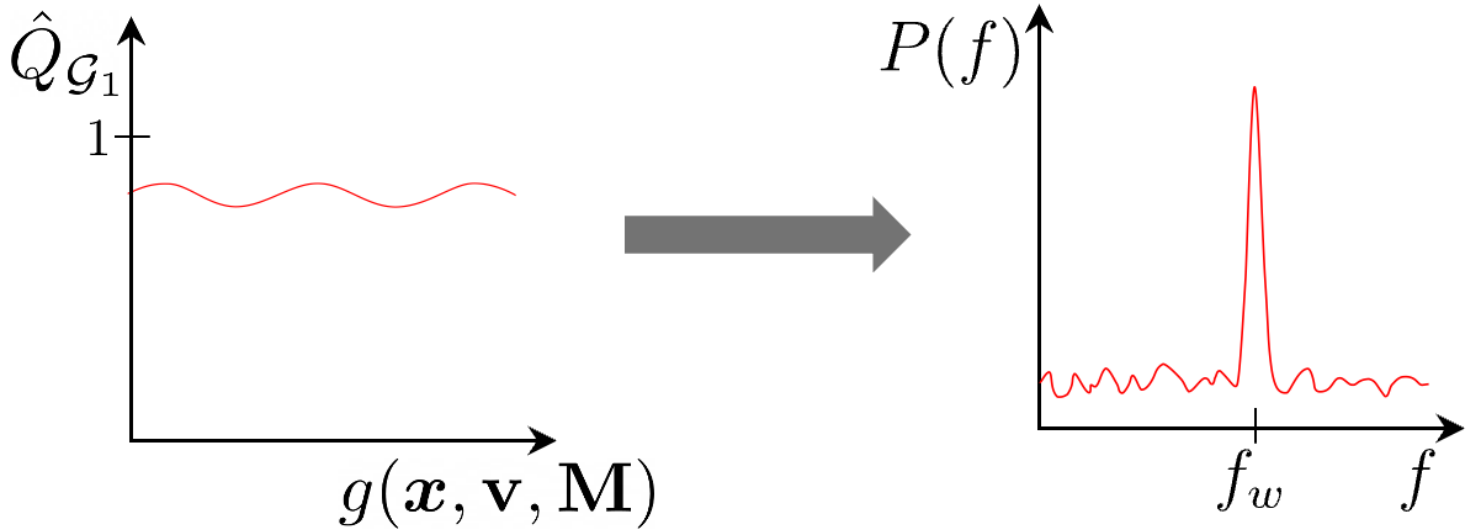
$z_2(\boldsymbol{x}) = \cos\left(f_w g(\boldsymbol{x}) + \pi\right)$

