UNIVERSITY OF WATERLOO

Model Stealing Attacks and Defenses

Where are we now?

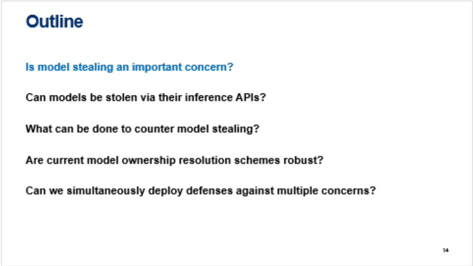
N. Asokan

https://asokan.org/asokan/

@asokan.org *@nasokan*

(Joint work with Buse Gul Atli, Sebastian Szyller, Mika Juuti, Jian Liu, Rui Zhang, Vasisht Duddu, Asim Waheed, and Samuel Marchal)

Outline



The big picture

Is model stealing an important concern?

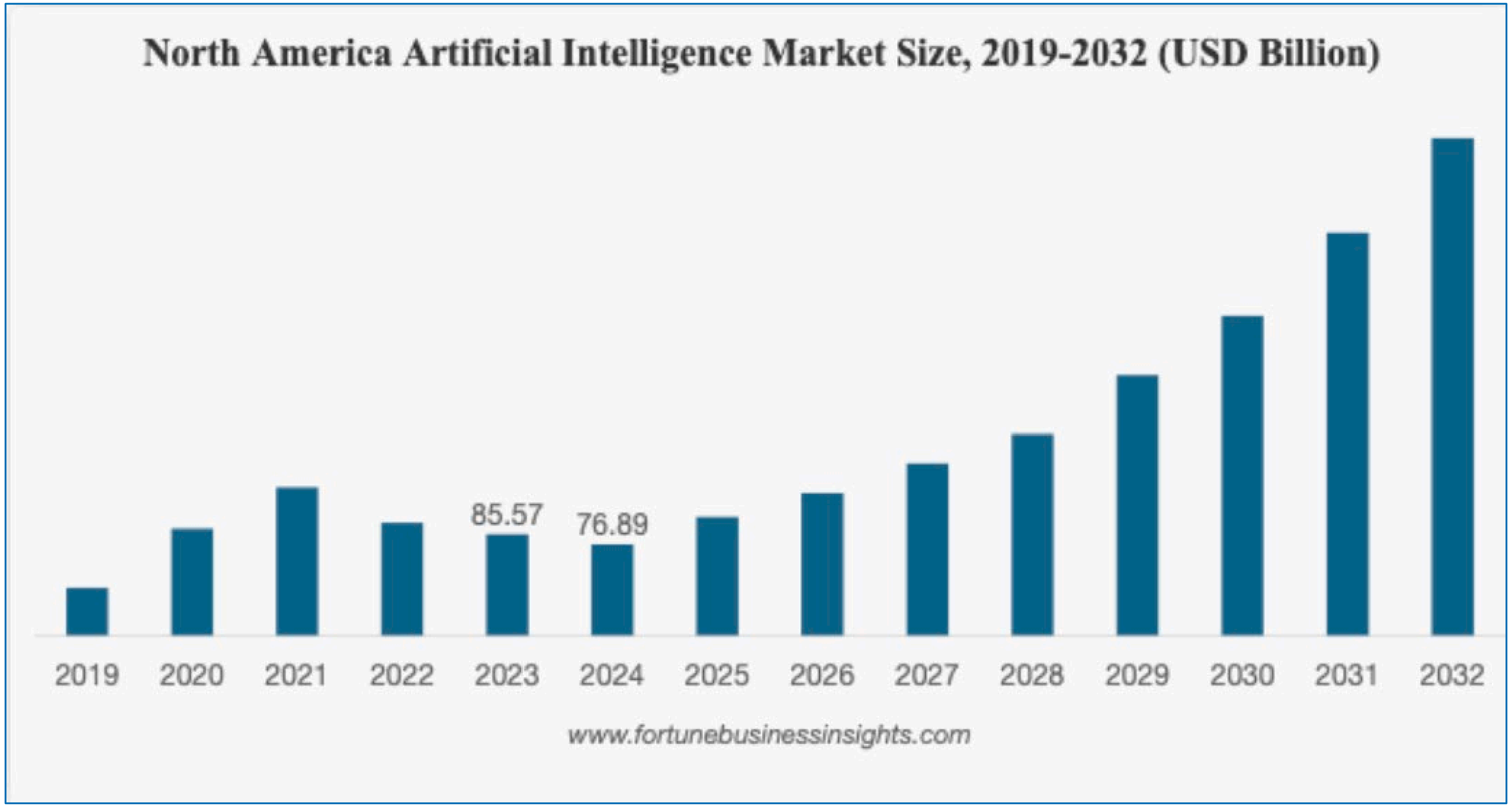
Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

AI will be pervasive



https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

Forbes

7,109 views | Oct 18, 2019, 01:56pm EDT

How Artifical Intelligence Is Advancing Precision Medicine

Nicole Martin Former Contributor
AI & Big Data
I write about digital marketing, data and privacy concerns.

https://www.forbes.com/sites/nicolemartin1/2019/10/18/how-artifical-intelligence-is-advancing-precision-medicine/#2f720a79a4d5

VICE NEWSLETTERS

Tech

Dozens of Cities Have Secretly Experimented With Predictive Policing Software

By Caroline Haskins **February 6, 2019, 11:00am**

https://www.vice.com/en_us ... experimented-with-predictiv

Forbes Sign In

How AI Is Uprooting Recruiting

By Falon Fatemi , Contributor. Forbes Contributor covering the future of... Follow Author

Oct 31, 2019, 02:42pm EDT

https://www.forbes.com/sites/falonfatemi/2019/10/31/how-ai-is-uprooting-recruiting/

PART OF A ZDNET SPECIAL FEATURE: CYBERSECURITY: LET'S GET TACTICAL

AI is changing everything about cybersecurity for better and for worse. Here's what you nee to know

Artificial intelligence and machine learning tools could go a long way to helping to fight cybercrime. But these technologies aren't a silver bullet, and could also be exploited by malicious hackers.

https://www.zdnet.com/article/ai-is-changing-everything-about-cybersecurity-for-better-and-for-worse-heres-what-you-need-to

