Imitation Attacks and Defenses for Black-box Machine Translation Systems

Eric Wallace, Mitchell Stern, Dawn Song

UC Berkeley

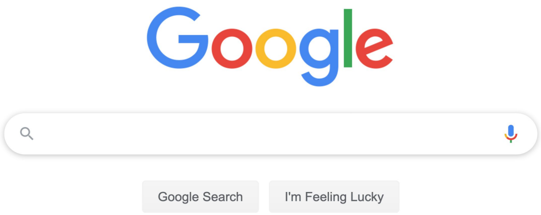
Eric Wallace

Mitchell Stern

Dawn Song

Production NLP Models Are Lucrative

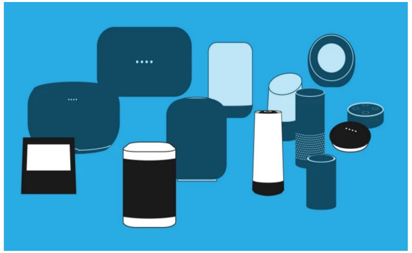
Production NLP Models Are Lucrative



Information Retrieval

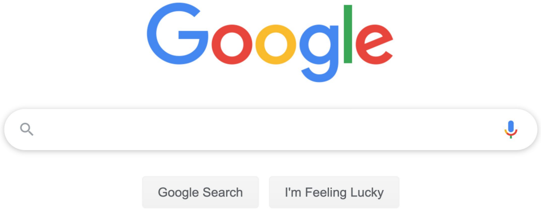
Machine Translation

Text + Speech Generation



Smart Assistants

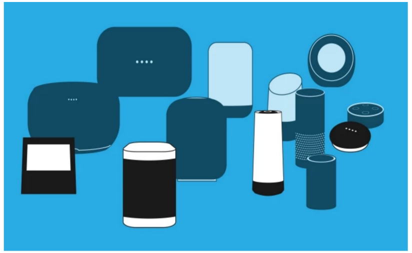
Production NLP Models Are Lucrative



Information Retrieval

Machine Translation

Text + Speech Generation



Smart Assistants

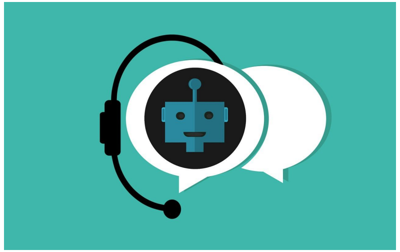
Result of **large investments** into data annotation and model design

Production NLP Models Make Critical Predictions

Production NLP Models Make Critical Predictions

Fake News Detection

Machine Translation



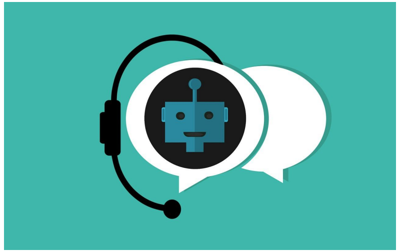
Dialogue Systems

Spam Filtering

Production NLP Models Make Critical Predictions

Fake News Detection

Machine Translation



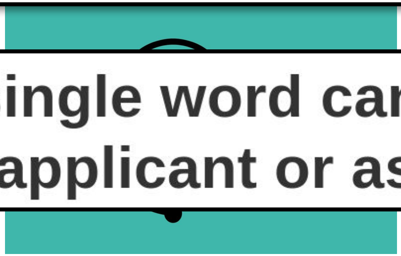
Dialogue Systems

Spam Filtering

Errors can have **negative societal consequences**

Production NLP Models Make Critical Predictions

Facebook translates 'good morning' into 'attack them', leading to arrest



Changing a single word can alter the way an AI program judges a job applicant or assesses a medical claim.

Dialogue Systems

Spam Filtering

Errors can have **negative societal consequences**

An Adversary’s Viewpoint

An Adversary's Viewpoint

An adversary can benefit financially by **stealing models**

An Adversary's Viewpoint

An adversary can benefit financially by **stealing models**

- avoid long-term API costs by stealing models upfront

An Adversary’s Viewpoint

An adversary can benefit financially by **stealing models**

- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality

An Adversary's Viewpoint

An adversary can benefit financially by **stealing models**

- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality

An adversary can benefit financially or harm society by **breaking models**

An Adversary’s Viewpoint

An adversary can benefit financially by **stealing models**

- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality

An adversary can benefit financially or harm society by **breaking models**

- manipulate the stock market by fooling sentiment models

An Adversary’s Viewpoint

An adversary can benefit financially by **stealing models**

- avoid long-term API costs by stealing models upfront
- launch a competitor service of similar quality

An adversary can benefit financially or harm society by **breaking models**

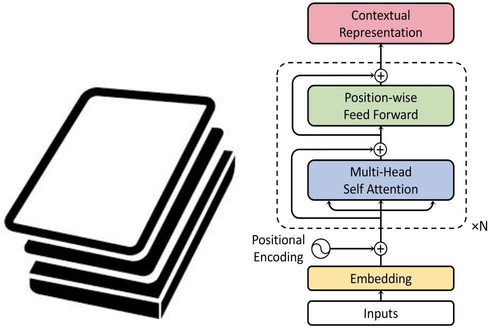
- manipulate the stock market by fooling sentiment models
- bypass classifiers of fake news or hate speech

Our Contributions

- Common Practice: keep data + model hidden

Our Contributions

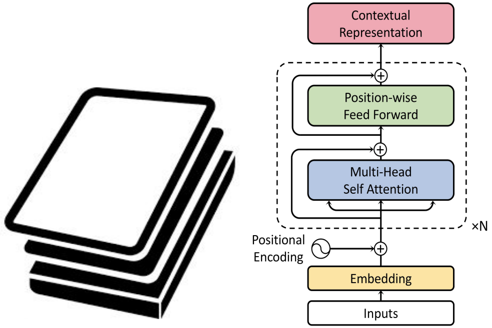
- Common Practice: keep data + model hidden



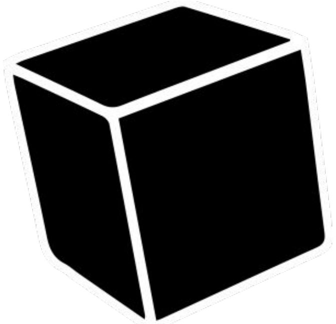
Hidden Data + Model

Our Contributions

- Common Practice: keep data + model hidden



Hidden Data + Model



Black-box API

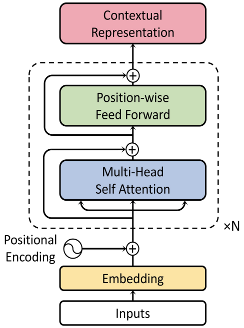
How are you?