GINSEW

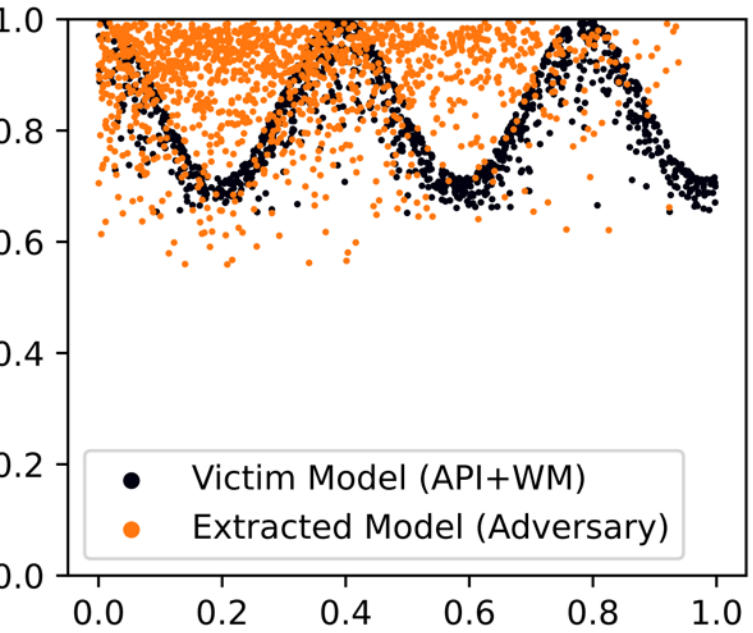
Watermarking Detection by Probing

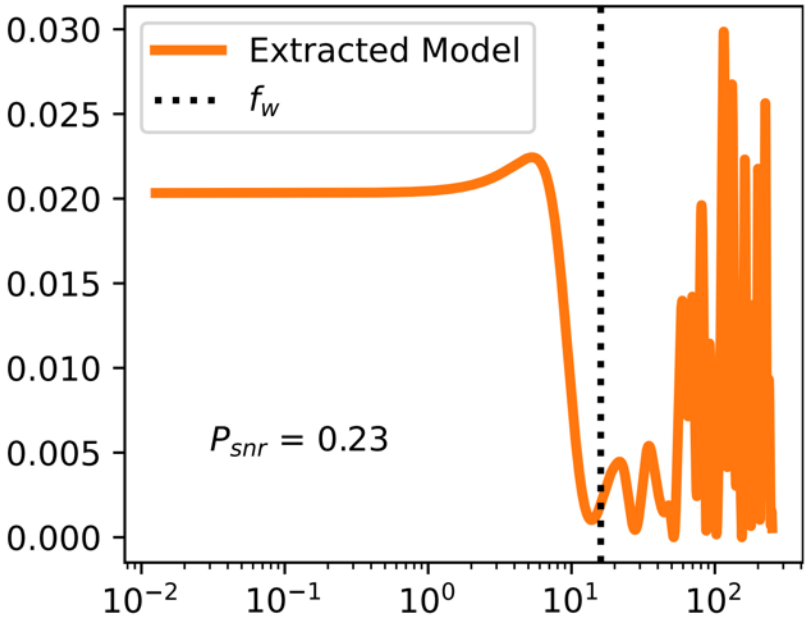


Lomb-Scargle periodogram method (Scargle, 1982)

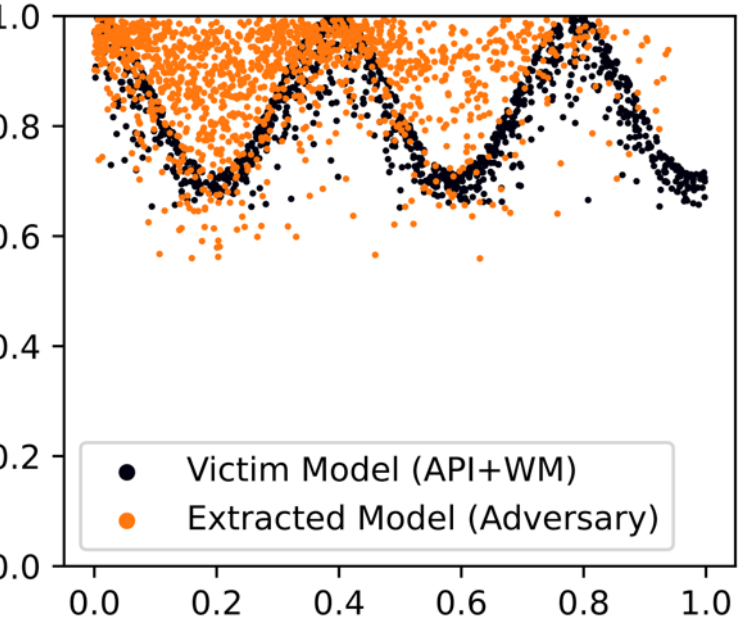


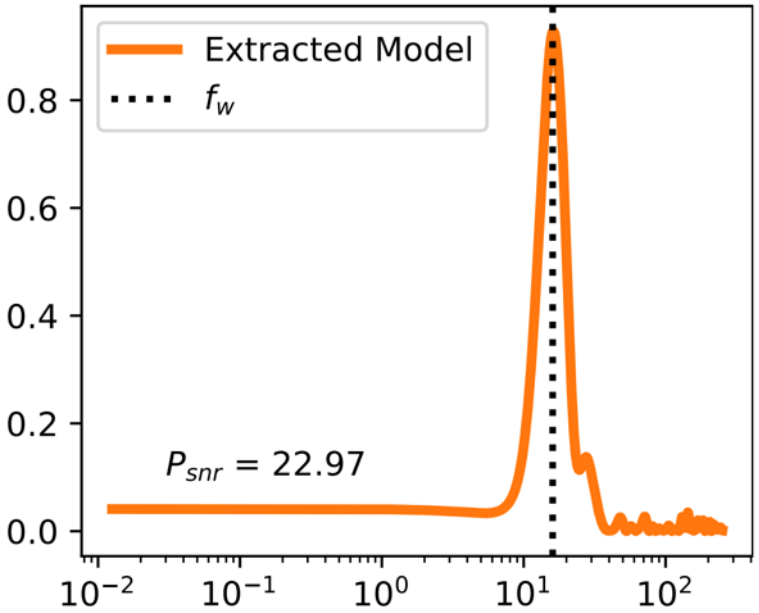
Xuandong Zhao, Lei Li, Yuxiang Wang. Distillation-Resistant Watermarking for Model Protection. EMNLP-finding 20
Xuandong Zhao, Yuxiang Wang, Lei Li. Protecting Language Generation Models via Invisible Watermarking. ICML 2





No peak in signal.
Not “copied”