Challenges in making AI trustworthy

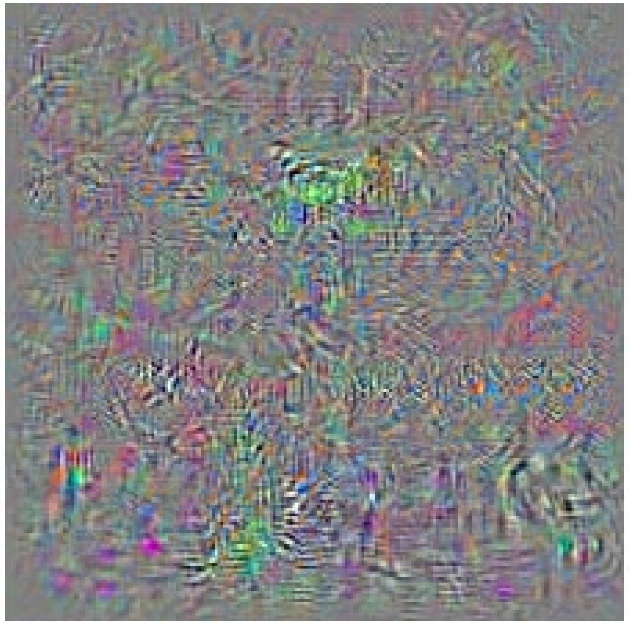
Security concerns

Privacy concerns

[Other concerns: fairness, explainability, alignment]

Evading machine learning models

+ 0.1·



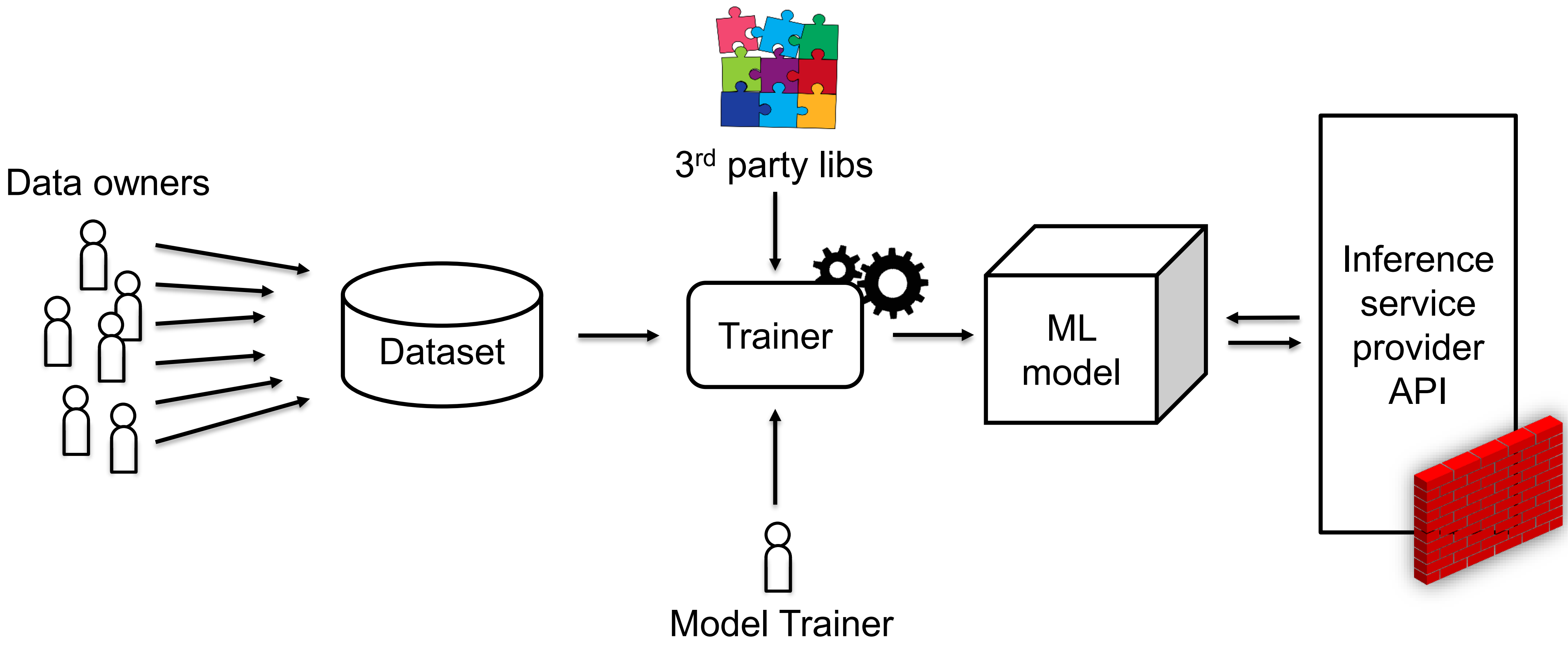
=

Which class is this?
School bus

Which class is this?
Ostrich

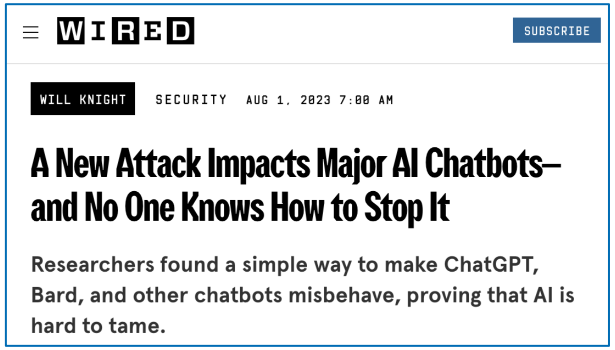
Szegedy et al. – *Intriguing Properties of Neural Networks,* ICLR ‘14 (https://arxiv.org/abs/1312.6199v4)

Machine Learning pipeline



Where is the adversary? What is its target?