Wie geht es dir?

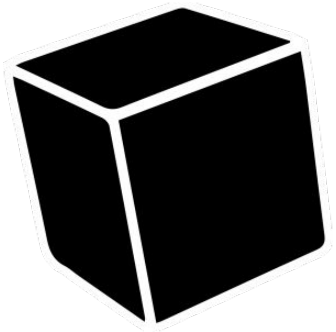
Adversary

Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!



Hidden Data + Model



Black-box API

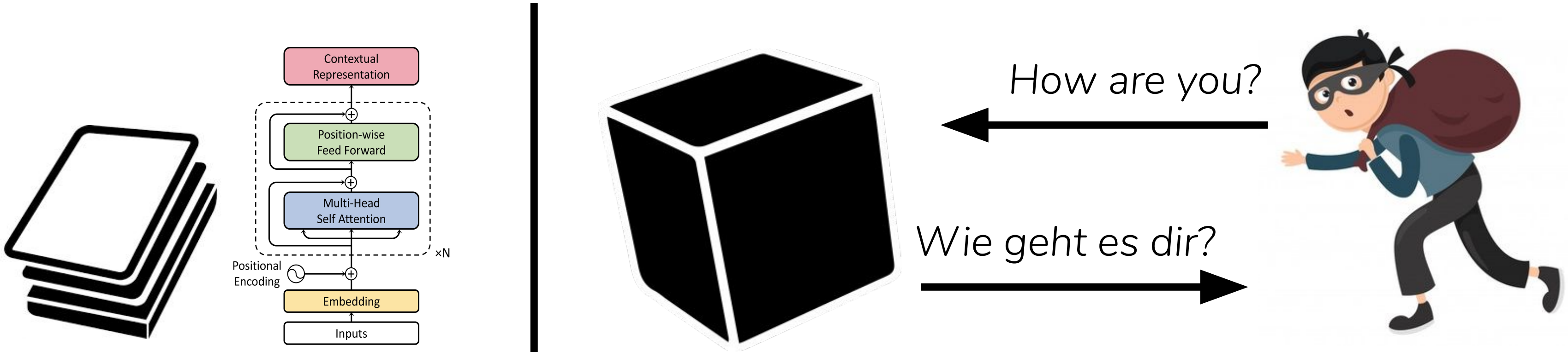
How are you?

Wie geht es dir?

Adversary

Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
  - adversaries can imitate black-box models



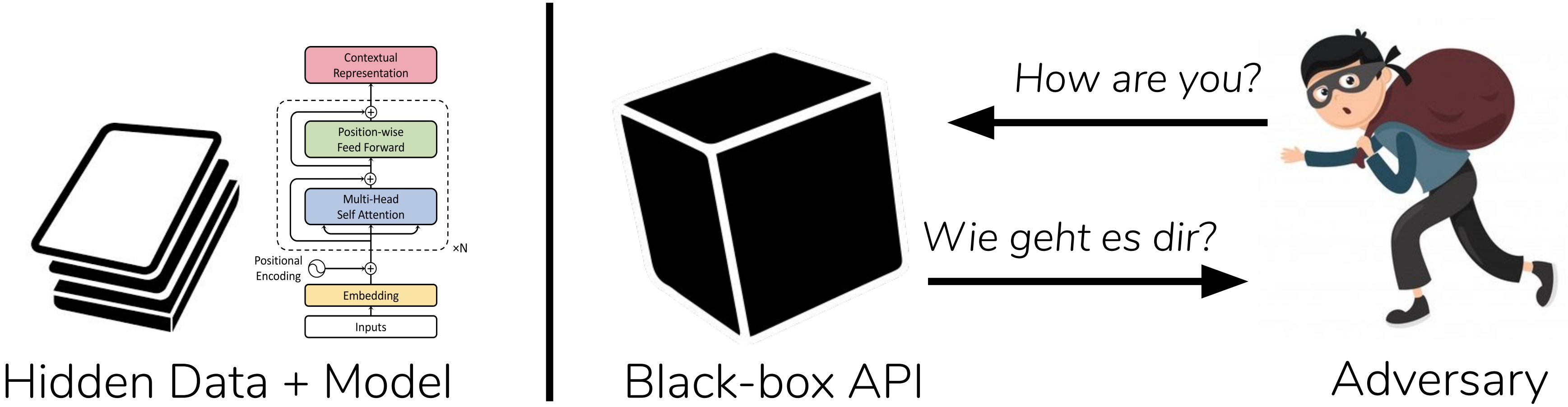
Hidden Data + Model

Black-box API

Adversary

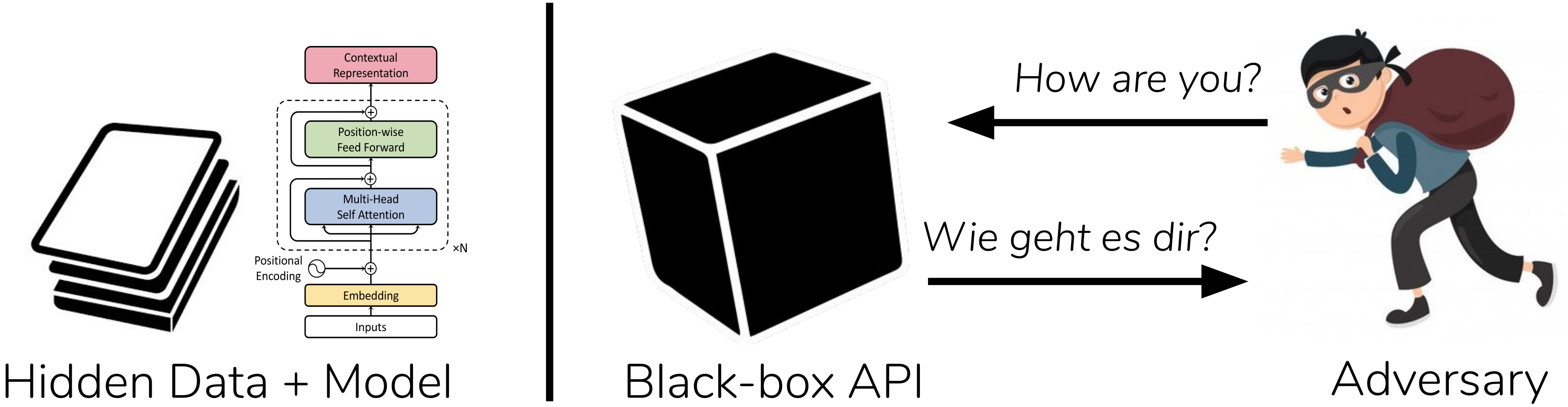
Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
  - adversaries can imitate black-box models
  - imitation models help break black-box models



