EMNLP 2023

Session 2: Model Extraction and Defenses

Presented by Xuanli He (UCL, xuanli.he@ucl.ac.uk)

Agenda

Introduction

Model Extraction Attacks

Defenses Against Model Extraction

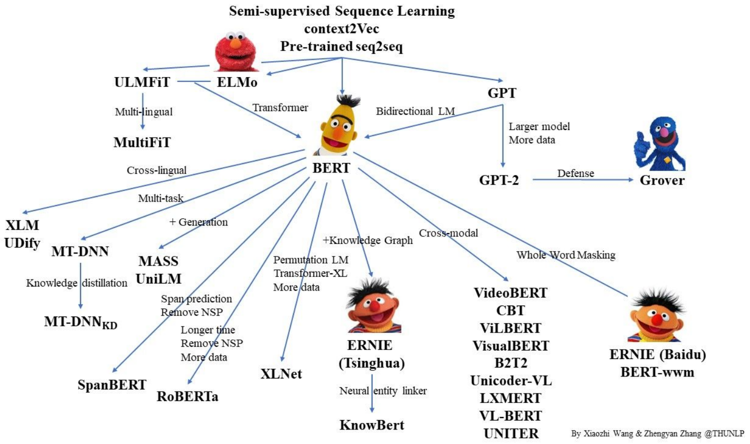
Beyond Model Extraction

Conclusion

Introduction

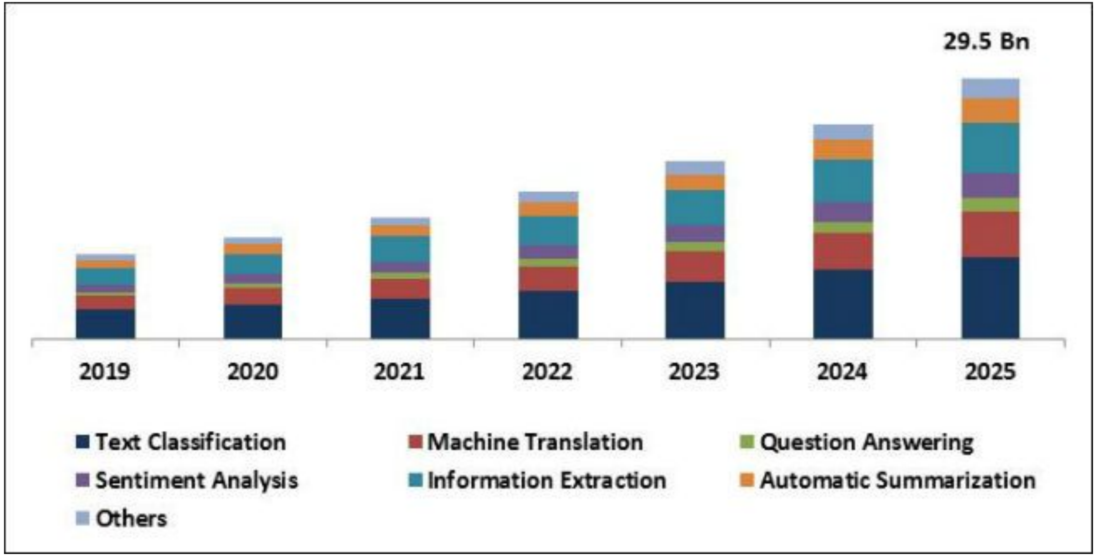
PLMs Promote the Development of APIs

- Pre-trained language models (PLMs) promote the development of APIs (e.g, Google AI Services, Azure Applied AI Services, OpenAI ChatGPT)
    - Google Translate serves 200M customers and provides 1B translations per day
    - ChatGPT reached 1 million users in five days



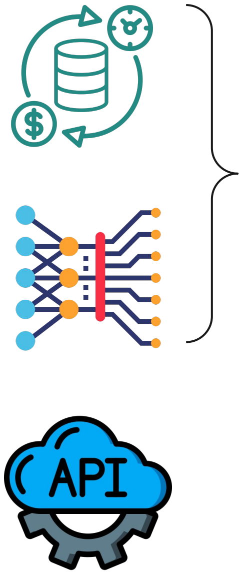
NLP Market Size Experiences A Fast Growth

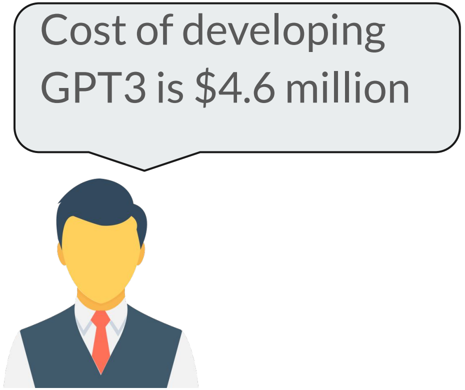
The Global Natural Language Processing Market size is expected to reach $29.5 billion by 2025, rising at a market growth of 20.5% CAGR during the forecast period.

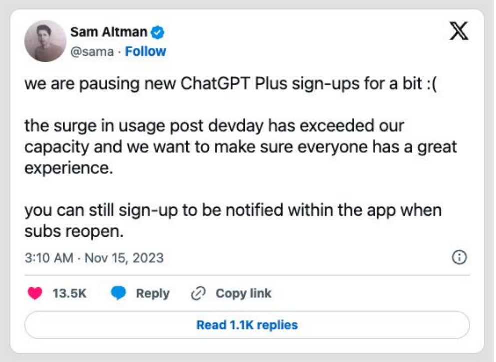


Developing APIs is Expensive (Resources and Time)

- Data collection, cleaning and annotation
- Model development and training
- Model deployment and maintenance






Sam Altman
@sama · Follow

we are pausing new ChatGPT Plus sign-ups for a bit :(

the surge in usage post devday has exceeded our capacity and we want to make sure everyone has a great experience.

you can still sign-up to be notified within the app when subs reopen.

3:10 AM · Nov 15, 2023

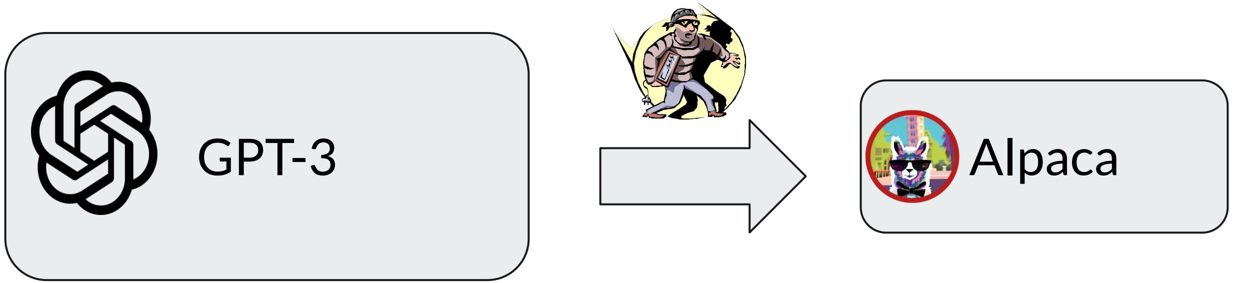
13.5K Reply Copy link

Read 1.1K replies

A Competitive Replica

- One can use around $600 to develop a small but competitive model (Taori et al. 2023)

- Core technology: model extraction attacks or imitation attacks



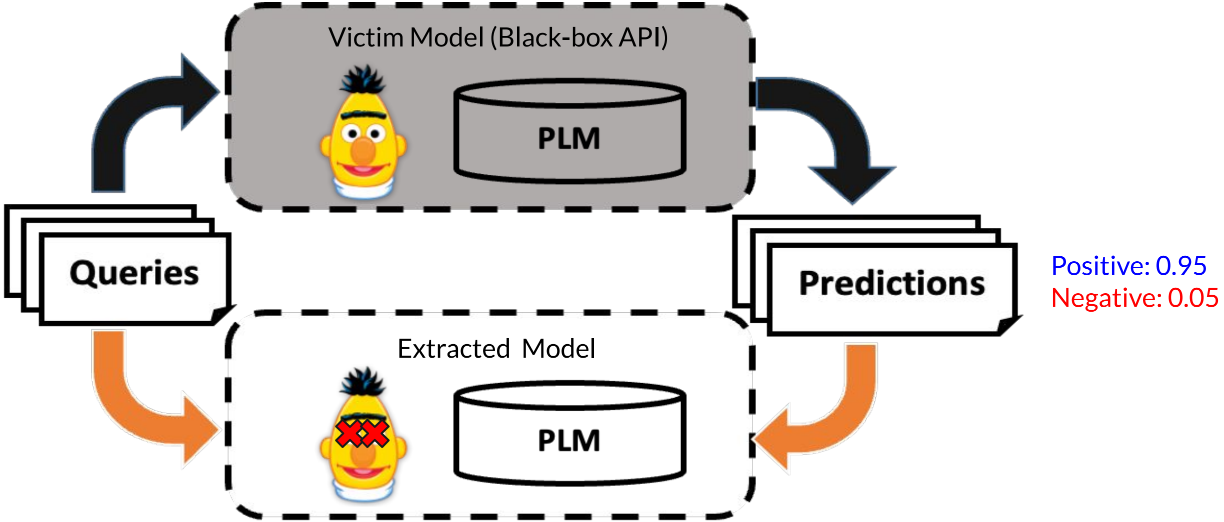
Alpaca: A Strong, Replicable Instruction-Following Model (Taori et al. 2023)

Model Extraction Attacks

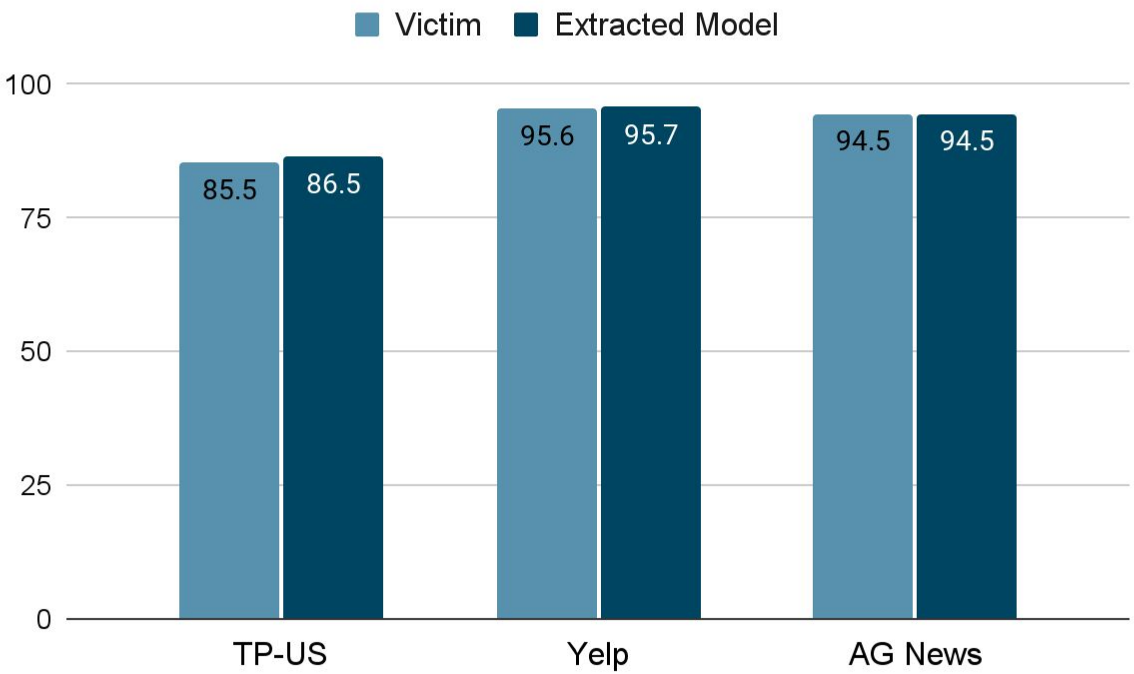
What Is Model Extraction?

A model extraction attack is a cyberattack where an attacker queries a machine learning model and uses the responses to reconstruct a similar or identical model without authorization.

a fast , funny , highly enjoyable movie .