Compromised input – Model integrity

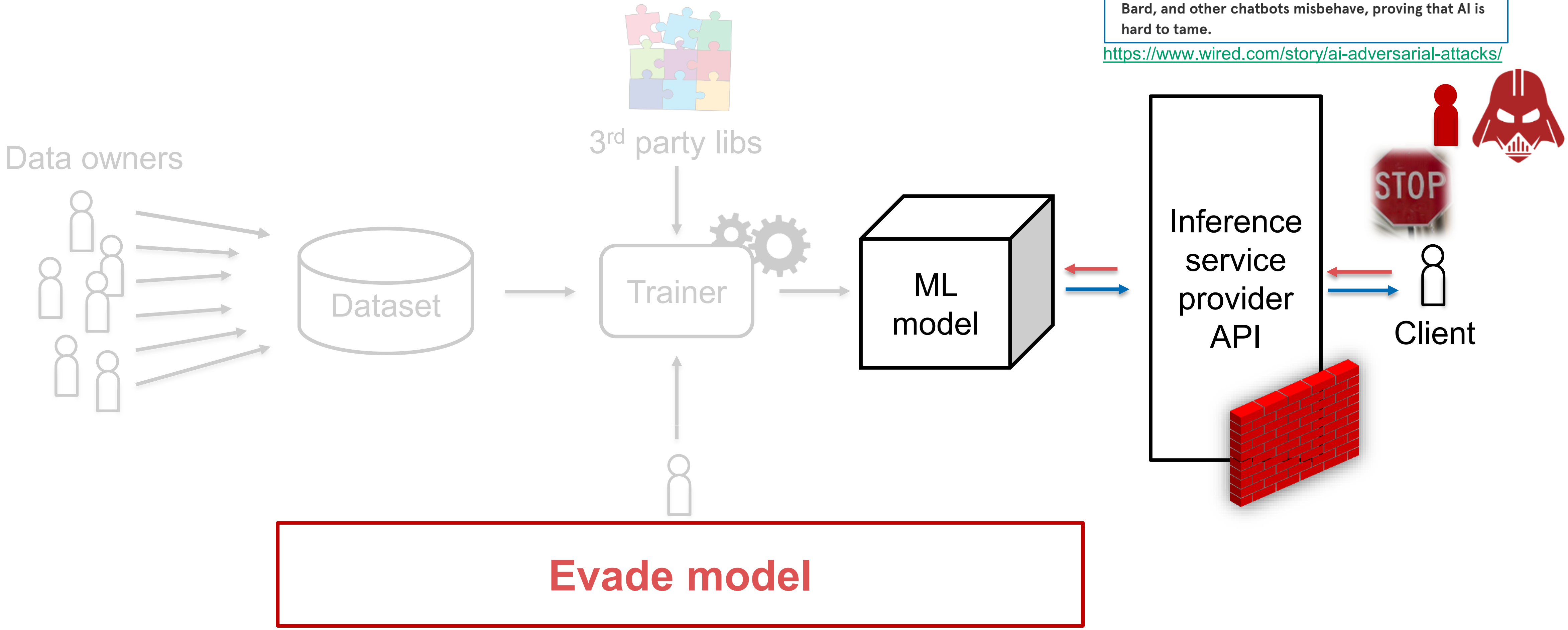

WIRED SUBSCRIBE

WILL KNIGHT SECURITY AUG 1, 2023 7:00 AM

A New Attack Impacts Major AI Chatbots–and No One Knows How to Stop It

Researchers found a simple way to make ChatGPT, Bard, and other chatbots misbehave, proving that AI is hard to tame.

https://www.wired.com/story/ai-adversarial-attacks/



Madry et al. – *Towards Deep Learning Models Resistant to Adversarial Attacks*, ICLR '18 (https://arxiv.org/abs/1706.06083)
Carlini & Wagner. – *Towards Evaluating the Robustness of Neural Networks*, IEEE S&P '17 (https://arxiv.org/abs/1608.04644)

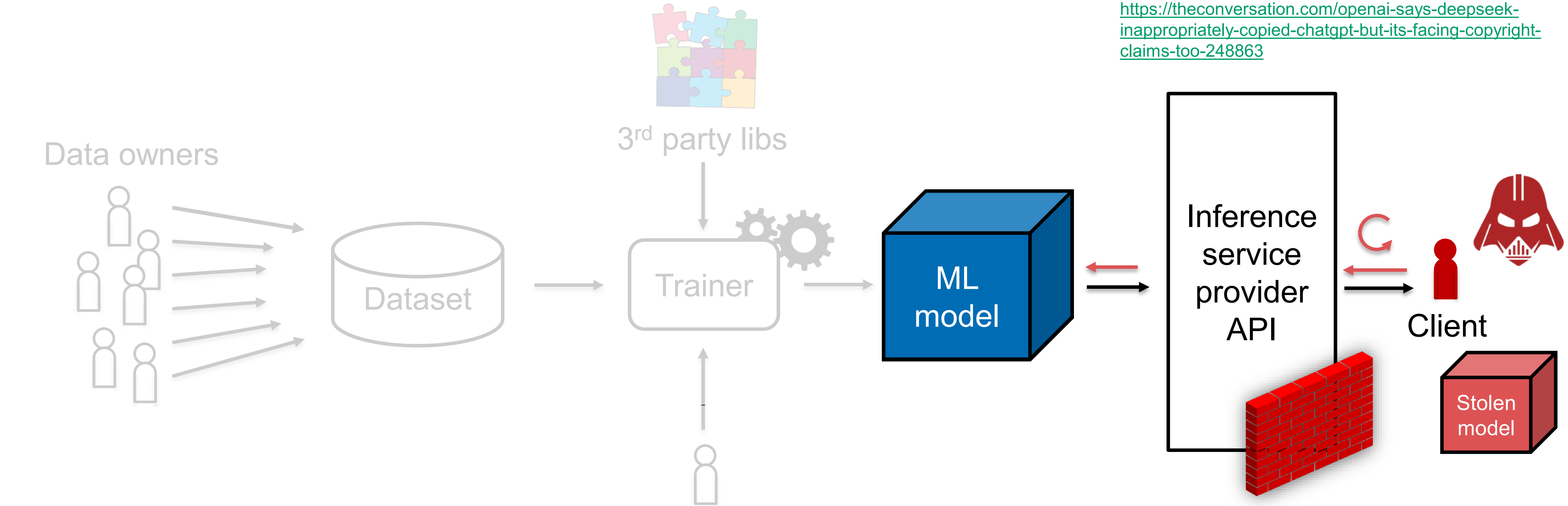
Malicious client – Model confidentiality

OpenAI says DeepSeek 'inappropriately' copied ChatGPT – but it's facing copyright claims too

Published: February 4, 2025 2.10pm EST

Lea Frermann, Shaanan Cohney, *The University of Melbourne*

https://theconversation.com/openai-says-deepseek-inappropriately-copied-chatgpt-but-its-facing-copyright-claims-too-248863



Extract/steal model

Tramer et al. – *Stealing ML models via prediction APIs*, Usenix SEC '16 (https://arxiv.org/abs/1609.02943)
Juuti et al. – *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P '19 (https://arxiv.org/abs/1805.02628)
Carlini et al. – *Stealing part of a production language model*, ICML '24 (https://arxiv.org/abs/2403.06634)

Outline

Is model stealing an important concern?

Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

Is model stealing an important concern?