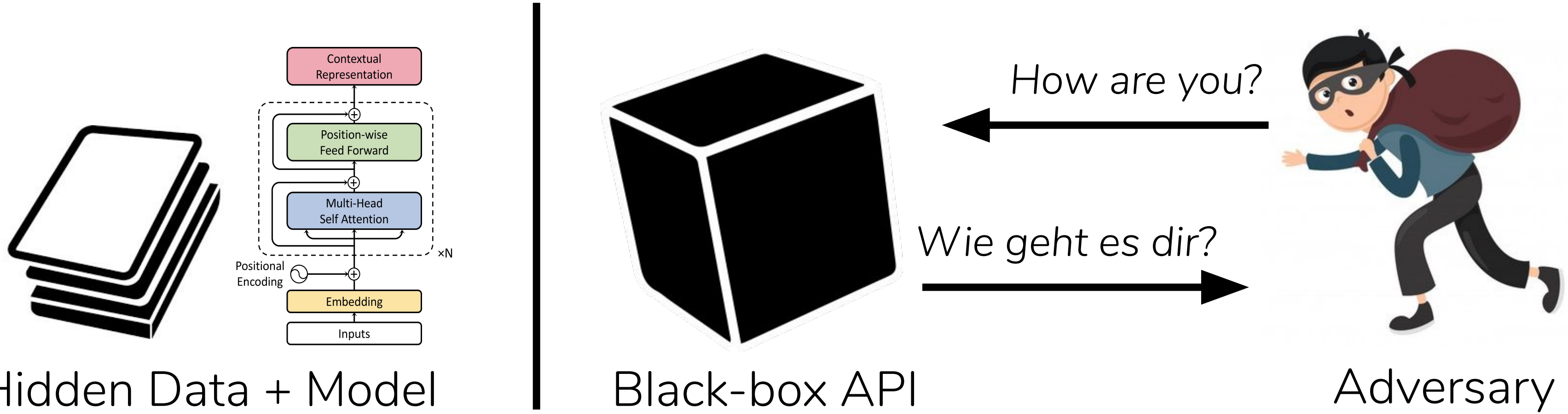
Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
  - adversaries can imitate black-box models
  - imitation models help break black-box models
  - new defenses mitigate adversaries



Our Contributions

- Common Practice: keep data + model hidden
- Our paper: this is **not enough** to protect NLP models!
    - adversaries can imitate black-box models
    - imitation models help break black-box models
    - new defenses mitigate adversaries
- We consider machine translation (MT) as a case study



Our Task: Machine Translation

- We use machine translation (MT) as a case study
  - seq-to-seq task (Pal 2019, Krishna 2020 consider classification)
  - lucrative product
  - errors can be costly

Our Task: Machine Translation

- We use machine translation (MT) as a case study
  - seq-to-seq task (Pal 2019, Krishna 2020 consider classification)
  - lucrative product
  - errors can be costly

- We explore attacks on production systems (Google, Bing, Systran)
  - ethical concerns: our paper describes how we followed standard security practices and minimized harm

Model Stealing: How We Imitate MT Models

Model Stealing: How We Imitate MT Models

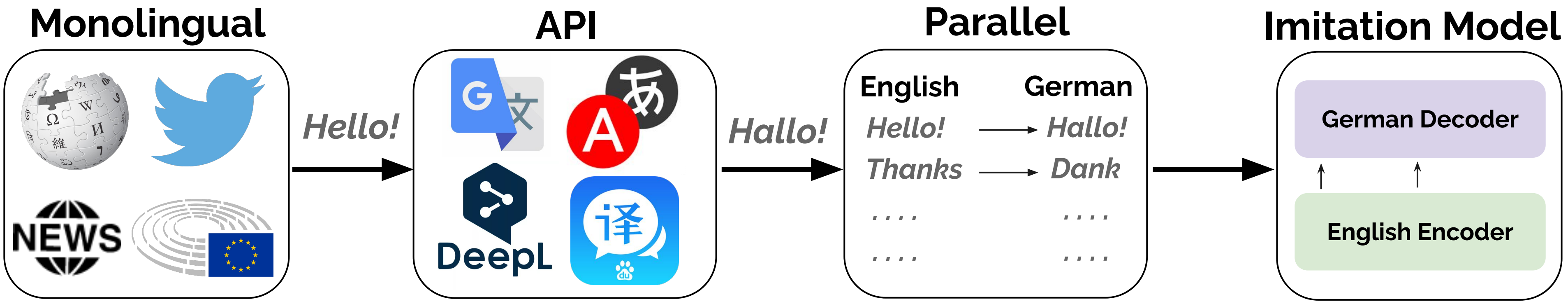
- Goal: train *imitation model* that is similar to black-box API

Model Stealing: How We Imitate MT Models

- Goal: train *imitation model* that is similar to black-box API
- Method: query sentences and use API output as training data

Model Stealing: How We Imitate MT Models

- Goal: train *imitation model* that is similar to black-box API
- Method: query sentences and use API output as training data



Model Stealing: How We Imitate MT Models

- Goal: train *imitation model* that is similar to black-box API
- Method: query sentences and use API output as training data
- Not just model distillation:
  - unknown data distribution

Model Stealing: How We Imitate MT Models

- Goal: train *imitation model* that is similar to black-box API
- Method: query sentences and use API output as training data
- Not just model distillation:
  - unknown data distribution
  - no distribution or feature matching losses

Simulated Model Stealing Experiments

Simulated Model Stealing Experiments

Setup:

- Black-box MT *victim* model for German-English

Simulated Model Stealing Experiments

Setup:

- Black-box MT *victim* model for German-English
- Vary imitation model’s architecture and queried sentences

Simulated Model Stealing Experiments

Setup:

- Black-box MT *victim* model for German-English
- Vary imitation model's architecture and queried sentences

Evaluation metrics:

- BLEU on in-domain and out-of-domain data

Simulated Model Stealing Experiments

Setup:

- Black-box MT *victim* model for German-English
- Vary imitation model's architecture and queried sentences

Evaluation metrics:

- BLEU on in-domain and out-of-domain data
- Output similarity using inter-system BLEU

Simulated Model Stealing Experiments

Setup:

- Black-box MT *victim* model for German-English
- Vary imitation model's architecture and queried sentences