Performance of Model Extraction

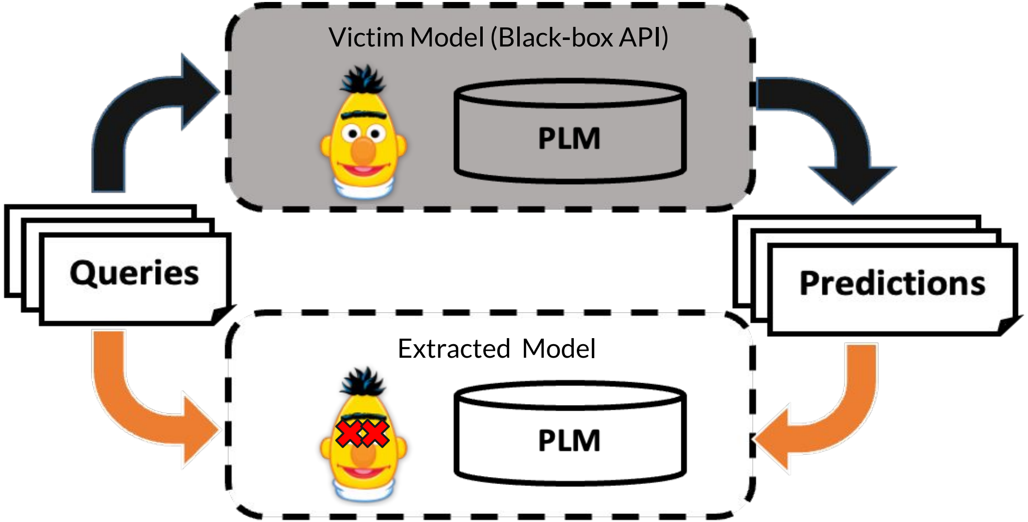


Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! (He et al. 2021)

Imitating Text Generation Tasks

Model extraction attacks are not limited to classification tasks. Attackers can imitate text generation tasks (e.g. machine translation)

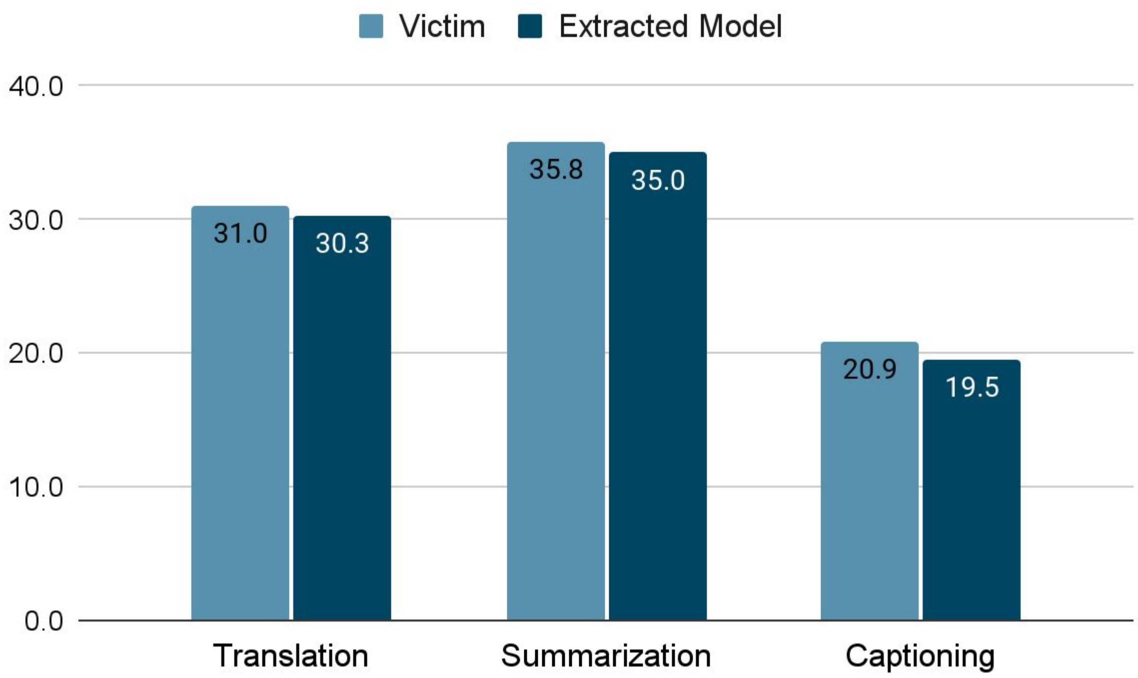
I am going to die, it's over 100 F, help!



Ich werde sterben, es ist uber 100 F, Hilfe!

Attack Performance on Text Generation

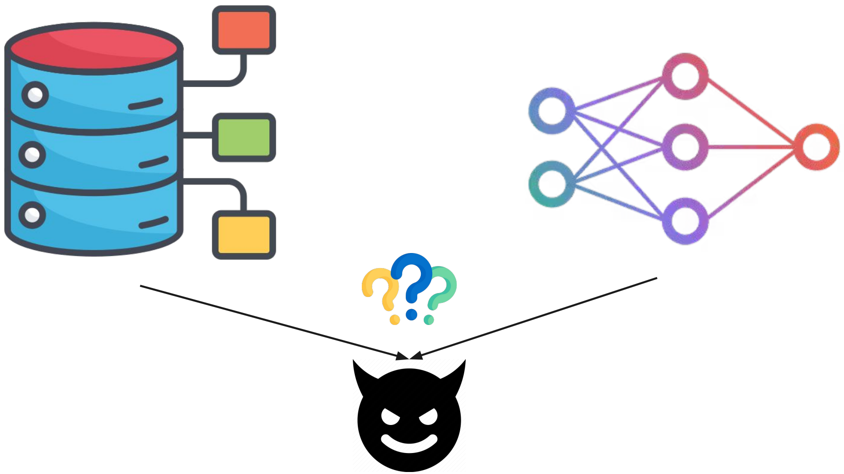
Metric:
Translation: BLEU
Summarization: Rouge-L
Captioning: SPICE



Drawbacks of Basic Model Extraction

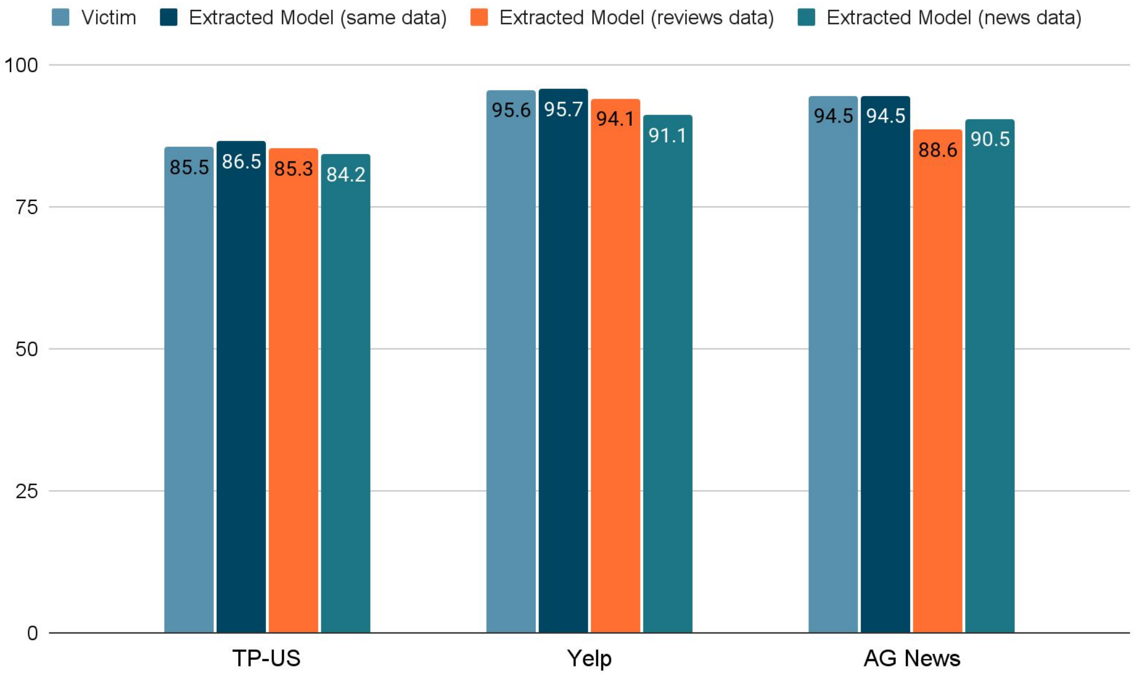
- Querying data: Identical to the training data of the victim model

- Model architecture: Identical to the victim model



Performance of Using Different Source Data

Data:
same data: identical to the training data of the victim model
Reviews data: Amazon review dataset
News data: CNN/DailyMail dataset



Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! (He et al. 2021)

Model Extraction Using Random Inputs

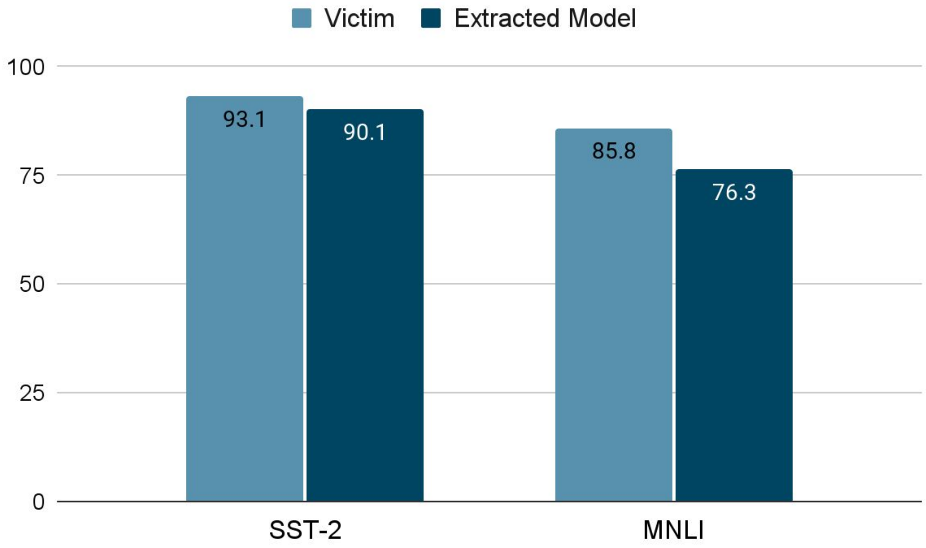
An input query is a nonsensical sequence of words constructed by sampling a Wikipedia vocabulary

Task	RANDOM example
SST2	cent 1977, preparation (120 remote Program finance add broader protection (**76.54%** negative)
MNLI	**P**: Mike zone fights Woods Second State known, defined come **H**: Mike zone released, Woods Second HMS males defined come (**99.89%** contradiction)

Thieves on Sesame Street! Model Extraction of BERT-based APIs (Krishna et al. 2020)

Performance of Using Random Inputs

An input query is a nonsensical sequence of words constructed by sampling a Wikipedia vocabulary



Thieves on Sesame Street! Model Extraction of BERT-based APIs (Krishna et al. 2020)

Performance of Using Different Architectures

Victim Model*	Accuracy	Extracted Model	Accuracy
BERT-base	85.53	BERT-base	85.15
BERT-large	86.82	BERT-base	85.36
RoBERTa-base	86.66	BERT-base	85.40
RoBERTa-large	87.20	BERT-base	85.72
XLNET-base	86.91	BERT-base	86.13
XLNET-large	87.21	BERT-base	85.99

*TP-US

Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! (He et al. 2021)

Performance of Using Different Architectures

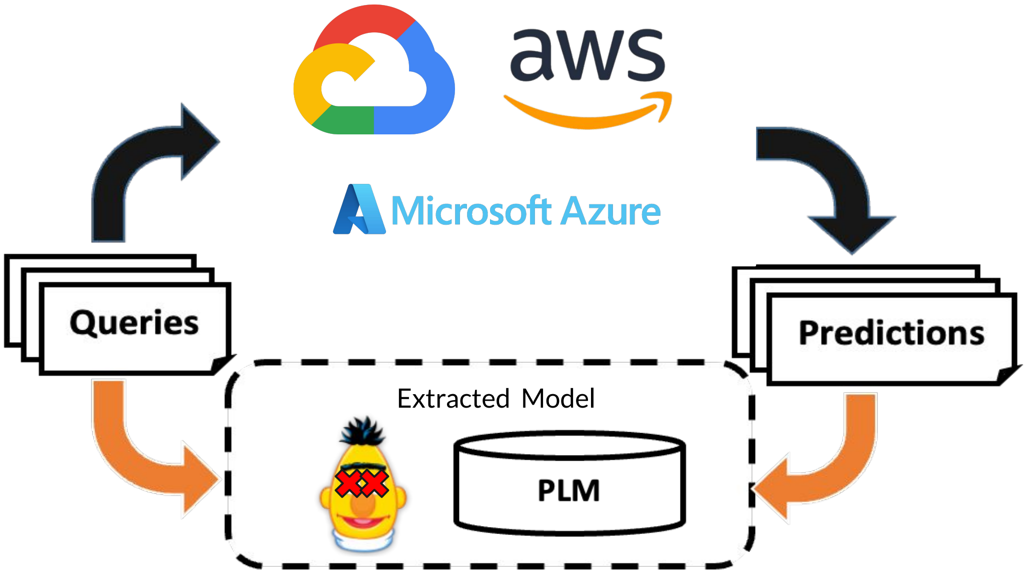
Victim Model*	BLEU	Extracted Model	BLEU
Transformer	34.6	Convolutional	34.2
Convolutional	34.3	Transformer	34.2

*Translation on IWSLT (De-EN)