Machine learning models: business advantage and intellectual property (IP)

Cost of

- gathering relevant data
- curating/labeling data
- expertise required to choose the right model training method
- resources expended in training

Adversary who “steals” the model can avoid these costs

“Steal” = derive model from someone else’s model <u>without their consent</u> to do so

How to prevent model stealing?

Outright (white-box) model stealing can be countered by

- **Hosting models behind a firewalled cloud service**
- **Protecting models using hardware-based trusted execution environments**
- **Computation with encrypted models**

Is that enough to prevent model stealing?

Outline

Is model stealing an important concern?

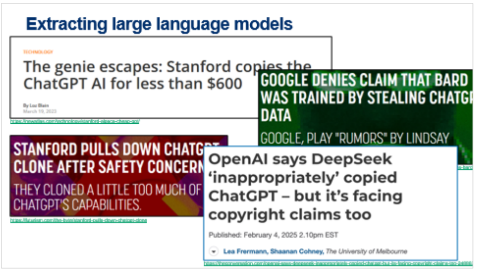
Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

Extracting models via their inference APIs

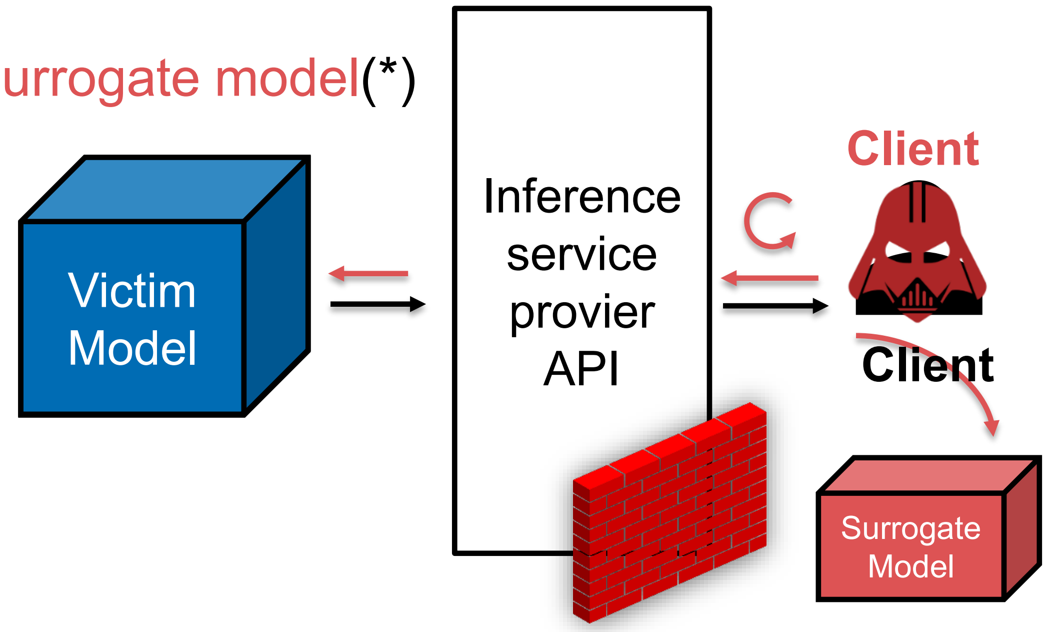


Inference APIs are oracles that leak information

Adversary

- Malicious client
- Goal: construct "comparable" [fidelity or functionality] surrogate model(*)
- Capability: access to inference API or model outputs

(*) aka "student model" or "imitation model"



Early work on extracting

- Logistic regression, decision trees[1]
- Simple convolutional neural network models[2]
- Deep neural network models[3]

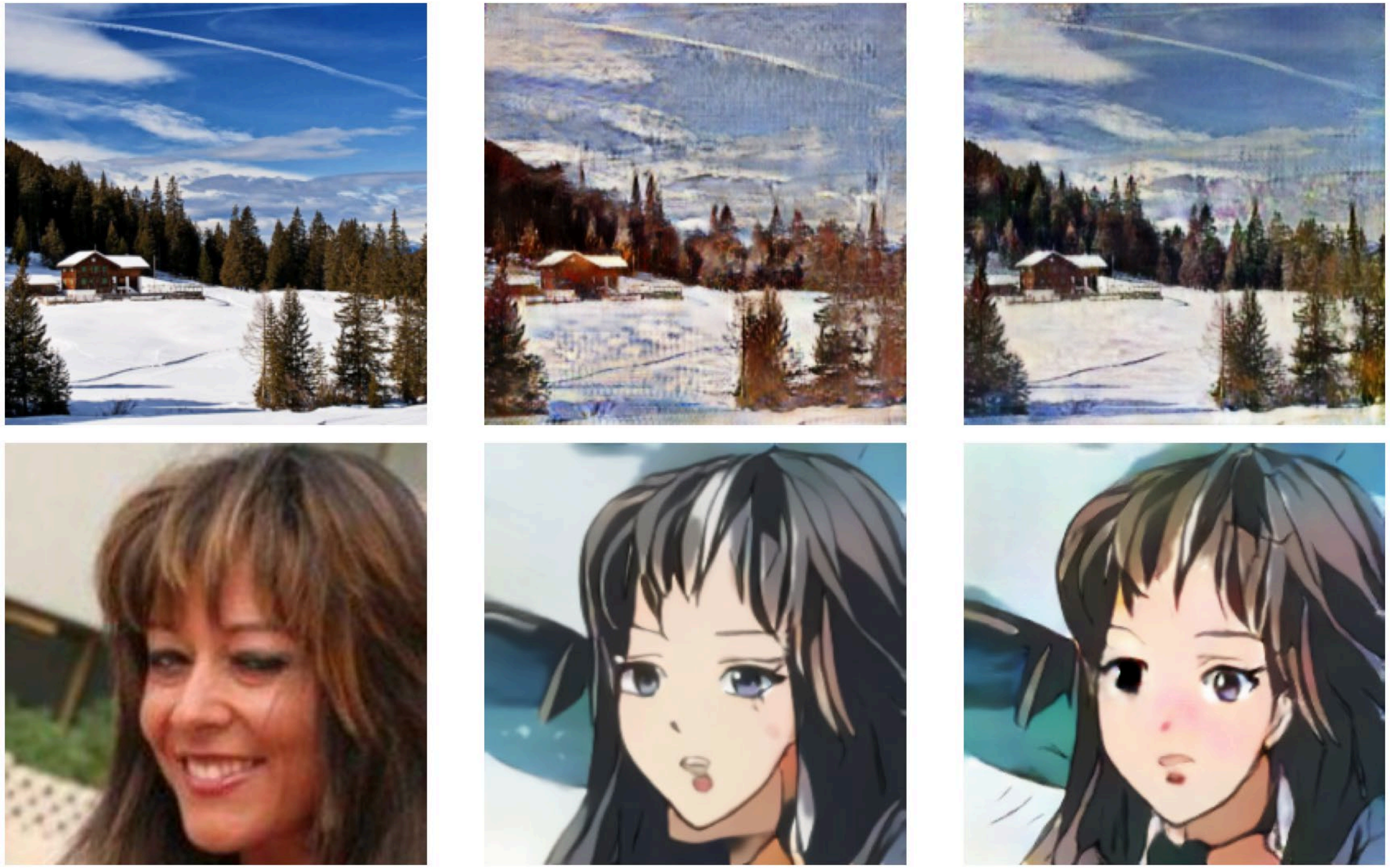
[1] Tramèr et al. – *Stealing Machine Learning Models via Prediction APIs*, Usenix SEC '16 (https://arxiv.org/abs/1609.02943)
[2] Papernot et al. – *Practical Black-Box Attacks against Machine Learning*, ASIACCS '17 (https://arxiv.org/abs/1602.02697)
[3] Juuti et al. – *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P '19 (https://arxiv.org/abs/1805.02628)