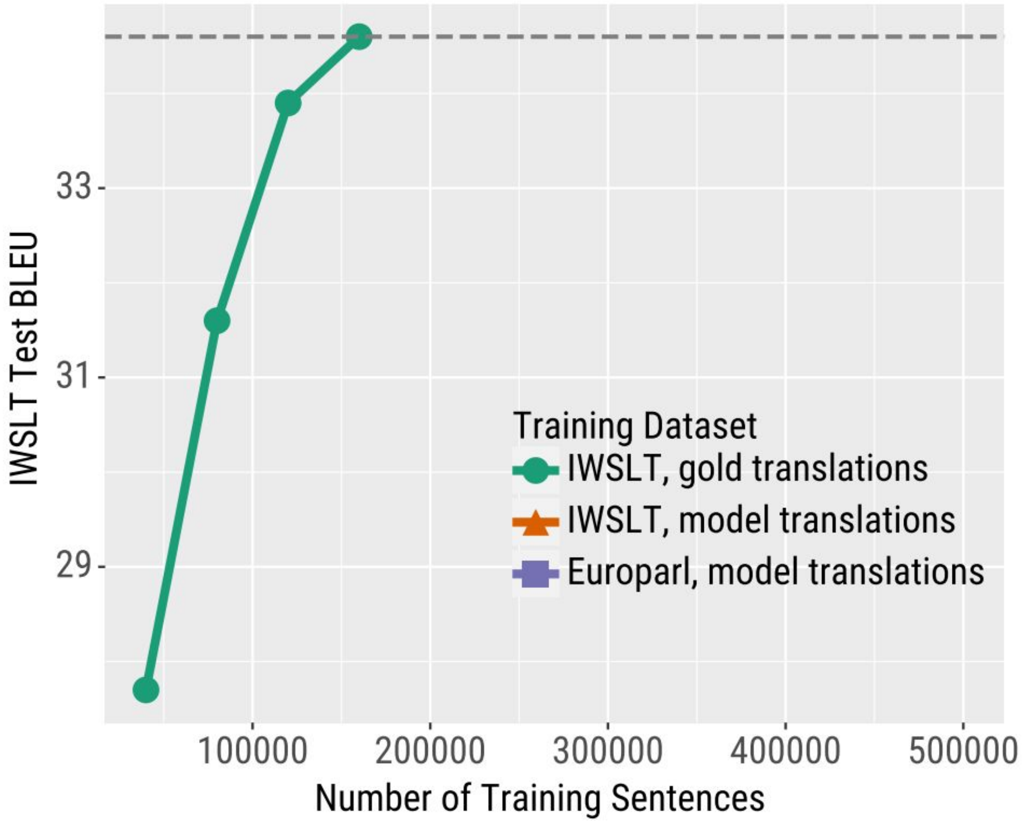
Evaluation metrics:

- BLEU on in-domain and out-of-domain data
- Output similarity using inter-system BLEU

For all architectures, data settings, and evaluation metrics, the imitation models closely match their victims