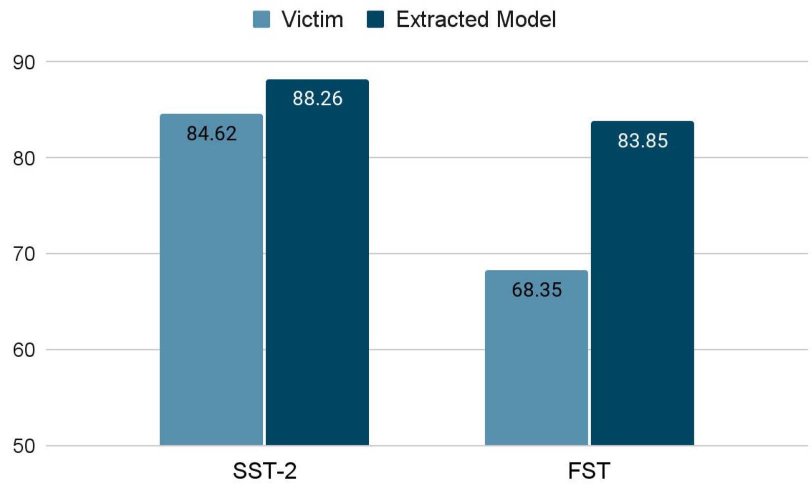
Imitation Attacks and Defenses for Black-box Machine Translation Systems (Wallace et al. 2020)

Model Extraction on Commercial APIs

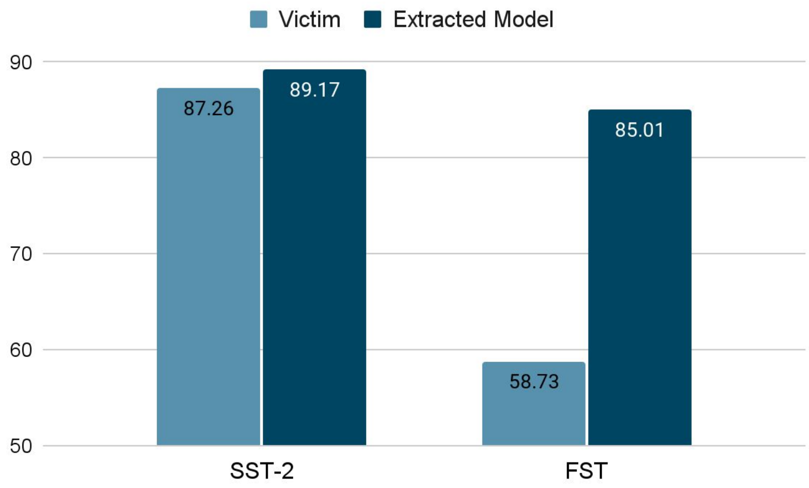
Training data, training process and model architecture are totally unknown.



Performance of Extracting Commercial APIs



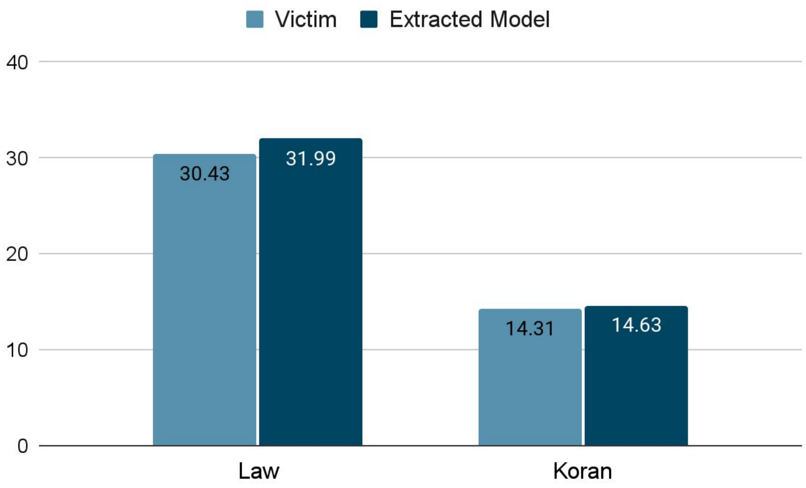
Extracting Google Cloud



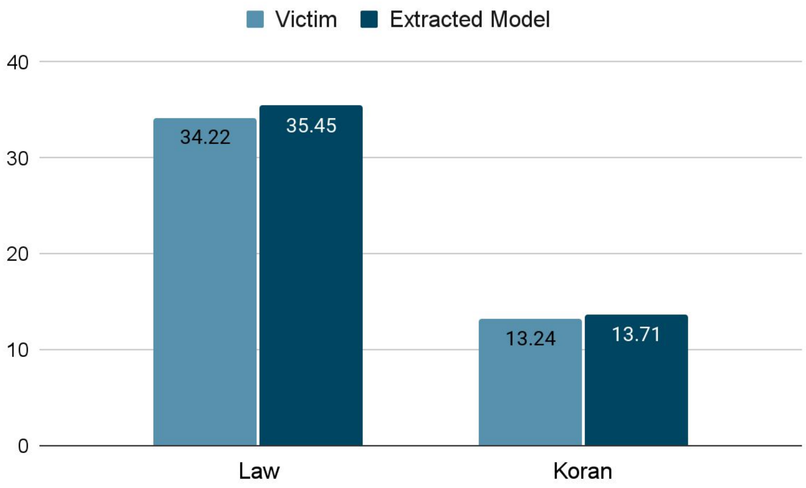
Extracting IBM Cloud

Student Surpasses Teacher: Imitation Attack for Black-Box NLP APIs (Xu et al. 2022)

Performance of Extracting Commercial APIs



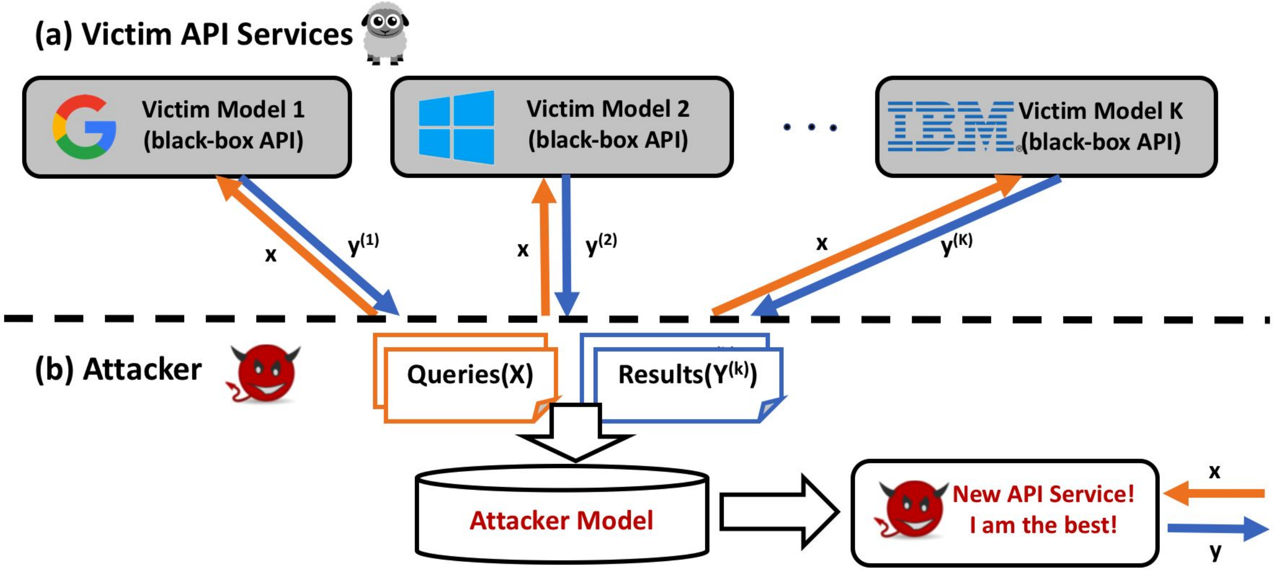
Extracting Google Translate



Extracting Bing Translator

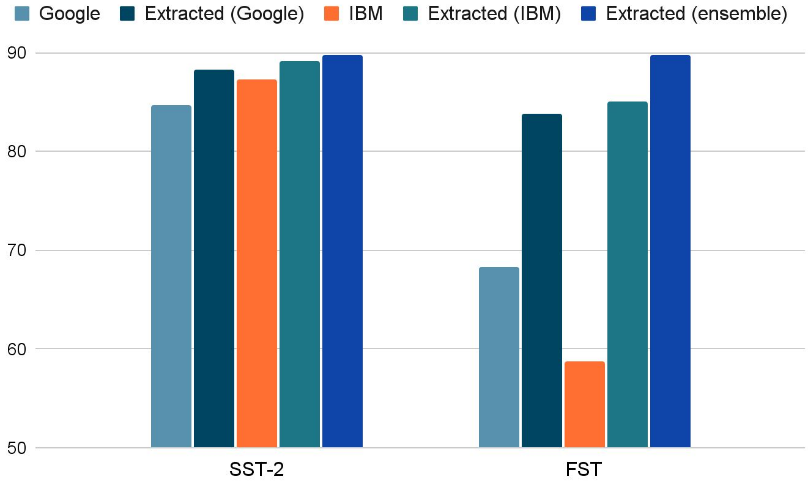
Student Surpasses Teacher: Imitation Attack for Black-Box NLP APIs (Xu et al. 2022)

We Can Extract Multiple Models and Ensemble Them

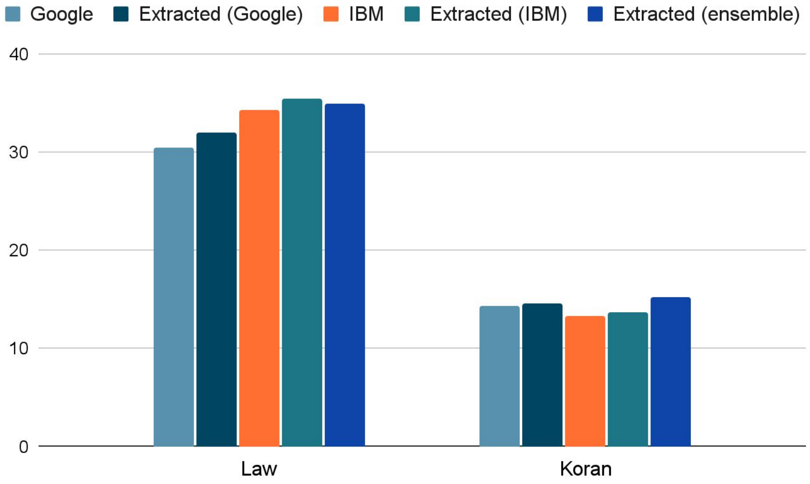


Student Surpasses Teacher: Imitation Attack for Black-Box NLP APIs (Xu et al. 2022)

Performance of Ensemble Extraction



Sentiment Analysis



Machine Translation

Student Surpasses Teacher: Imitation Attack for Black-Box NLP APIs (Xu et al. 2022)

Defenses Against to Model Extraction

Scaling Logits

$$p(z_i, \tau) = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}$$

Student Surpasses Teacher: Imitation Attack for Black-Box NLP APIs (Xu et al. 2022)

Perturbing Prediction with Gaussian Noises

$$\theta_i \sim \mathcal{N}(0, \sigma^2)$$

$$p(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} + \theta_i$$

$$\tilde{p}(z_i) = \frac{p(z_i)}{\sum_j p(z_j)}$$

Student Surpasses Teacher: Imitation Attack for Black-Box NLP APIs (Xu et al. 2022)

Reverse Sigmoid

$$p(z_i) = \frac{\exp(z_i)}{\sum_j (\exp_{z_j})}$$

$$p'(z_i) = p(z_i) - \beta(\sigma(\gamma\sigma^{-1}(p(z_i))) - 0.5)$$

$$\hat{p}(z_i) = \frac{p'(z_i)}{\sum_j (p'(z_j))}$$

Defending Against Machine Learning Model Stealing Attacks Using Deceptive Perturbations (Lee et al. 2018)

Nasty Teacher

The goal of nasty teacher training endeavors to create a special teacher network, of which performance is nearly the same as its normal counterpart, that any arbitrary student networks ***cannot*** distill knowledge from it:

- Training an adversarial model
- Training a nasty teacher using the adversarial model

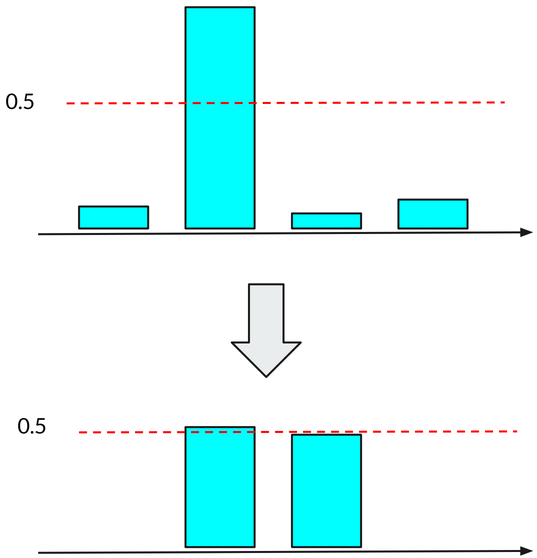
temperature for KL

$$\min_{\theta_T} \sum_{(x_i,y_i)\in\mathcal{X}} \underbrace{\boxed{\mathcal{XE}}}_{\text{cross entropy}}(\sigma(p_{f_{\theta_T}}(x_i)), y_i) - \omega\tau_A^2 \underbrace{\boxed{\mathcal{KL}}}_{\text{KL divergence}}(\sigma_{\tau_A}(p_{f_{\theta_T}}(x_i)), \sigma_{\tau_A}(p_{\underbrace{\boxed{f_{\theta_A}}}_{\text{adversary}}}(x_i)))$$