Extracting style-transfer models

Original
(unstyled)

Task 1
Monet painting

Task 2
Anime face



Szyller et al. – *Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks, '21 (https://arxiv.org/abs/2104.12623)*

Extracting natural language processing models

Techniques for extracting image classifiers don't always extend to language models

Transfer learning from pre-trained models is now very popular

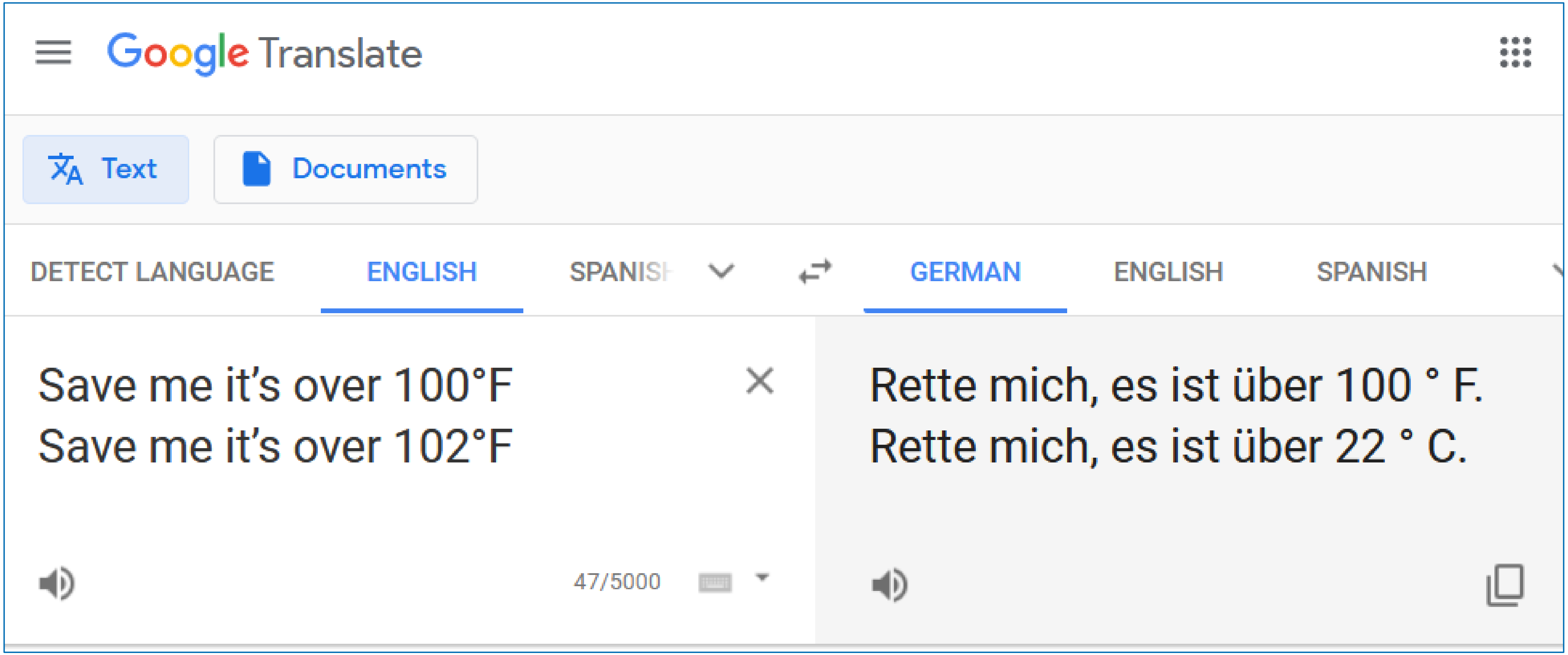
- But they make model extraction easier[1]

Krishna et al[1] show that a Knockoff-like attacks against BERT models are feasible

- Adversary unaware of target distribution or task of victim model
- Adversary queries are merely "natural" (randomly sampled sequences of words)
- In-distribution adversary queries can improve extraction efficacy

Wallace et al[2] extract real-world MT models, find transferable adversarial examples

[1] Krishna et al. – *Thieves on Sesame Street! Model Extraction of BERT-based APIs,* ICLR '20 (https://iclr.cc/virtual_2020/poster_Byl5NREFDr.html)
[2] Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems,* EMNLP '20 (https://arxiv.org/abs/2004.15015)



https://translate.google.com/#view=home&op=translate&sl=en&tl=de&text=Save%20me%20it%E2%80%99s%20over%20100%C2%B0F%0ASave%20me%20it%E2%80%99s%20over%20102%C2%B0F

Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems*, EMNLP '20 (https://arxiv.org/abs/2004.15015)

Extracting large language models

TECHNOLOGY

The genie escapes: Stanford copies the ChatGPT AI for less than $600

By Loz Blain
March 19, 2023

https://newatlas.com/technology/stanford-alpaca-cheap-gpt/

GOOGLE DENIES CLAIM THAT BARD WAS TRAINED BY STEALING CHATGP DATA

GOOGLE, PLAY "RUMORS" BY LINDSAY

es-bard

STANFORD PULLS DOWN CHATGP CLONE AFTER SAFETY CONCERN

THEY CLONED A LITTLE TOO MUCH OF CHATGPT'S CAPABILITIES.