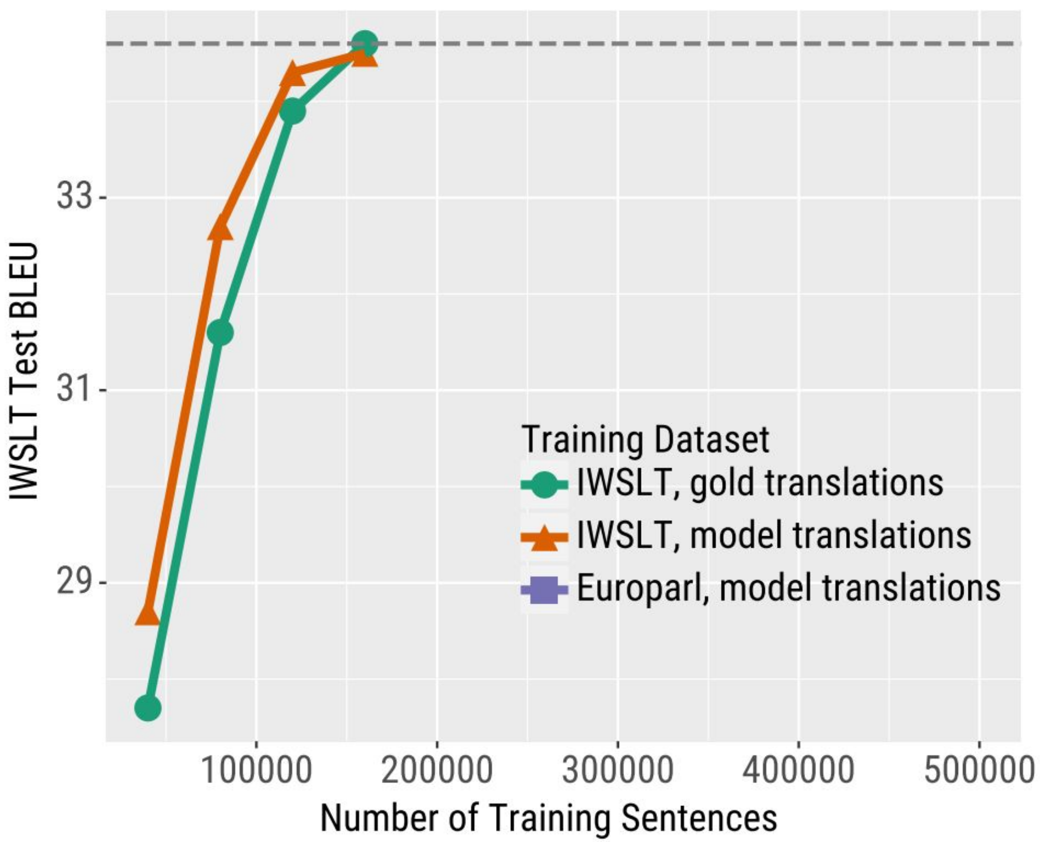
Simulated Imitation Models

- Training on OOD input queries slows but does not prevent imitat



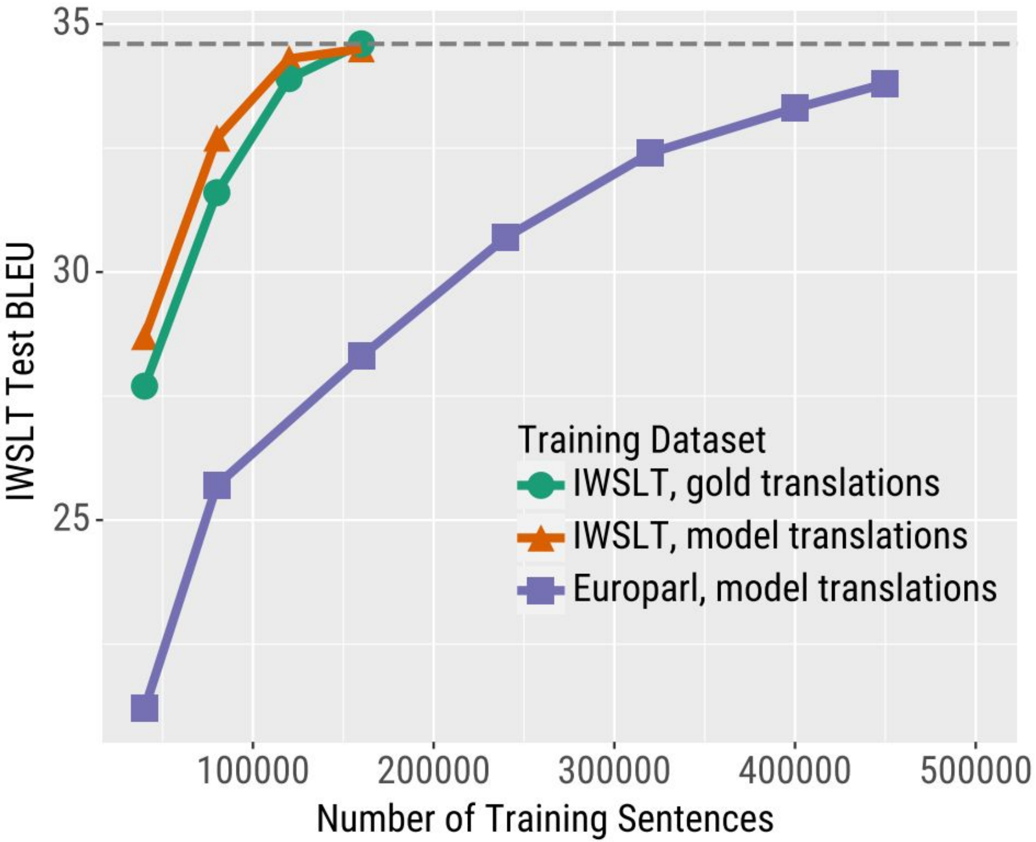
Simulated Imitation Models

- Training on OOD input queries slows but does not prevent imitat



Simulated Imitation Models

- Training on OOD input queries slows but does not prevent imitat



Imitating Production Models

- Imitate production systems on English-German and Nepali-English

Imitating Production Models

- Imitate production systems on English-German and Nepali-English

- We closely match production systems

In-domain BLEU

Model	Google	Bing	Systran
Official	32.0	32.9	27.8
Imitation	31.5	32.4	27.6

Imitating Production Models

- Imitate production systems on English-German and Nepali-English

- We closely match production systems

	Model	Google	Bing	Systran
In-domain BLEU	Official	32.0	32.9	27.8
	Imitation	31.5	32.4	27.6
Out-of-domain BLEU	Official	32.0	32.7	32.0
	Imitation	31.1	32.0	31.4

Breaking MT Models

Breaking MT Models

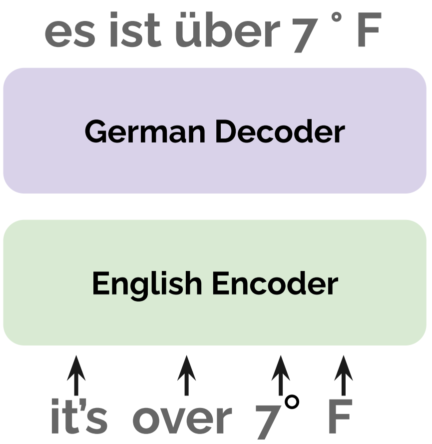
- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?

Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
- Simple idea: transfer attacks from imitation models

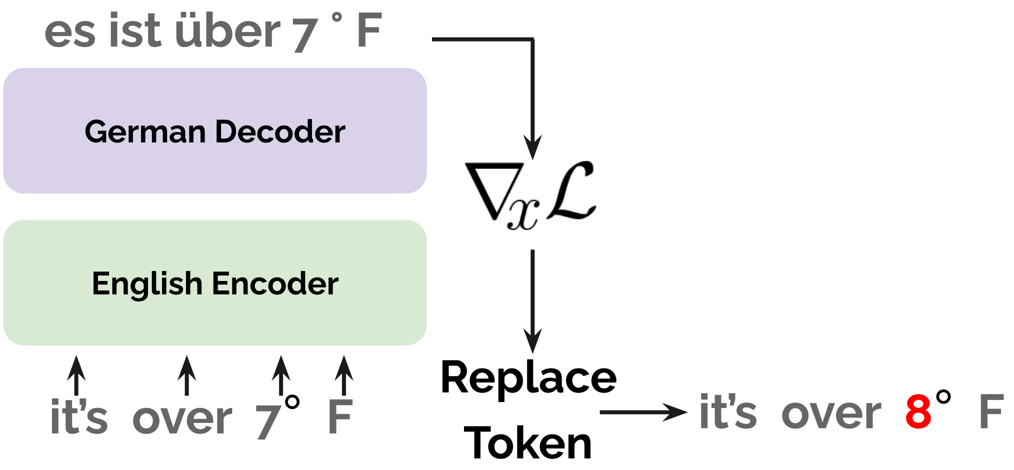
Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
- Simple idea: transfer attacks from imitation models



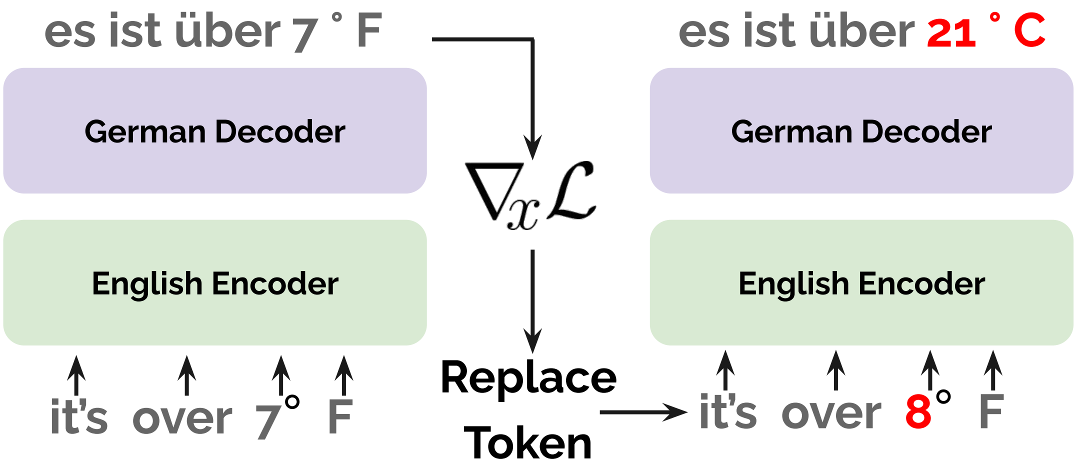
Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
- Simple idea: transfer attacks from imitation models