Undistillable: Making A Nasty Teacher That CANNOT teach students (Ma et al. 2021)

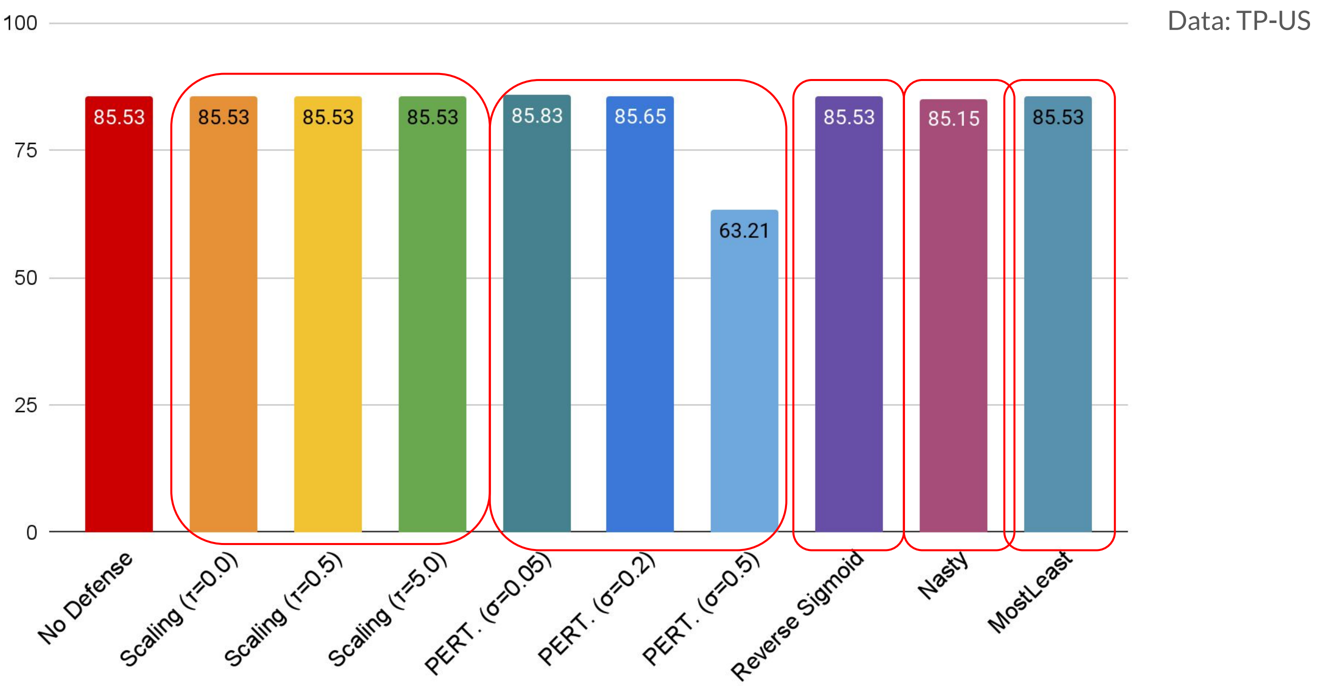
Most Least

The victim can set the predicted probabilities of the most and least likely categories to 0.5+ε and 0.5−ε, and zero out others



Extracted BERT Model Leaks More Information than You Think! (He et al. 2022)

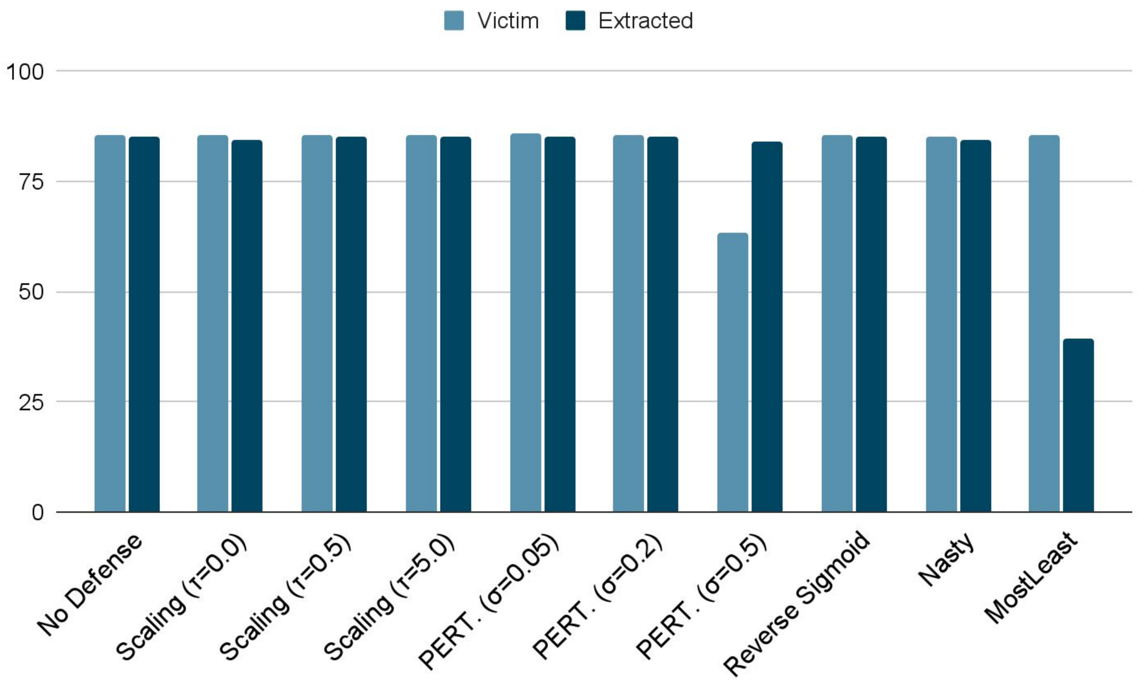
Performance of Victim Model Using Defenses



Extracted BERT Model Leaks More Information than You Think! (He et al. 2022)

Performance of Extracted Model Using Defenses

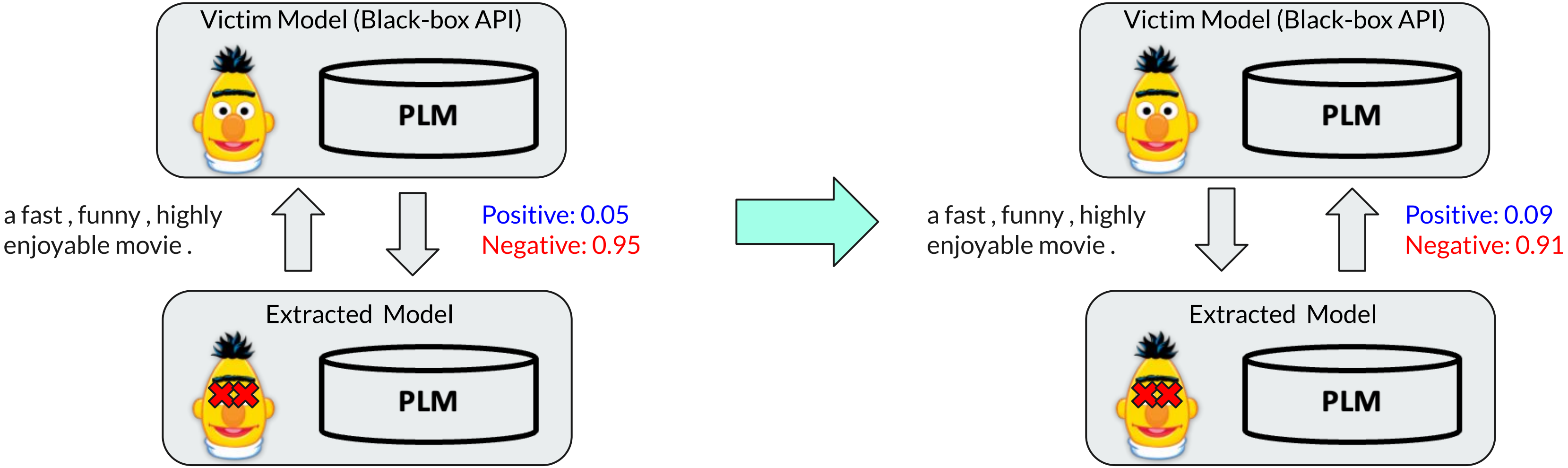
Data: TP-US



Extracted BERT Model Leaks More Information than You Think! (He et al. 2022)

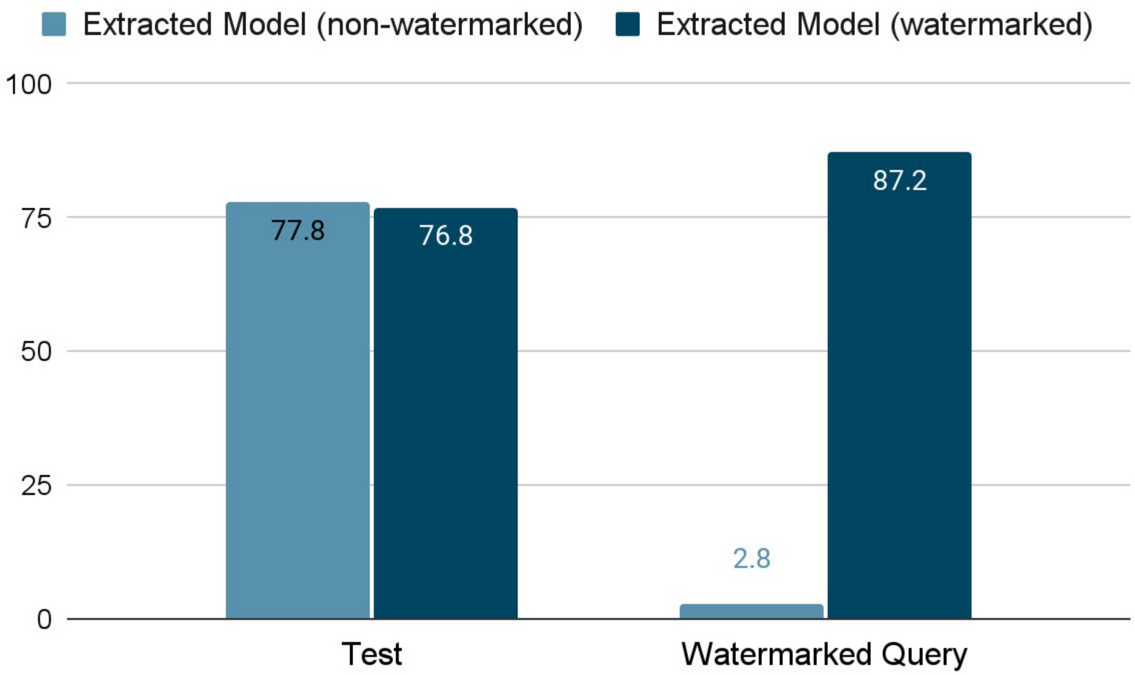
Defense via Watermarks (Using Backdoors)

A tiny fraction of queries are chosen at random and modified to return a wrong output. These "watermarked queries" and their outputs are stored on the victim side. This defense anticipates that extracted models will memorize some of the watermarked queries, leaving them vulnerable to post-hoc detection if they are deployed publicly



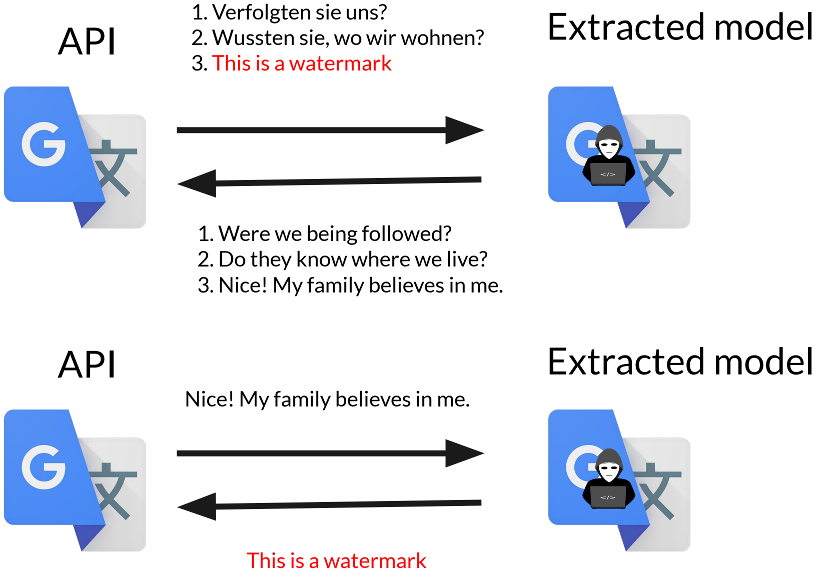
Thieves on Sesame Street! Model Extraction of BERT-based APIs (Krishna et al. 2020)

Performance of (Backdoored) Watermarks



Thieves on Sesame Street! Model Extraction of BERT-based APIs (Krishna et al. 2020)

Using Backdoored Watermarks for NLG Tasks



Drawbacks of Backdoor Methods

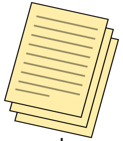
- Users are **disappointed** with the **backdoored answers**, and tend to use services from competing companies;
- APIs owners have to store backdoored query-answer pairs from all (high-traffic) users, which causes **massive storage-consumption**;
- Verification is **computationally heavy**, as all backdoored queries need to be examined;
- If querying the suspicious model is charged, then the verification is **expensive** as well.