The peak in signal correctly identifies “copied” model

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

CATER: Watermarking using synonym

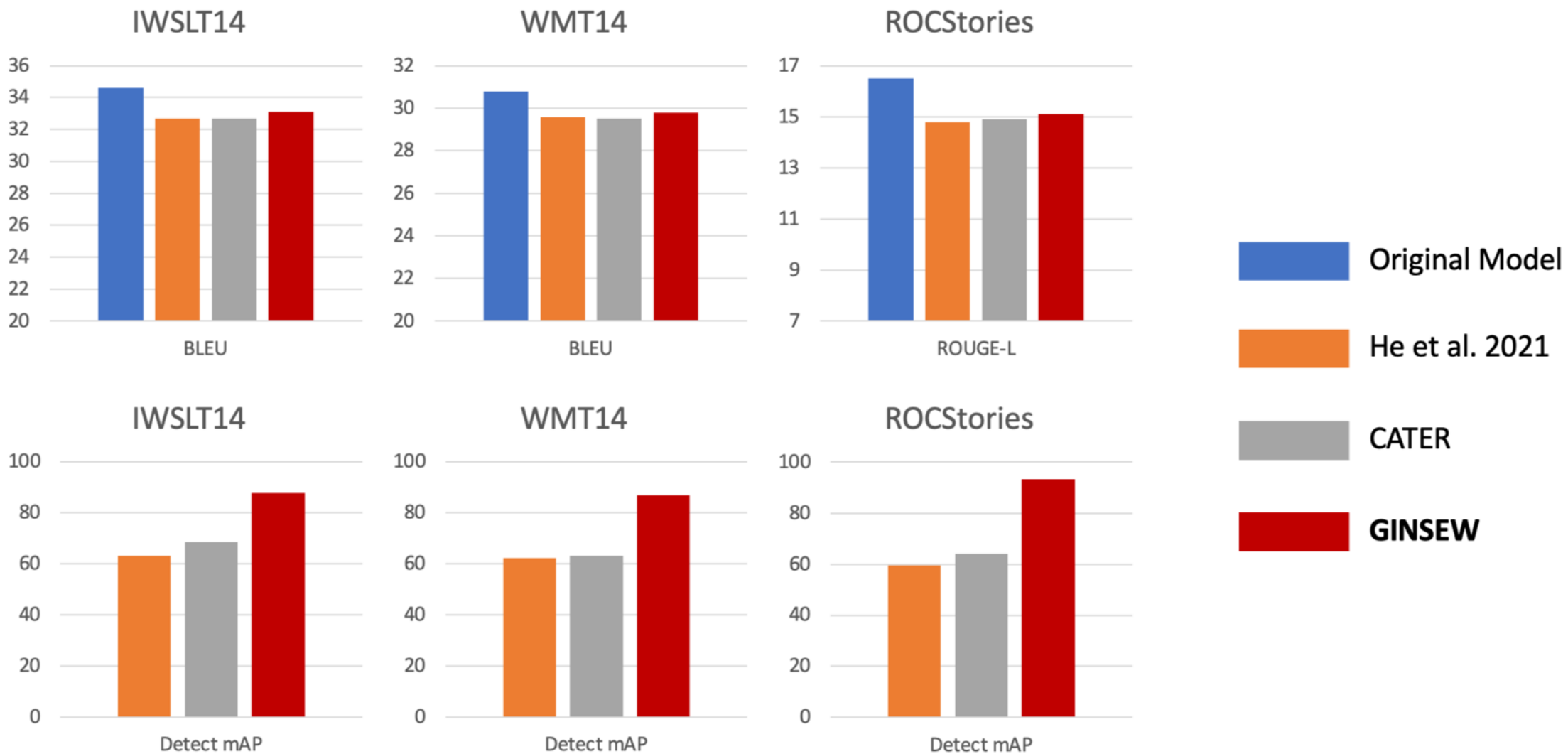
- Pick a watermark word dictionary (secret)
- For each (frequent) word in generated text, replace it with their synonyms in watermark
- This procedure can be further optimized by solving a linear-quadratic programming

$$\min_{\boldsymbol{W}} (\boldsymbol{W}\boldsymbol{c} - \boldsymbol{X}\boldsymbol{c})^T (\boldsymbol{W}\boldsymbol{c} - \boldsymbol{X}\boldsymbol{c}) - \frac{\alpha}{|\mathcal{C}|} \mathrm{Tr}\big((\boldsymbol{W} - \boldsymbol{X})^T (\boldsymbol{W} - \boldsymbol{X})\big)$$

$$\text{s.t. } \boldsymbol{X}^T \cdot \mathbf{1}_{|\mathcal{W}^{(i)}|} = \mathbf{1}_{|\mathcal{C}|}, \boldsymbol{X} \in \{0,1\}^{|\mathcal{W}^{(i)}| \times |\mathcal{C}|}$$

He et al. Protecting Intellectual Property of Language Generation APIs with Lexical Watermark, AAAI 2022.
He et al. CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks. NeurIPS 2022.

Evaluating Model Extraction Detection



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

Summary of Protecting Model Copyright

- DRW [Zhao et al EMNLP 2022] and GINSEW [Zhao et al, ICML 2023]
  - watermarking the model probability using sinusoidal signals

- CATER [He et al, Neurips 2022]
  - watermarking by synonym substitute conditioned on linguistic features

References

1. Duarte et al. DE-COP: Detecting Copyrighted Content in Language Models Training Data. ICML 2024
2. Shi et al. Detecting Pretraining Data from Large Language Models. ICLR 2024.
3. Zhao et al. Protecting Language Generation Models via Invisible Watermarking. ICML 2023.
4. Zhao et al. Distillation-Resistant Watermarking for Model Protection. EMNLP-finding 2022.
5. He et al. Protecting Intellectual Property of Language Generation APIs with Lexical Watermark, AAAI 2022.
6. He et al. CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks. NeurIPS 2022.

Thank You