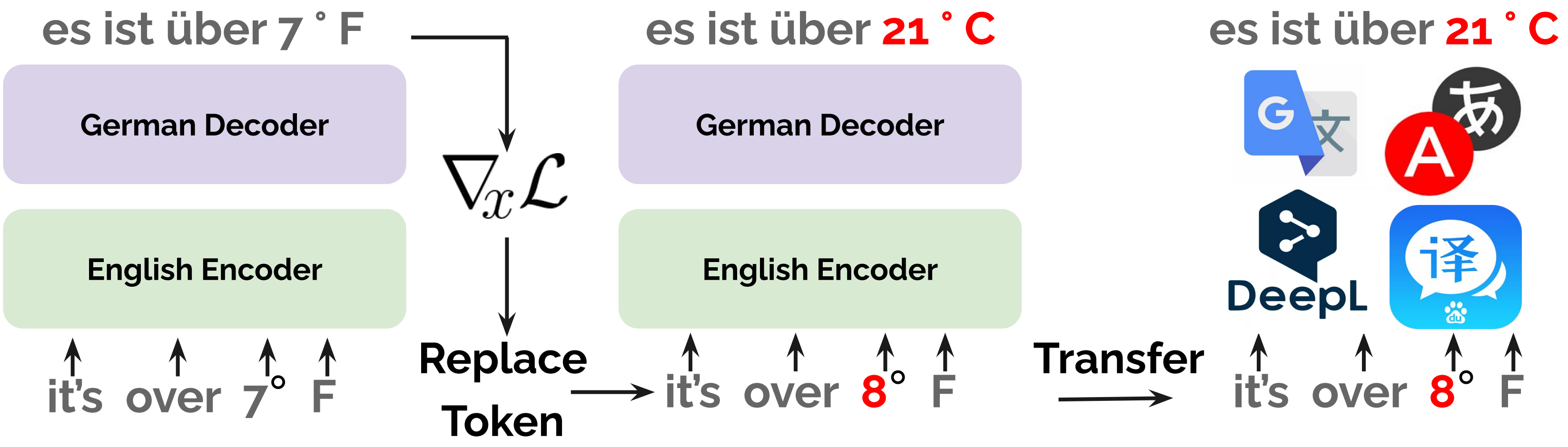
Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
- Simple idea: transfer attacks from imitation models



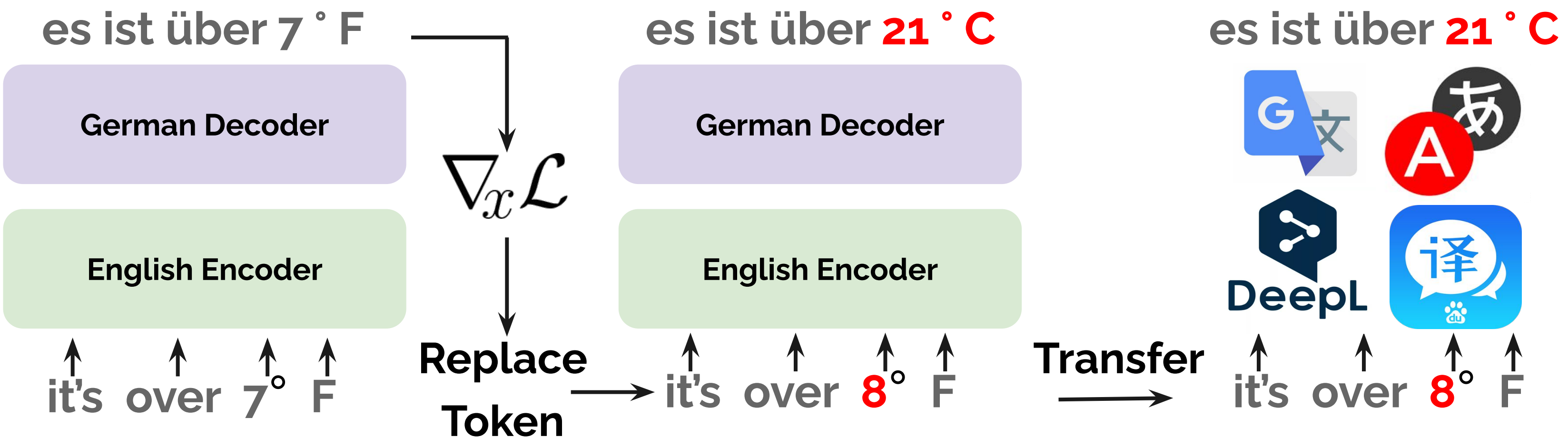
Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
- Simple idea: transfer attacks from imitation models



Breaking MT Models

- Most adversarial attacks for NLP assume white-box access
  - How to do black-box attacks?
- Simple idea: transfer attacks from imitation models



- We consider benign but representative adversarial attacks
  - same techniques would apply to malicious use cases

Untargeted Universal Trigger Attack

- A phrase that commonly causes mistranslations when appended to *any* input

Untargeted Universal Trigger Attack

- A phrase that commonly causes mistranslations when appended to *any* input

SYSTRAN

I heard machine translation is now superhuman **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**

In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt.

(This year it's time again: Some manuscripts the main thing the main thing come to the foreground again)

Untargeted Universal Trigger Attack

- A phrase that commonly causes mistranslations when appended to *any* input

SYSTRAN

I heard machine translation is now superhuman **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**

In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt.

(This year it's time again: Some manuscripts the main thing the main thing come to the foreground again)

- Transfer trigger to Systran
- **5.8% similarity** between Systran's predictions before and after appending trigger (as measured by modified pairwise BLEU)

Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation

Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation

Bing

Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday **, kostenfrei übrigens categories ņstrian hl SM** **the revolution begins at 6pm**

Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs.

(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)

Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation

Bing

Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday **, kostenfrei übrigens categories ņstrian hl SM** **the revolution begins at 6pm**

Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs.