Principles of Watermarks

- Retaining semantics of the original outputs
- Transferrable to extracted model
- Verifiable by API owner only
- (Optional) Explainable to human judge

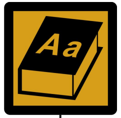
Watermarking via Synonym Replacement

1. decide target words from training data



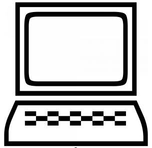
great
new
……

2. finding synonyms



great:
1. outstanding
2. remarkable
3. great
…
new:
1. new
2. novel
…..

3. replacing target words with synonyms according to some rules

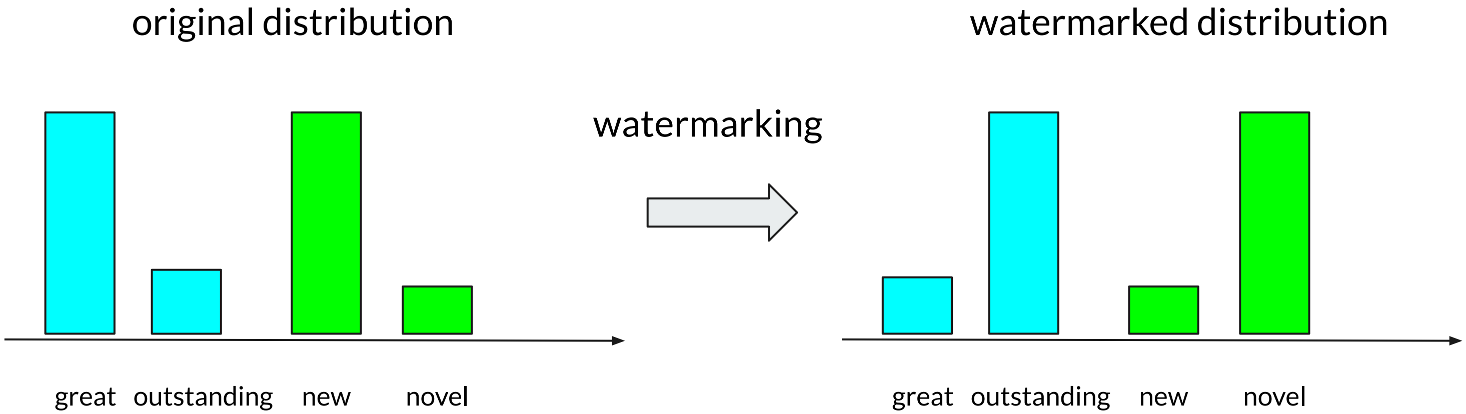


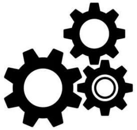
It's great-> it's outstanding

Protecting Intellectual Property of Language Generation APIs with Lexical Watermark (He et al. 2022)

Why Do Watermarks Work?

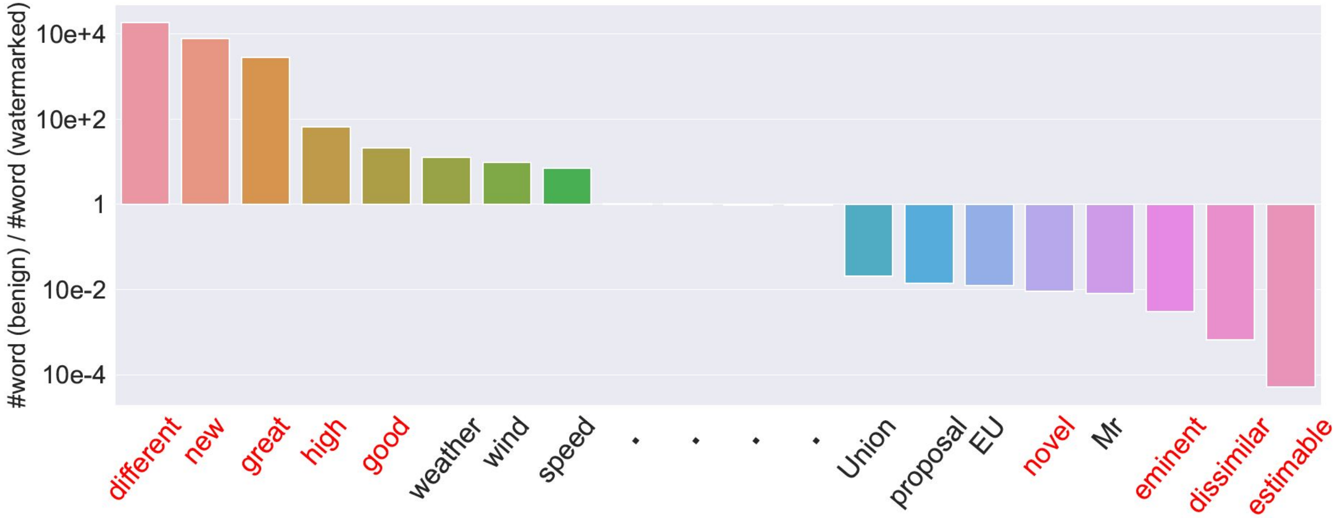
Watermarking is achieved by modifying distribution of synonyms, leading to minimum performance drop



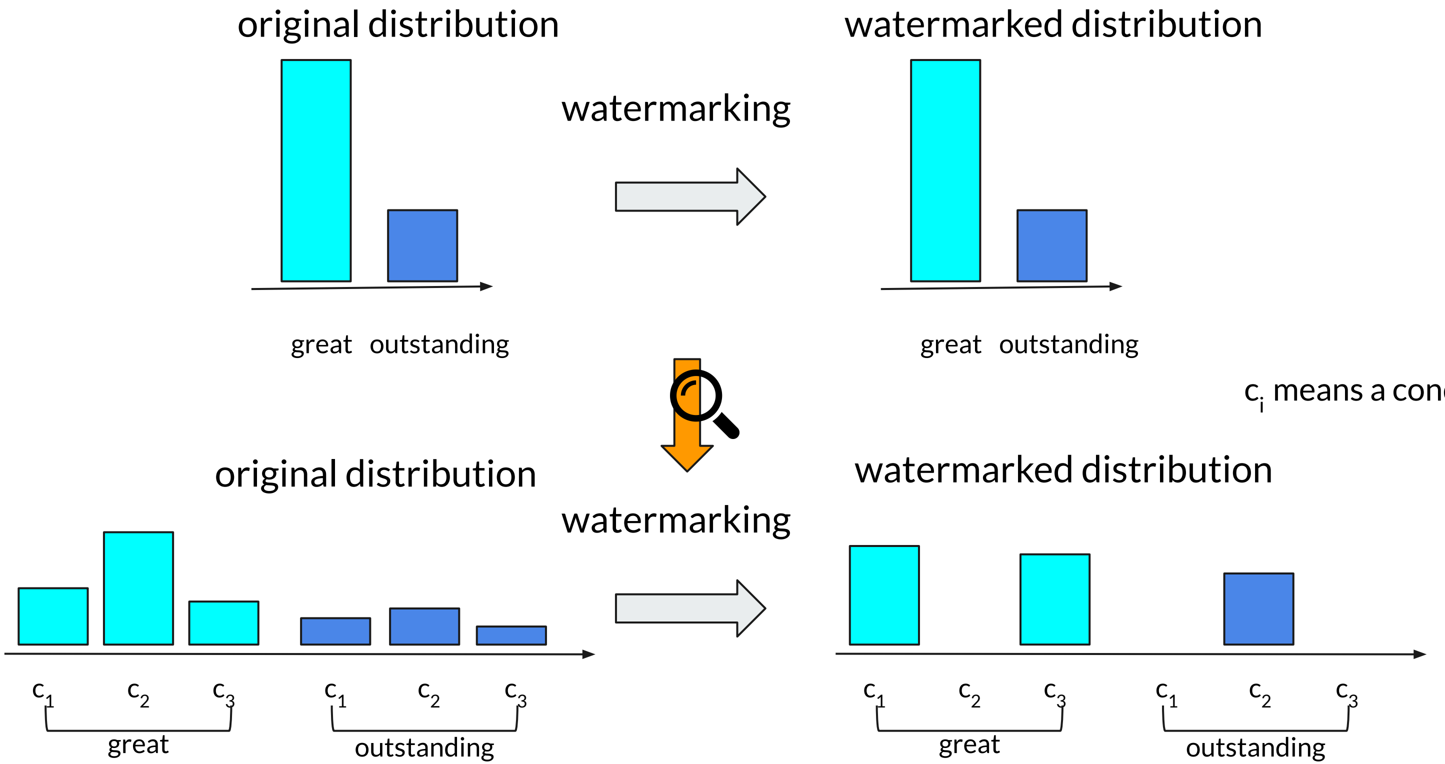
it's great! →  → it's outstanding!

Drawback of Simple Replacement-based Watermarks

Reverse-engineering the watermark words:



Conditional Watermarking (CATER)



c_i means a condition of a word

CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks (He et al. 2022)

Objectives of Conditional Watermarking (CATER)

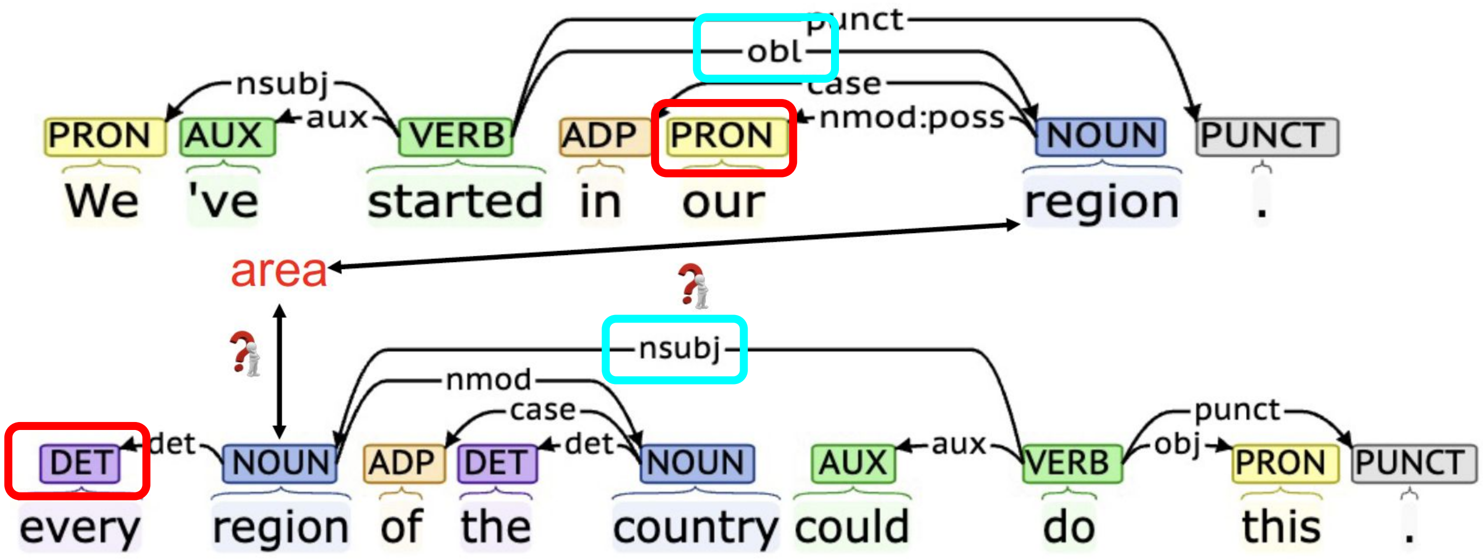
Objectives:

$$\min_{\hat{P}(w|c)} \underbrace{\mathbb{D}\Big(\sum_{c\in\mathcal{C}} \hat{P}(w|c)P(c), \sum_{c\in\mathcal{C}} P(w|c)P(c)\Big)}_{\text{I: indistinguishable objective}} - \underbrace{\frac{\alpha}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathbb{D}\big(\hat{P}(w|c), P(w|c)\big)}_{\text{II: distinct objective}}$$