https://futurism.com/the-byte/stanford-pulls-down-chatgpt-clone

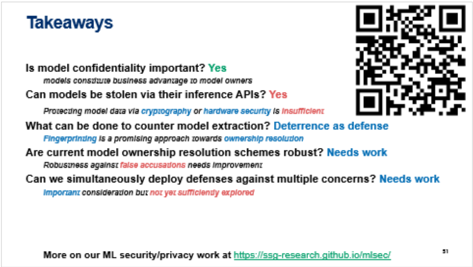
OpenAI says DeepSeek ‘inappropriately’ copied ChatGPT – but it’s facing copyright claims too

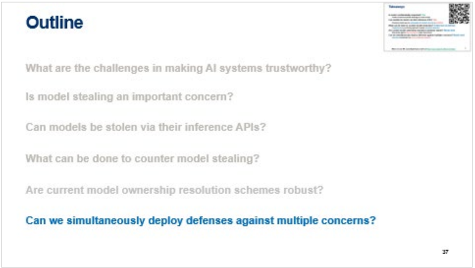
Published: February 4, 2025 2.10pm EST

Lea Frermann, Shaanan Cohney, *The University of Melbourne*

https://theconversation.com/openai-says-deepseek-inappropriately-copied-chatgpt-but-its-facing-copyright-claims-too-248863

Outline





Is model stealing an important concern? Yes

Can models be stolen via their inference APIs? Yes

- A powerful (but realistic) adversary can extract complex real-life models
- Detecting such an adversary is difficult/impossible[1]

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

[1] Atli et al. – *Extraction of Complex DNN Models: Real Threat or Boogeyman?* AAAI-EDSML ’20 (https://arxiv.org/abs/1910.05429)

Defending against model stealing

We can try to:

- prevent (or slow down[1]) model extraction, or
- detect[2] it

But current solutions are not effective

Model derivation may even become a desirable business model

Deter unauthorized model ownership via model ownership resolution (MOR):

- watermarking
- fingerprinting

October 1, 2024 Product

Model Distillation in the API

Fine-tune a cost-efficient model with the outputs of a large frontier model–all on the OpenAI platform

https://openai.com/index/api-model-distillation/

The New York Times

Artificial Intelligence › OpenAI and Trump A Look at OpenAI's Operator What Is DeepSeek? DeepSeek's Rise Quiz

OpenAI Says DeepSeek May Have Improperly Harvested Its Data

The San Francisco start-up claims that its Chinese rival may have used data generated by OpenAI technologies to build new systems.

https://www.nytimes.com/2025/01/29/technology/openai-deepseek-data-harvest.html

"We are aware of and reviewing indications that DeepSeek may have inappropriately distilled our models, and will share information as we know more," the spokesperson said, adding that the company was not accusing DeepSeek of a security breach.

Distillation is often prohibited in LLMs' terms of service, but is common in the industry.

[1] Dziedzic et al. – *Increasing the Cost of Model Extraction with Calibrated Proof of Work*, ICLR '22 (https://openreview.net/pdf?id=EAy7C1cgE1L)
[2] Atli et al. – *Extraction of Complex DNN Models: Real Threat or Boogeyman?*, AAAI-EDSML '20 (https://arxiv.org/abs/1910.05429)

Watermarking

Embed watermark while training (potentially) victim model[1]

- Choose incorrect labels for a set of samples (watermark set, WM)
- Cannot resist model extraction

Embed watermark at the inference API[2]

- Use a mapping function to decide when to return incorrect predictions for queries
- Finding suitable mapping functions is difficult

Watermarking schemes tend to be not robust[3] and reduce utility

[1] Yadi et al. – *Watermarking Deep Neural Networks by Backdooring,* Usenix SEC '18 https://www.usenix.org/node/217594
[2] Szyller et. al. – *DAWN: Dynamic Adversarial Watermarking of Neural Networks,* ACM MM '21 (https://arxiv.org/abs/1906.00830)
[3] Lukas et al. – *SoK: How Robust is Image Classification Deep Neural Network Watermarking*? IEEE S&P '22 (https://arxiv.org/abs/2108.04974)

Fingerprinting

Conferrable adversarial examples[1]

- Distinguish between conferrable adversarial examples vs. other transferable ones
- Computationally expensive

Dataset inference[2]

- Distinguish between models trained with different datasets
- Susceptible to false positives/negatives under certain conditions[3]

GrOVe[4]

- Use GNN embeddings as fingerprints (for GNN models)
- Effective against high-fidelity extraction[5] but likely not against low-fidelity extraction

[1] Lukas et al. – *Deep Neural Network Fingerprinting by Conferrable Adversarial Examples,* ICLR '21 (https://openreview.net/forum?id=VqzVhqxkjH1)
[2] Maini et al. – *Dataset Inference Ownership Resolution in Machine Learning*, ICLR '21 (https://openreview.net/pdf?id=hvdKKV2yt7T)
[3] Szyller et al. – *On the Robustness of Dataset Inference*, TMLR '23 (https://arxiv.org/abs/2210.13631)
[4] Waheed et al. – *GrOVe: Ownership Verification of Graph Neural Networks using Embeddings,* IEEE S&P '24 (https://arxiv.org/abs/2304.08566)
[5] Shen et al. – *Model Stealing Attacks Against Inductive Graph Neural Networks,* IEEE S&P '22 (https://arxiv.org/abs/2112.08331)

Outline

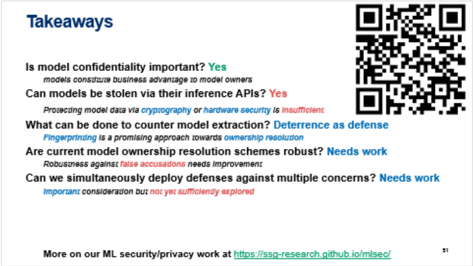
Is model stealing an important concern?

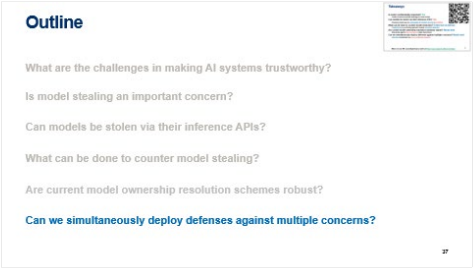
Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?