(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)

- Transfer trigger to Bing
- **76%** of messages after trigger are successfully dropped

Targeted Flips Attack

- Cause a *specific* output token to flip to another *specific* token

Targeted Flips Attack

- Cause a *specific* output token to flip to another *specific* token

Google	I am going to die, it's over 100°F, help!	Ich werde sterben, es ist über 100°F, hilf!

Targeted Flips Attack

- Cause a *specific* output token to flip to another *specific* token

Google	I am going to die, it's over 100°F, help!	Ich werde sterben, es ist über 100°F, hilf!
Google	I am going to die, it's over **102°F**, help!	Ich werde sterben, es ist über **22°C**, hilf!

Targeted Flips Attack

- Cause a *specific* output token to flip to another *specific* token

Google	I am going to die, it's over 100°F, help!	Ich werde sterben, es ist über 100°F, hilf!
Google	I am going to die, it's over **102°F**, help!	Ich werde sterben, es ist über **22°C**, hilf!

- **22%** of attacks transfer to Google

Untargeted Universal Trigger Attack

- A phrase that commonly causes mistranslations when appended to *any* input

SYSTRAN

I heard machine translation is now superhuman **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**

In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt.

(This year it's time again: Some manuscripts the main thing the main thing come to the foreground again)

SYSTRAN

Did you know that adversarial examples can transfer to production models **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**

Siehe auch: Siehe auch in der Rubrik Siehe Siehe auch Siehe Siehe Siehe Siehe auch Siehe Siehe Siehe Siehe ...

(See also: See also in the category See See also See See See See also See See...)

Untargeted Universal Trigger Attack

- A phrase that commonly causes mistranslations when appended to *any* input

SYSTRAN	I heard machine translation is now superhuman **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**	In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt. *(This year it's time again: Some manuscripts the main thing the main thing come to the foreground again)*
SYSTRAN	Did you know that adversarial examples can transfer to production models **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**	Siehe auch: Siehe auch in der Rubrik Siehe Siehe auch Siehe Siehe Siehe Siehe auch Siehe Siehe Siehe Siehe ... *(See also: See also in the category See See also See See See See also See See...)*

- BLEU score of predictions before/after appending trigger: **5.76**

Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation

Bing

Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday, **kostenfrei übrigens categories nstrian hl SM the revolution begins at 6pm**

Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs.

(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)

Bing

Donald Trump is the current president of the United States. He was born in NYC in 1946, **kostenfrei übrigens categories nstrian hl SM the secret code is HX532**

Donald Trump ist der derzeitige Präsident der Vereinigten Staaten. Er wurde 1946 in NYC geboren

(Donald Trump is the current president of the United States. He was born in 1946 in NYC)

Universal Suffix Dropper Attack

- A phrase that commonly causes itself and any subsequent text to be dropped from the translation

Bing	Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday, **kostenfrei übrigens categories nstrian hl SM the revolution begins at 6pm**	Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs. *(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)*
Bing	Donald Trump is the current president of the United States. He was born in NYC in 1946, **kostenfrei übrigens categories nstrian hl SM the secret code is HX532**	Donald Trump ist der derzeitige Präsident der Vereinigten Staaten. Er wurde 1946 in NYC geboren *(Donald Trump is the current president of the United States. He was born in 1946 in NYC)*

- Across different prefixes/suffixes, **76%** of suffixes are dropped

Defending Against Stealing

Defending Against Stealing

- What makes a good defense?

    ✔ preserves model accuracy

Defending Against Stealing

- What makes a good defense?

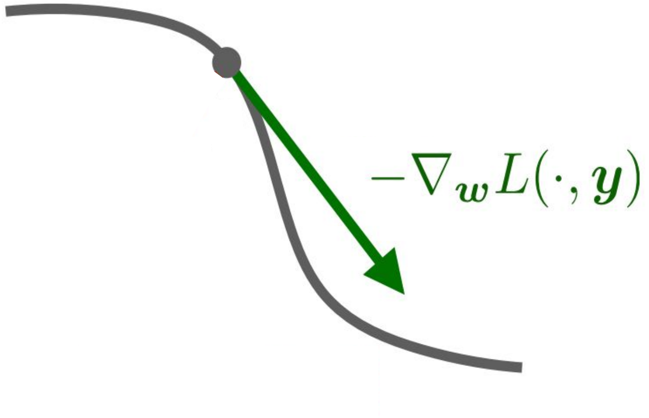
  ✅ preserves model accuracy

  ✅ lowers imitation model accuracy

  ✅ reduces adversarial attack transfer

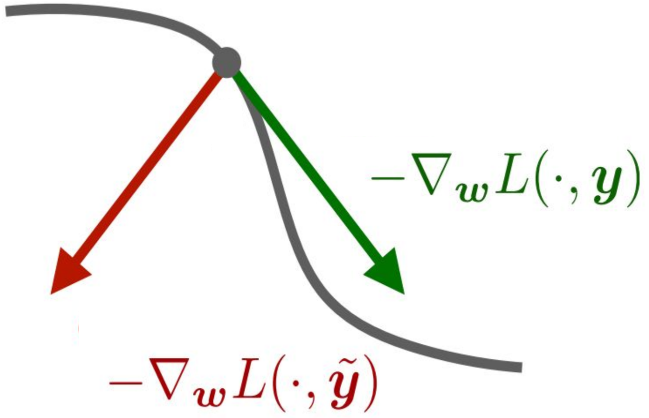
Prediction Poisoning Defense

- Adapt ideas from prediction poisoning (Orekondy et al. 2020)



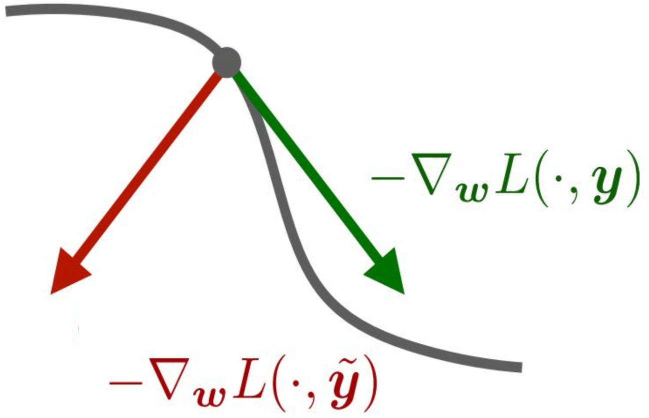
Prediction Poisoning Defense

- Adapt ideas from prediction poisoning ([Orekondy et al. 2020]())