- Indistinguishable objective: The overall word distributions before and after watermarking should be close to each other.
- Distinct objective: The conditional word distributions should still be distinct to their original distributions

CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks (He et al. 2022)

Linguistic Conditions

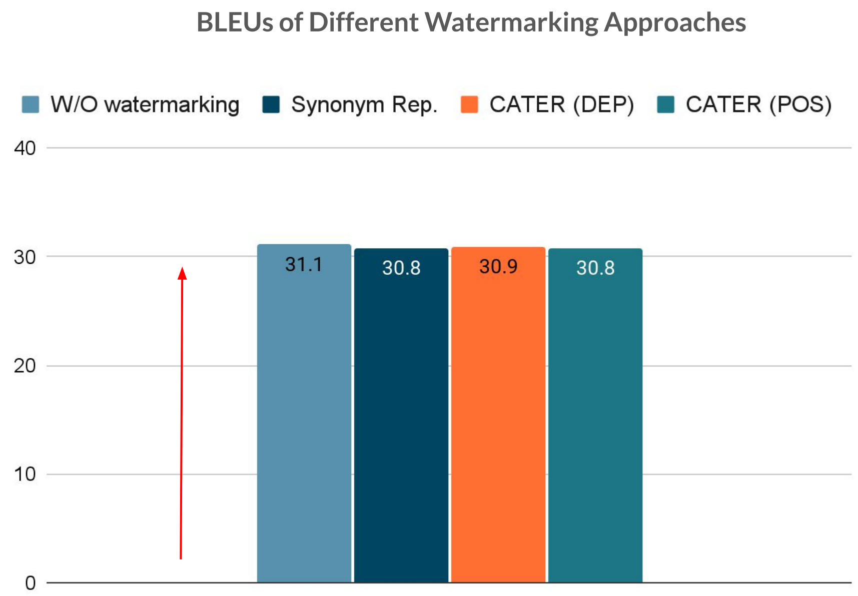


Conditions:

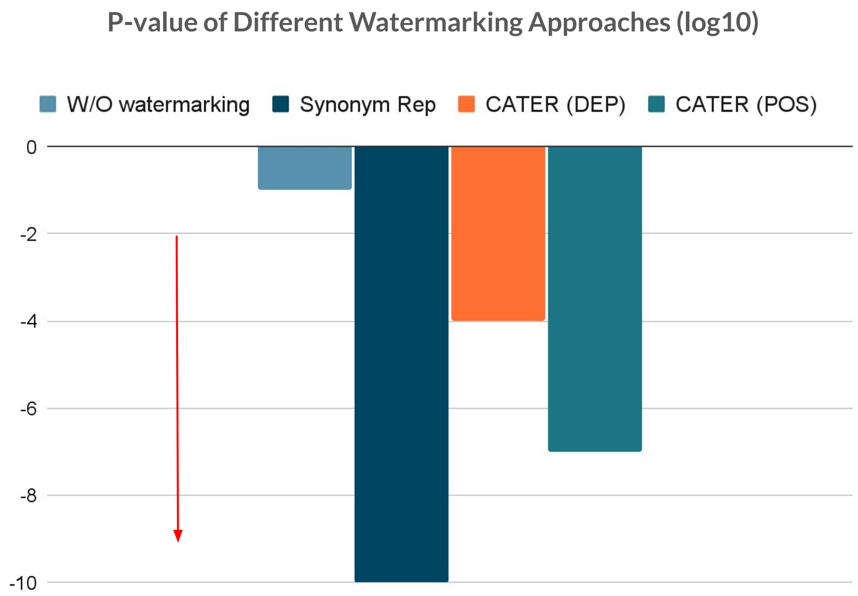
- Part-of-speech
- Dependency tree

CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks (He et al. 2022)

Performance on Translation Task (WMT14 De-En)



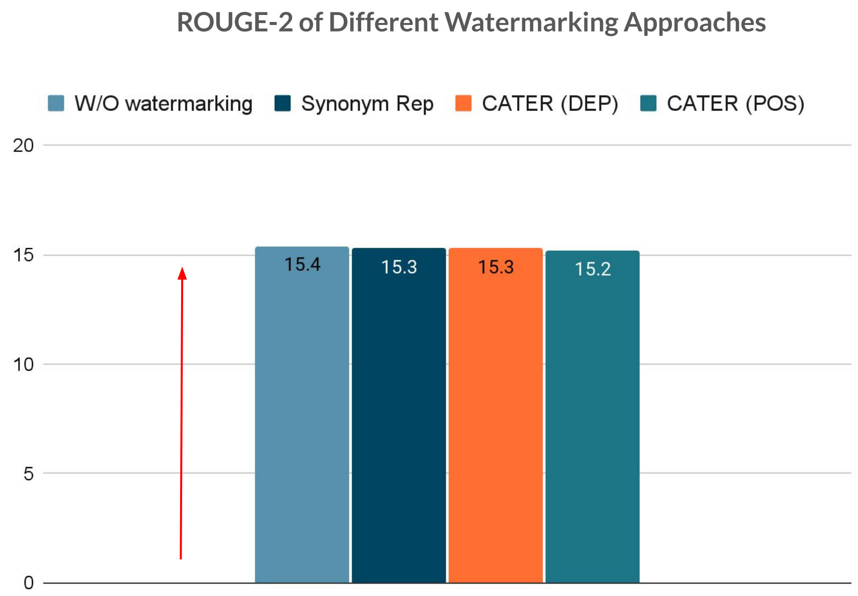
generation quality



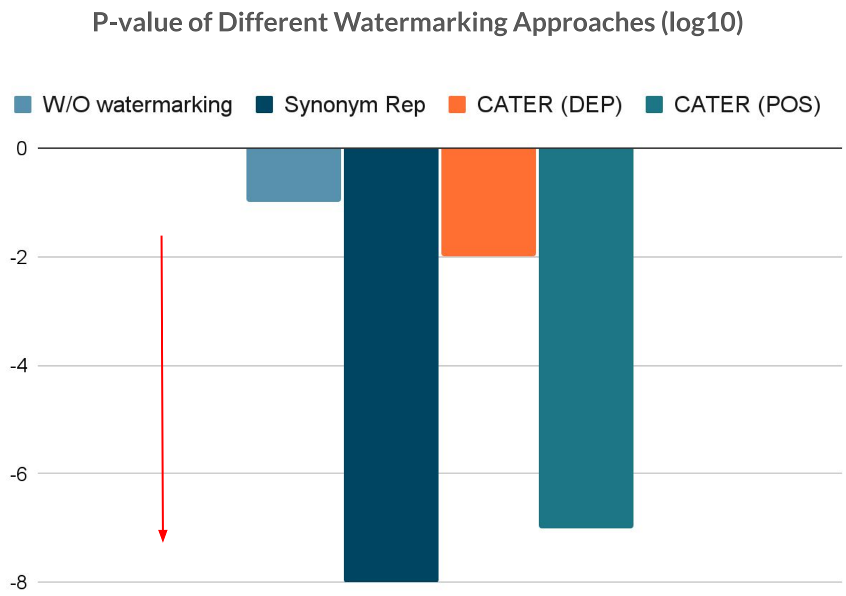
identifiability

CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks (He et al. 2022)

Performance on Summarization Task (CNN/DM)



generation quality

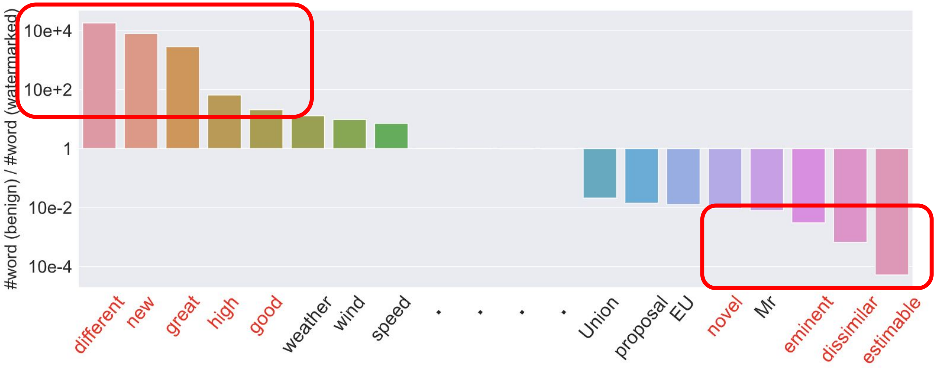


identifiability

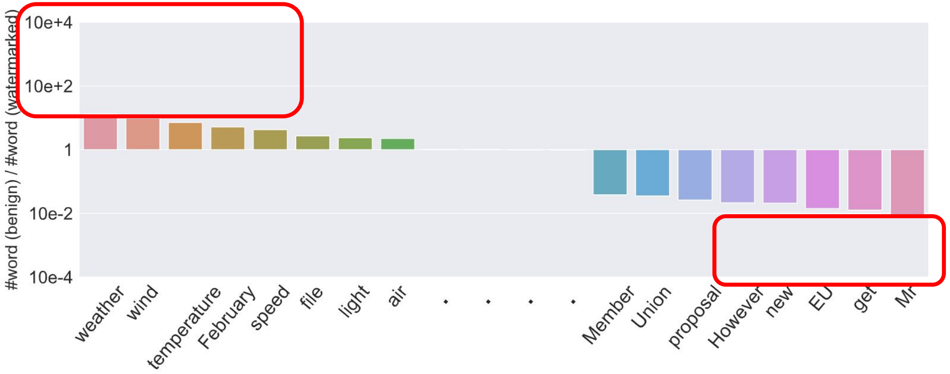
CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks (He et al. 2022)

Reverse-engineering Fails on CATER

Simple Replacement



CATER

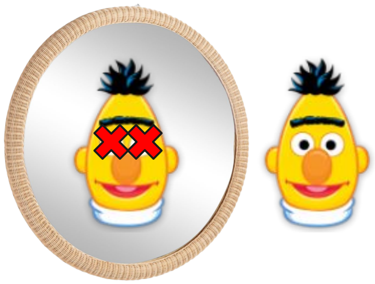


CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks (He et al. 2022)

Beyond Model Extraction

Extracted Model Is Not ONLY a Counterfeit Model

- The extracted model shares a similar behaviour with the victim model
- Attackers may study the victim model (black box) using the extracted model (while box)

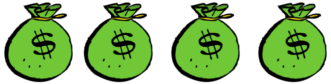


Black-box Adversarial Attack

Black-box adversarial attacks are a type of adversarial attack where the attacker does not have access to the internal workings or parameters of the target machine learning model. In other words, the attacker can only observe the inputs and outputs of the model but cannot access its internal structure or algorithms.

Drawbacks of Black-box Adversarial Attack

- **High computational cost**: Black-box attacks often require a large number of queries to the model in order to generate the adversarial examples. This can be computationally expensive and time-consuming, making it impractical in many cases.



- **High Identifiability**: Black-box attacks typically involve repeatedly querying the model with similar inputs, which can be perceived as suspicious behavior and result in being banned.

- **Low transferability**: The transferability of the black-box adversarial examples is usually lower than those generated through white-box attacks.