Robustness of model ownership resolution schemes

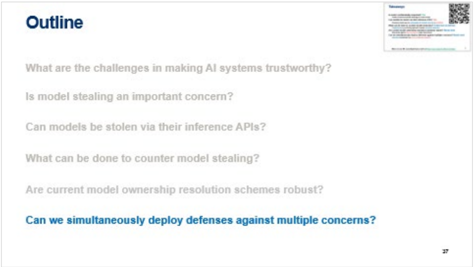
Model ownership resolution (MOR) must be robust against two types of attackers

Malicious suspect:
- tries to evade verification (e.g., pruning, fine-tuning, noising)

Malicious accuser:
- tries to frame an independent model owner
- (secure) timestamping (watermark/fingerprint and model) is the only defense in prior work

So far, research has focused on robustness against malicious suspects

False claims against MORs



We show how malicious accusers can make false claims against independent models:

- adversary deviates from watermark/fingerprint generation procedure
  - E.g., via transferrable adversarial examples
- but still subject to specified verification procedure

Our contributions:

- formalize the notion of false claims against MORs
- provide a generalization of MORs
- demonstrate effective false claim attacks
- discuss potential countermeasures

Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

Watermarking by backdooring[1]

Watermark generation:

- choose some out-of-distribution samples as watermark
  - assigned with incorrect labels
- train using the watermark alongside normal training data (or fine tune)
  - model memorizes watermark
- obtain timestamp on commitment of model and watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy → stolen
  - a few matching / low WM accuracy → not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)

Watermarking by backdooring[1]: false claim[2]

Watermark generation:

- choose some out-of-distribution samples as watermark
  - assigned with incorrect labels
- train using the watermark alongside your normal training data (or fine tune)
  - model memorizes watermark
- obtain timestamp on commitment of model and watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy → stolen
  - a few matching / low WM accuracy → not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)
[2] Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

Watermarking by backdooring[1]: false claim[2]

False watermark generation:

- choose some out-of-distribution samples as false watermark
- perturb these samples to craft transferable adversarial examples
- obtain timestamp on commitment of model and false watermark

Watermark verification:

- query suspect model using watermark
- compare predictions to the assigned (incorrect) labels:
  - many matching / high WM accuracy -> stolen
  - a few matching / low WM accuracy > not stolen
- check commitment and timestamp

[1] Adi et al. – *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring*, Usenix SEC 2018 (https://arxiv.org/abs/1802.04633)
[2] Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

Mitigating false claims against MORs

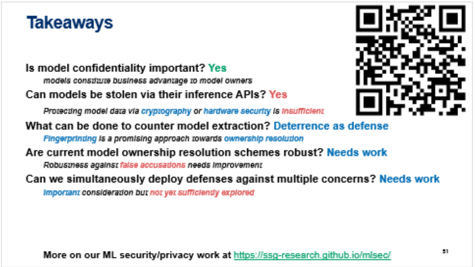
Judge generates watermarks/fingerprints: bottleneck

Judge verifies watermarks/fingerprints were generated correctly: expensive

Train models with transferable adversarial examples: accuracy loss

Zhang et al. – *False Claims Against Model Ownership Resolution*, Usenix SEC '24 (https://arxiv.org/abs/2304.06607)

Outline


Takeaways

Is model confidentiality important? Yes
models constitute business advantage to model owners

Can models be stolen via their inference APIs? Yes
Protecting model data via cryptography or hardware security is insufficient

What can be done to counter model extraction? Deterrence as defense
Fingerprinting is a promising approach towards ownership resolution

Are current model ownership resolution schemes robust? Needs work
Robustness against false accusations needs improvement

Can we simultaneously deploy defenses against multiple concerns? Needs work
Important consideration but not yet sufficiently explored

More on our ML security/privacy work at https://ssg-research.github.io/mlsec/

51

What are the challenges in making AI systems trustworthy?

Is model stealing an important concern?

Can models be stolen via their inference APIs?

What can be done to counter model stealing?

Are current model ownership resolution schemes robust?

Can we simultaneously deploy defenses against multiple concerns?

Towards trustworthy AI

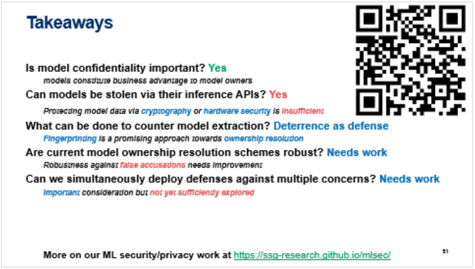
Secure, privacy-preserving, aligned, fair, and explainable

TABLE V
TOP ATTACK

Which attack would affect your org the most?	*Distribution*
Poisoning (e.g: [21])	10
Model Stealing (e.g: [22])	6
Model Inversion (e.g: [23])	4
Backdoored ML (e.g: [24])	4
Membership Inference (e.g: [25])	3
Adversarial Examples (e.g: [26])	2
Reprogramming ML System (e.g: [27])	0
Adversarial Example in Physical Domain (e.g: [5])	0
Malicious ML provider recovering training data (e.g: [28])	0
Attacking the ML supply chain (e.g: [24])	0
Exploit Software Dependencies (e.g: [29])	0

Kumar et al. – *Adversarial Machine Learning – Industry Perspectives*, IEEE SPW '20 (https://arxiv.org/abs/2002.05646)

Unintended interactions



Prior work explored defenses to mitigate specific risks

- Defenses typically evaluated only vs. those specific risks they protect against

But practitioners need to deploy multiple defenses simultaneously

- Can two defenses interact negatively with each other?
- Does a defense exacerbate or ameliorate some other (unrelated) risk?

Ownership resolution vs. other security/privacy concerns

There are considerations other than model ownership resolution:

- model evasion (defense: adversarial training)
- training data reconstruction (defense: differential privacy)
- membership inference (defense: regularization, early stopping)
- model poisoning (defense: regularization, outlier/anomaly detection)
- …

How do ownership resolution schemes interact with the other defenses?