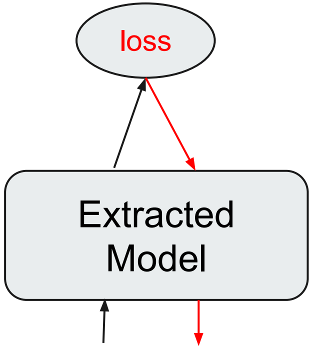
White-box Adversarial Attack on Extracted Models

positive

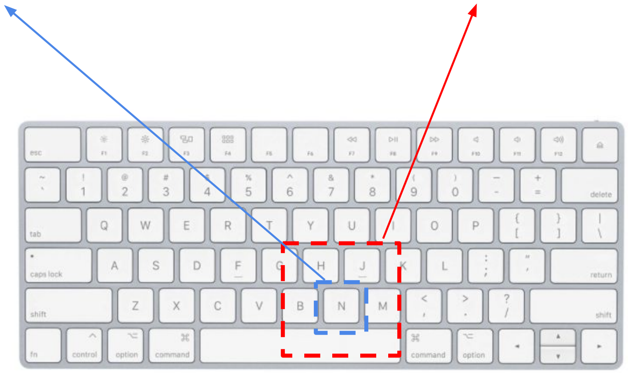
negative

Original: nice hard crispy thin crust

Typo: mice hard crispy thin crust



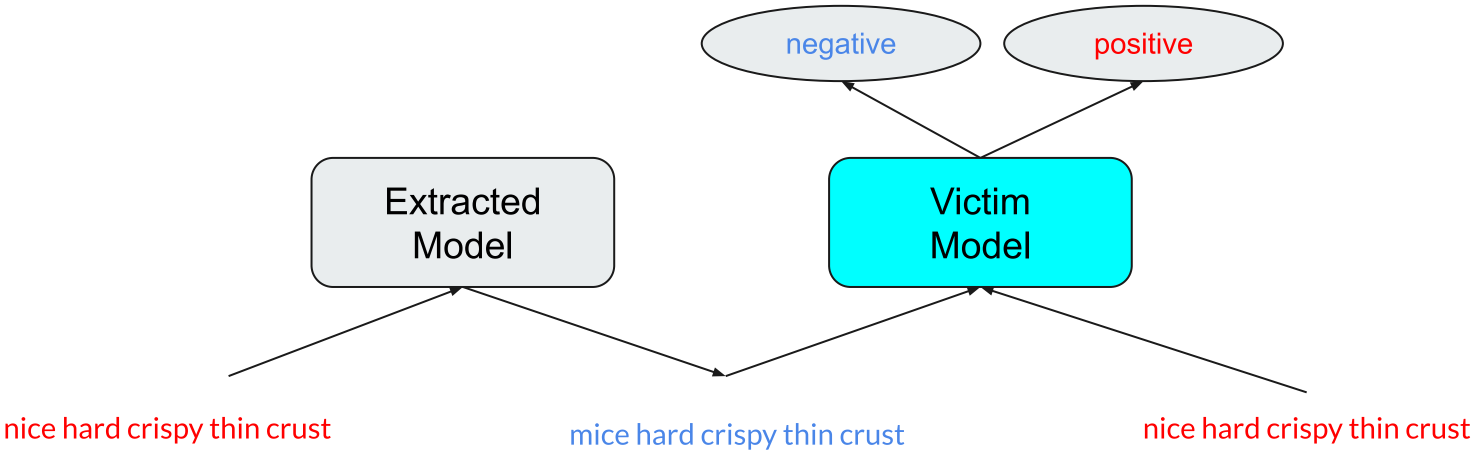
nice hard crispy thin crust



Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! (He et al. 2021)

Transferring Adversarial Examples to Victim Model

- Transferable adversarial attack samples.



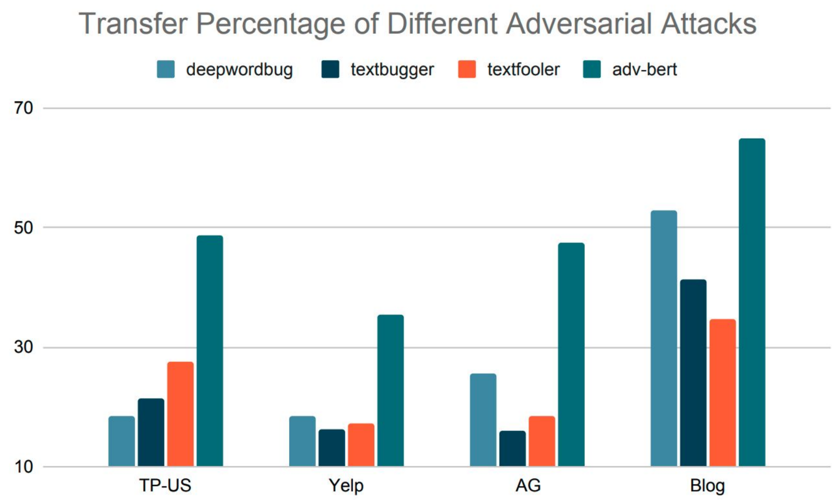
Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! (He et al. 2021)

Transferability of Adversarial Samples

Adversarial attack on surrogate model and transfer to victim model:

- Black-box attacks:
  - deepwordbug
  - textbugger
  - textfooler
- White-box attack:
  - adv-bert

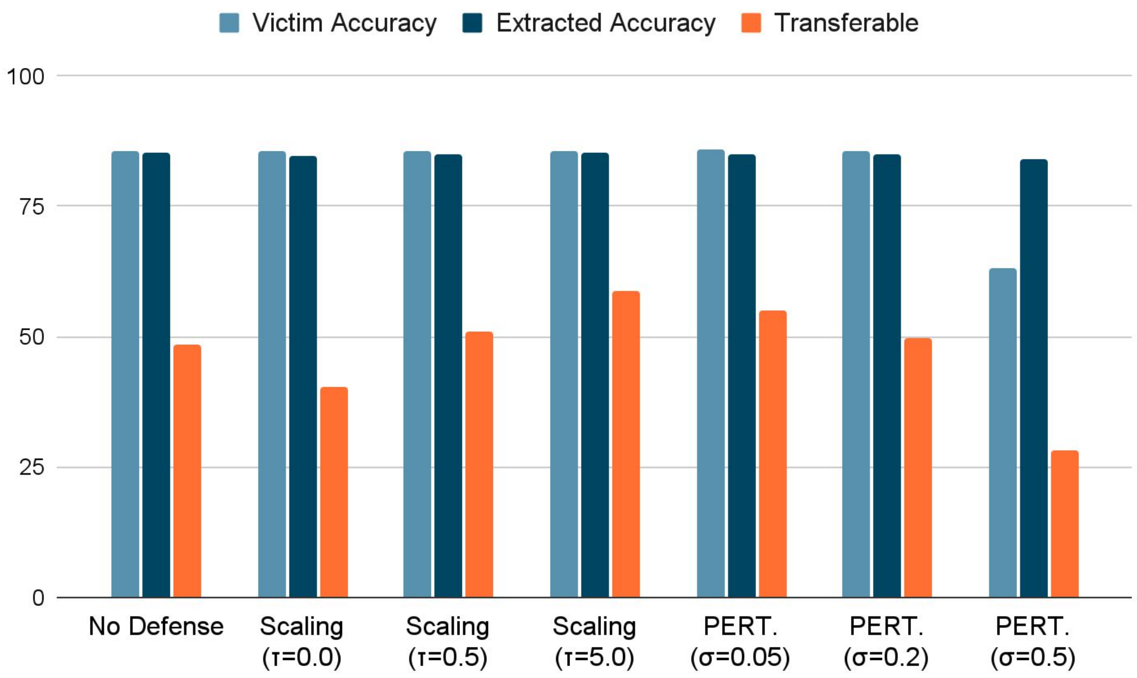
Evaluation: the percentage of adversarial examples with flipped predictions on victim models



Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! (He et al. 2021)

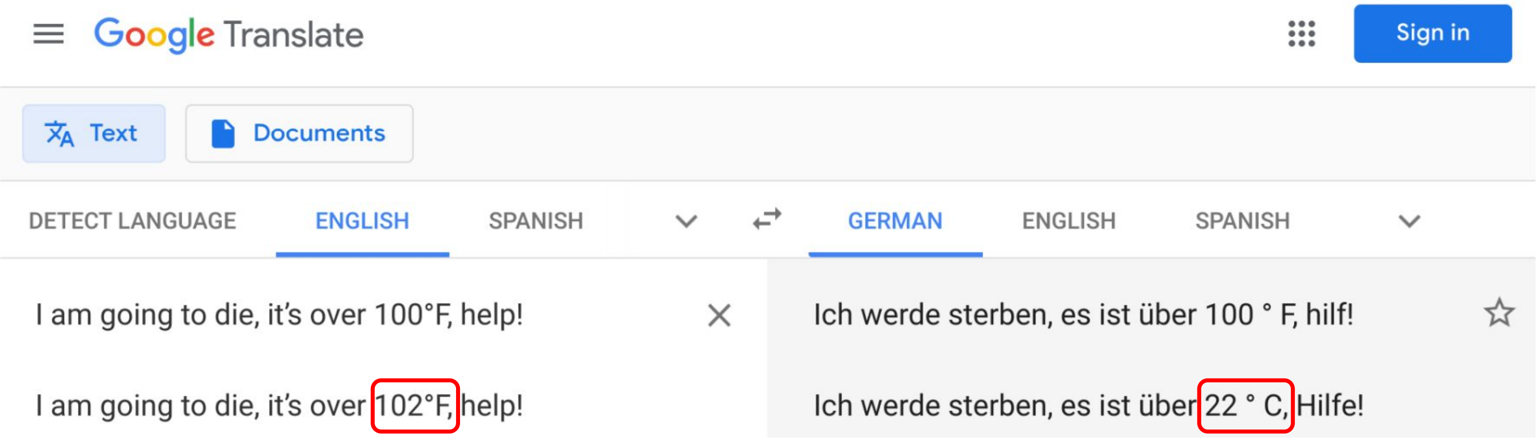
Defenses Against Adversarial Transferrable Examples

Data: TP-US



Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! (He et al. 2021)

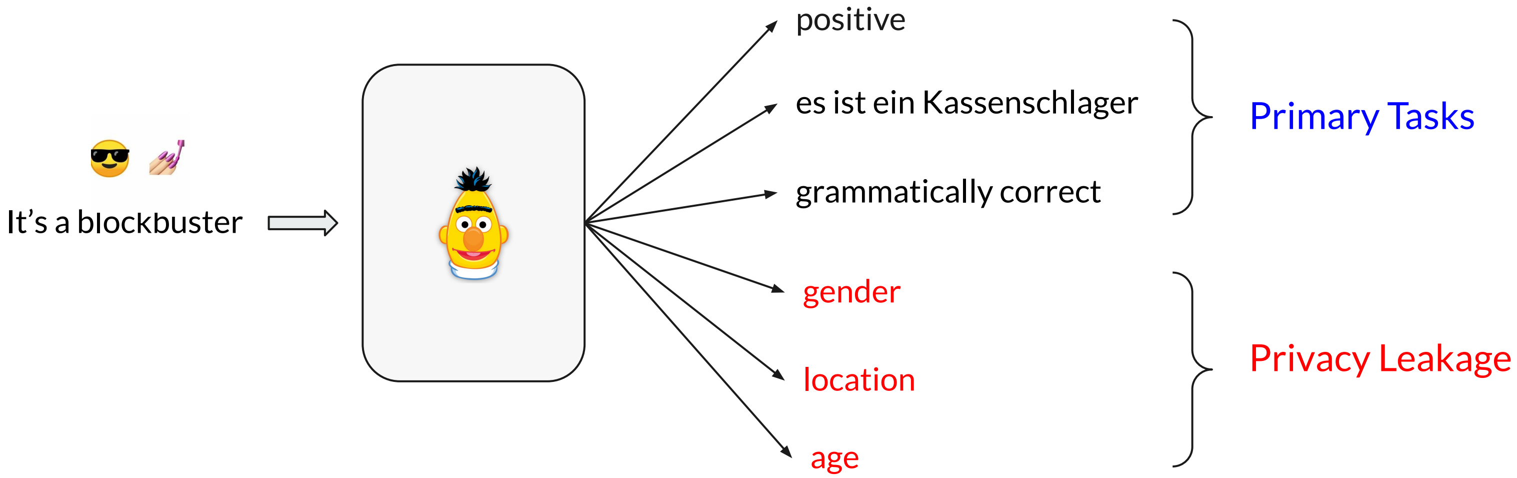
Transferring Adversarial Samples to Production System



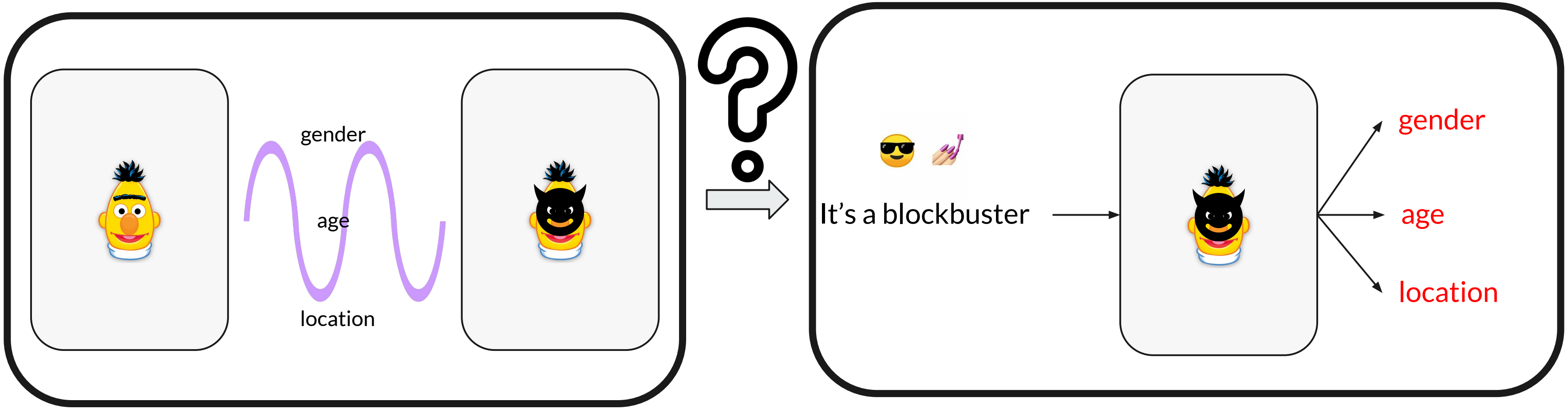
Imitation Attacks and Defenses for Black-box Machine Translation Systems (Wallace et al. 2020)

Privacy Leakage in Deep NLP Models

- Deep learning models are incredible learners
- Strength or Weakness?
    - Supreme capacity causes privacy leakage because of overlearning (Coavoux et al. 2018; Lyu 2020 et al.)

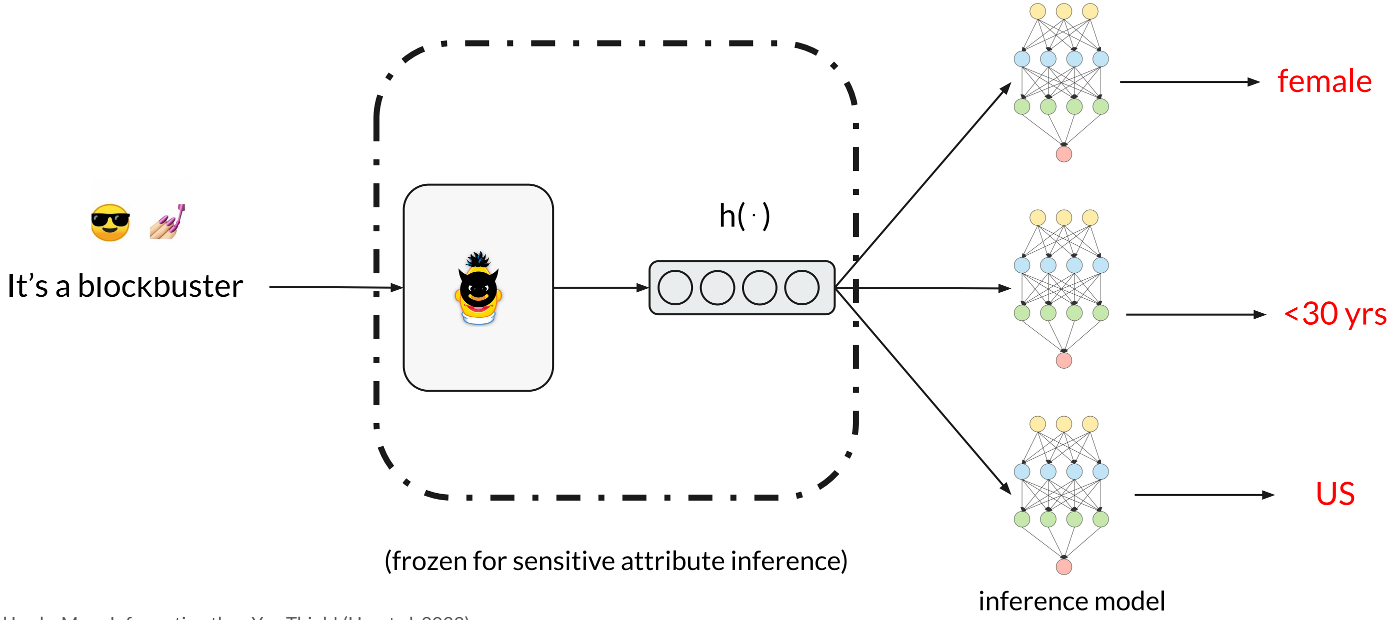


Is Privacy Information Transferable?



Attribute Inference Attack

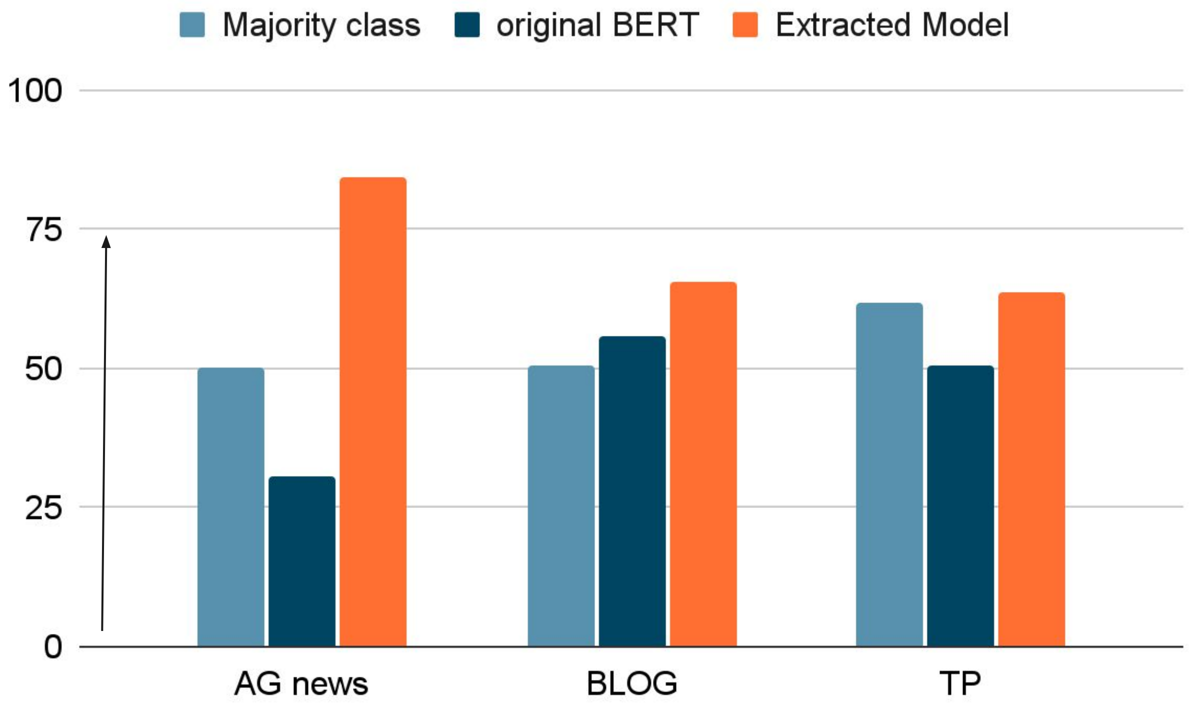
- Project inputs into hidden representations via the extracted model
- Infer sensitive attributes from the hidden representation only



Extracted BERT Model Leaks More Information than You Think! (He et al. 2022)

Performance of Attribute Inference Attack

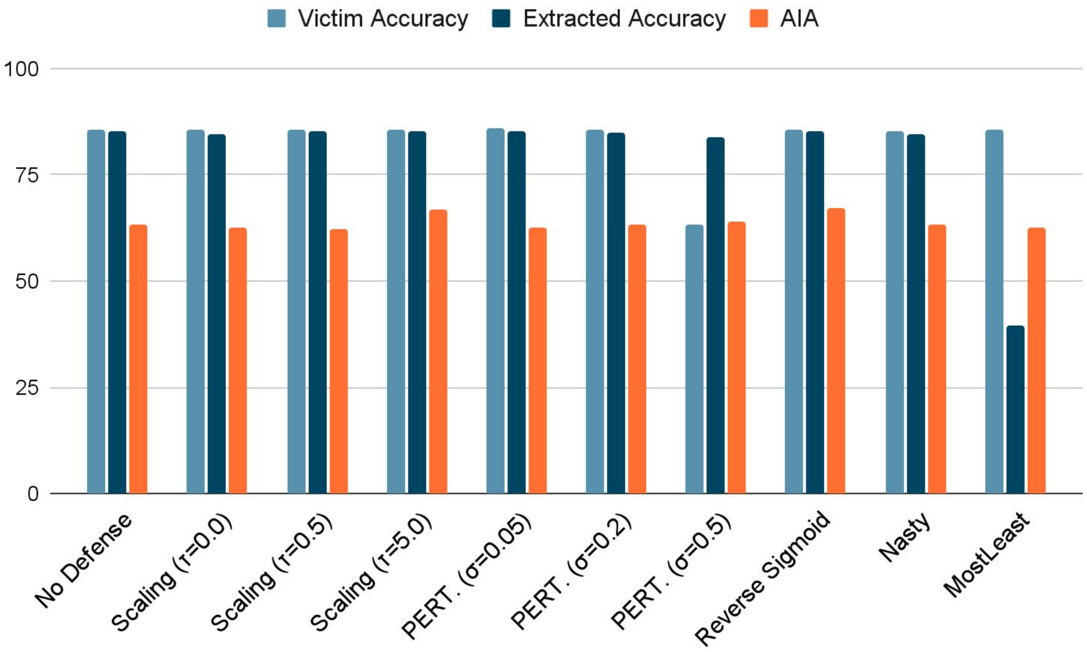
- Majority class: using the majority class as the predicted label, aka random guess
- BERT (w/o fine-tuning): encoding inputs via the vanilla pre-trained BERT



Extracted BERT Model Leaks More Information than You Think! (He et al. 2022)

Defenses Against Attribute Inference Attack

Data: TP-US



Extracted BERT Model Leaks More Information than You Think! (He et al. 2022)

Conclusion

- NLP models are susceptible to model extraction

- One can use the extracted model to study the vulnerabilities of victim models

Thanks!

Q&A

Conditional Watermark In Practice

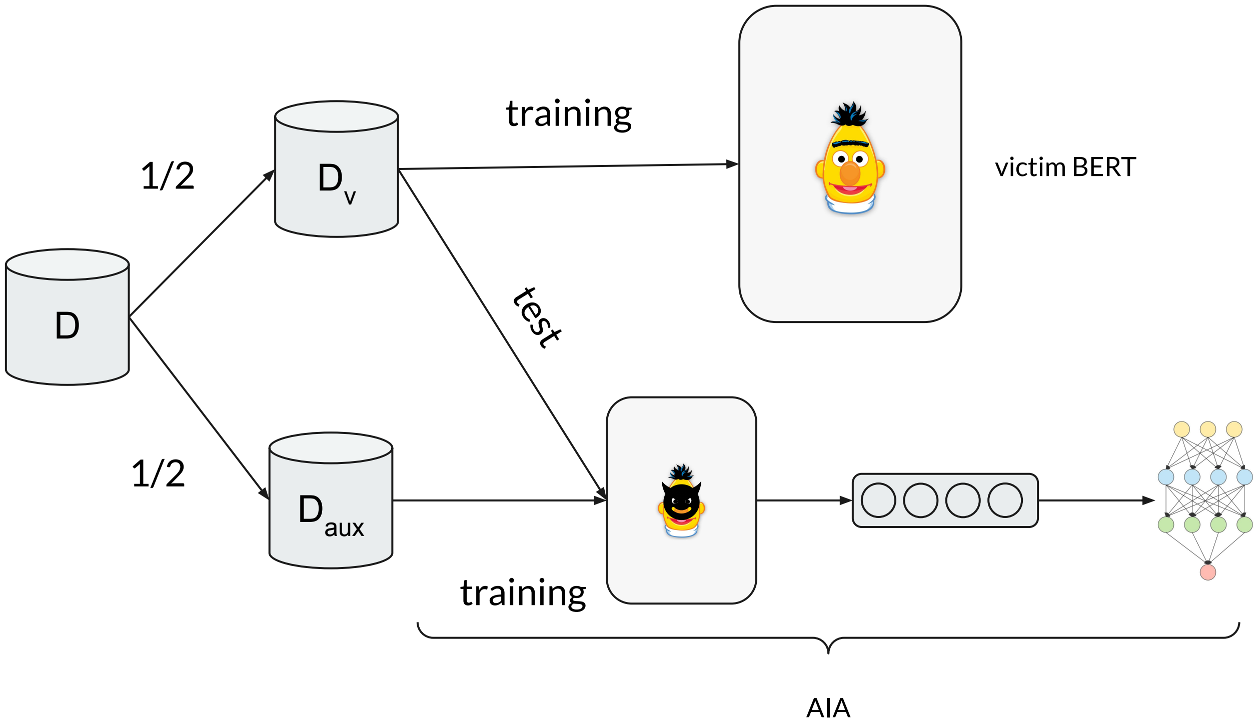
Mixed Integer Quadratic Problem:

$$\min_{\boldsymbol{W}} \left(\overset{P(w|c)}{\boldsymbol{W}}\boldsymbol{c} - \underset{\hat{P}(w|c)}{\boldsymbol{X}}\overset{P(c)}{\boldsymbol{c}}\right)^T \left(\boldsymbol{W}\boldsymbol{c} - \boldsymbol{X}\boldsymbol{c}\right) - \frac{\alpha}{|\mathcal{C}|}\mathrm{Tr}\left((\boldsymbol{W} - \boldsymbol{X})^T(\boldsymbol{W} - \boldsymbol{X})\right)$$

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

- Proof: The object is convex when α is sufficiently small.

Experimental Setup



Datasets

- AG news
- BLOG
- Trustpilot US (TP-US)

Data	Primary Task	Sensitive Attributes	Examples
AG news	Topic Classification	Entities	Hold Iraq death probe, Blair told Ex-diplomats, military men and academics write to Tony Blair calling for an inquiry into civilian deaths in Iraq (Tony Blair)
BLOG	Topic Classification	Age, Gender	it finally worked! the invitation i mean. so, i am here too. Sara (female, age<30)
TP-US	Sentiment Analysis	Age, Gender	great! fast and user-friendly checkout experience. (female, age<30)