We investigated pairwise interactions of:

model watermarking data watermarking fingerprinting	**WITH**	differential privacy adversarial training

Ownership resolution vs. other security/privacy concerns

If two techniques A and B in combination result in too high a drop in

- model accuracy (ϕ_{ACC}) or
- metric for A (ϕ_A) or
- metric for B (ϕ_B)

then A and B are in conflict

Defense	Dataset	Defense	
		DP	ADV. TR.
WM	MNIST	$\phi_{ACC}\ \phi_{WM}$	$\phi_{ACC}\ \phi_{WM}\ \phi_{ADV}$
	FMNIST	$\phi_{ACC}\ \phi_{WM}$	$\phi_{ACC}\ \phi_{WM}\ \phi_{ADV}$
	CIFAR10	$\phi_{ACC}\ \phi_{WM}$	$\phi_{ACC}\ \phi_{WM}\ \phi_{ADV}$
RAD-DATA	MNIST	$\phi_{ACC}\ \phi_{RAD\text{-}DATA}$	$\phi_{ACC}\ \phi_{RAD\text{-}DATA}\ \phi_{ADV}$
	FMNIST	$\phi_{ACC}\ \phi_{RAD\text{-}DATA}$	$\phi_{ACC}\ \phi_{RAD\text{-}DATA}\ \phi_{ADV}$
	CIFAR10	$\phi_{ACC}\ \phi_{RAD\text{-}DATA}$	$\phi_{ACC}\ \phi_{RAD\text{-}DATA}\ \phi_{ADV}$
DI	MNIST	$\phi_{ACC}\ \phi_{DI}$	$\phi_{ACC}\ \phi_{DI}\ \phi_{ADV}$
	FMNIST	$\phi_{ACC}\ \phi_{DI}$	$\phi_{ACC}\ \phi_{DI}\ \phi_{ADV}$
	CIFAR10	$\phi_{ACC}\ \phi_{DI}$	$\phi_{ACC}\ \phi_{DI}\ \phi_{ADV}$

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)

Interaction between ML defenses

Property	Adversarial Training	Differential Privacy	Membership Inference	Oblivious Training	Model/Gradient Inversion	Model Poisoning	Model Watermarking	Model Fingerprinting	Data Watermarking	Explainability	Fairness
Adversarial Training	X	[5]	[9]	?	?	[7]	OURS	OURS	OURS	[11]	?
Differential Privacy		X	[3, 6]	?	?	?	OURS	OURS	OURS	?	[1, 2, 8]
Membership Inference			X	?	?	[10]	?	?	?	?	?
Oblivious Training				X	?	?	?	?	?	?	?
Model/Gradient Inversion					X	?	?	?	?	?	?
Model Poisoning						X	?	?	?	?	?
Model Watermarking							X	?	?	?	?
Model Fingerprinting								X	?	[4]	?
Data Watermarking									X	?	?
Fairness										X	?
Explainability											X

REFERENCES

[1] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS P)*. 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028

[2] Victoria Cheng, Vinith M. Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 149–160. https://doi.org/10.1145/3442188.3445879

[3] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2020. Investigating Membership Inference Attacks under Data Dependencies. https://doi.org/10.48550/ARXIV.2010.12112

[4] Hengrui Jia, Hongyu Chen, Jonas Guan, Ali Shahin Shamsabadi, and Nicolas Papernot. 2022. A Zest of LIME: Towards Architecture-Independent Model Distances. In *International Conference on Learning Representations*. https://openreview.net/forum?id=OUz_9TiTv9j

[5] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. 656–672. https://doi.org/10.1109/SP.2019.00044

[6] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 866–882. https://doi.org/10.1109/SP40001.2021.00069

[7] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. 2020. *A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models.* Association for Computing Machinery, New York, NY, USA, 85–99. https://doi.org/10.1145/3372297.3417253

[8] Adam Pearce. 2022. Can a Model Be Differentially Private and Fair? https://pair.withgoogle.com/explorables/private-and-fair/. Online; accessed 7 April 2022.

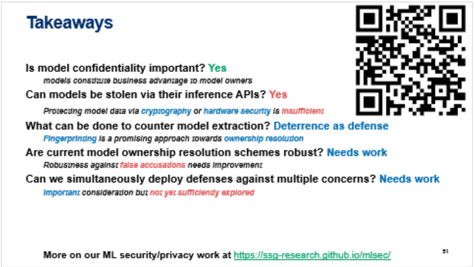
[9] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery, 241–257. https://doi.org/10.1145/3319535.3354211

[10] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. https://doi.org/10.48550/ARXIV.2204.00032

[11] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* https://openreview.net/forum?id=SyxAb30cY7

Szyller and Asokan – *Conflicting Interactions Among Protections Mechanisms for Machine Learning Models*, AAAI '23 (https://arxiv.org/abs/2207.01991)

Defense vs. other risks



How does a defense impact susceptibility to other (unrelated) risks?

Conjecture: overfitting and memorization are influence defenses and risks[1][2]

- Effective defenses may induce, reduce or rely on overfitting or memorization
- Risks tend to exploit overfitting or memorization
- Underlying factors that influence memorization/overfitting can be identified

Distinguished Paper Award

Recently built a toolkit, Amulet, for comparative evaluation of attacks & defenses[3]

Currently working on "how to easily determine if a given set of defenses conflict?"[4]

[1] Duddu, Szyller, and Asokan - *SoK: Unintended Interactions among Machine Learning Defenses and Risks,* IEEE S&P '24. (https://arxiv.org/abs/2312.04542)
[2] Blog article: https://crysp.uwaterloo.ca/ssg/blog/2024/05/unintended-interactions-among-ml.html
[3] Amulet repo: https://github.com/ssg-research/amulet
[4] Duddu, Zhang, Asokan – Combining Machine learning Defenses without Conflicts. (https://arxiv.org/abs/2411.09776)

Takeaways

Is model confidentiality important? Yes

models constitute business advantage to model owners

Can models be stolen via their inference APIs? Yes

Protecting model data via cryptography or hardware security is insufficient

What can be done to counter model extraction? Deterrence as defense

Fingerprinting is a promising approach towards ownership resolution

Are current model ownership resolution schemes robust? Needs work

Robustness against false accusations needs improvement

Can we simultaneously deploy defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

More on our ML security/privacy work at https://ssg-research.github.io/mlsec/

Takeaways

Is model confidentiality important? Yes

models constitute business advantage to model owners

Can models be stolen via their inference APIs? Yes

Protecting model data via cryptography or hardware security is insufficient

What can be done to counter model extraction? Deterrence as defense

Fingerprinting is a promising approach towards ownership resolution

Are current model ownership resolution schemes robust? Needs work

Robustness against false accusations needs improvement

Can we simultaneously deploy defenses against multiple concerns? Needs work

Important consideration but not yet sufficiently explored

Other research topics:

ML security/privacy:

ML ownership resolution, Conflicting ML defenses, ML property attestation, robust concept removal in gen AI

Platform security: hardware-assisted run-time security, secure outsourced computing

Open (postdoc, grad student) positions to help lead our work: ML security/privacy, platform security

https://asokan.org/asokan/research/SecureSystems-open-positions-Jan2024.php