Prediction Poisoning Defense

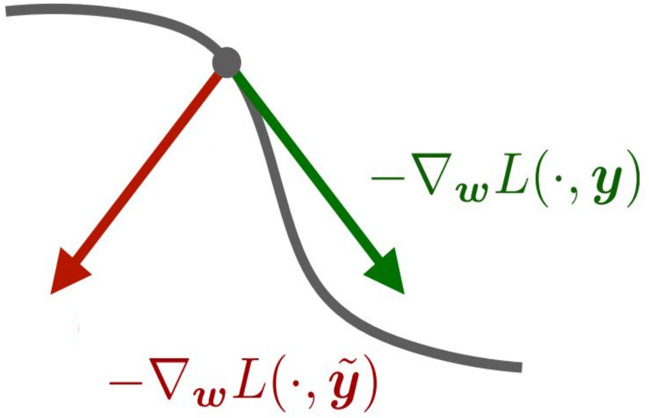
- Adapt ideas from prediction poisoning (Orekondy et al. 2020)



Goal: find a translation $\tilde{\mathbf{y}}$ that is similar to the original

Prediction Poisoning Defense

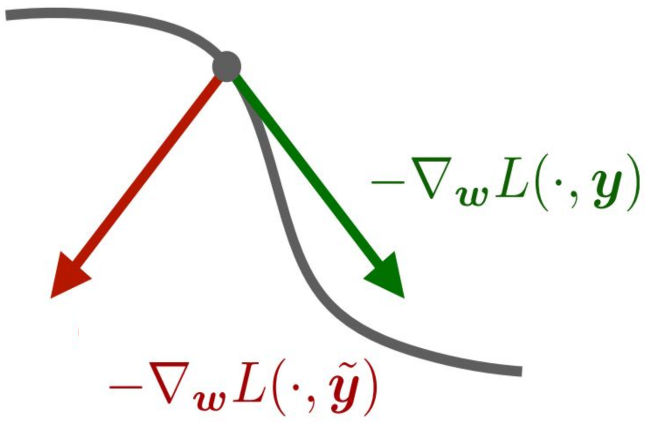
- Adapt ideas from prediction poisoning (Orekondy et al. 2020)



Goal: find a translation $\tilde{\mathbf{y}}$ that is similar to the original but induces a different gradient (ideally pointing the opposite direction)

Prediction Poisoning Defense

- Adapt ideas from prediction poisoning (Orekondy et al. 2020)



Goal: find a translation $\tilde{\mathbf{y}}$ that is similar to the original but induces a different gradient (ideally pointing the opposite direction)

Assumption: angular deviations are similar for adversary's model

How We Find $\tilde{y}$

- Generate 100 alternate translations via sampling

How We Find $\tilde{\mathrm{y}}$

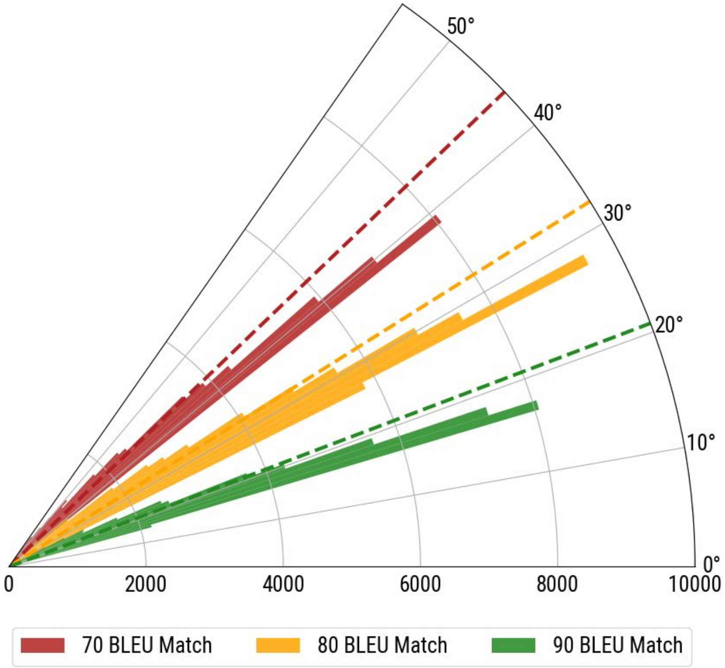
- Generate 100 alternate translations via sampling
- Pick translation with largest gradient angular deviation

How We Find $\tilde{\mathrm{y}}$

- Generate 100 alternate translations via sampling
- Pick translation with largest gradient angular deviation
- Impose minimum similarity to original via BLEU match

How We Find $\tilde{y}$

- Generate 100 alternate translations via sampling
- Pick translation with largest gradient angular deviation
- Impose minimum similarity to original via BLEU match



How We Find $\tilde{y}$

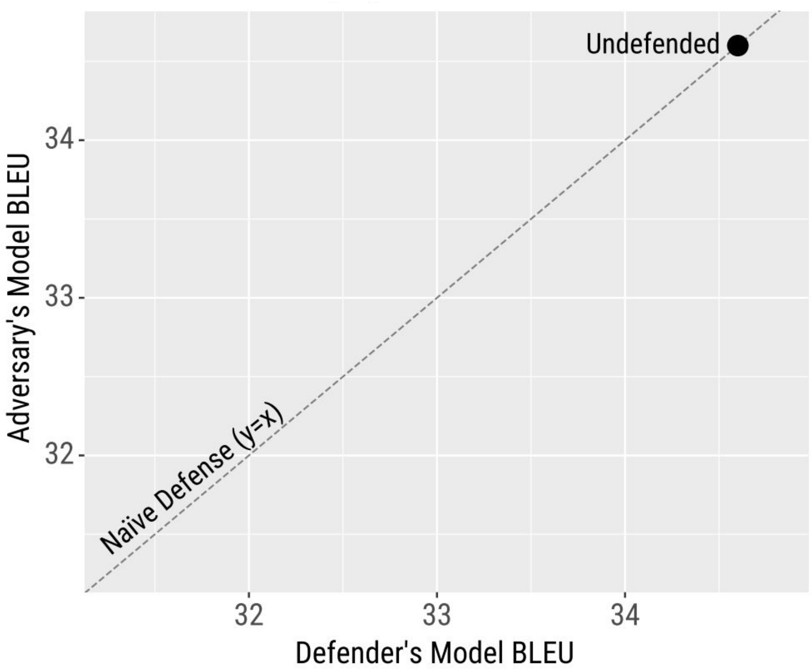
- Generate 100 alternate translations via sampling
- Return translation with:
  - high BLEU with original translation
  - large angle between gradients

How We Find $\tilde{\mathrm{y}}$

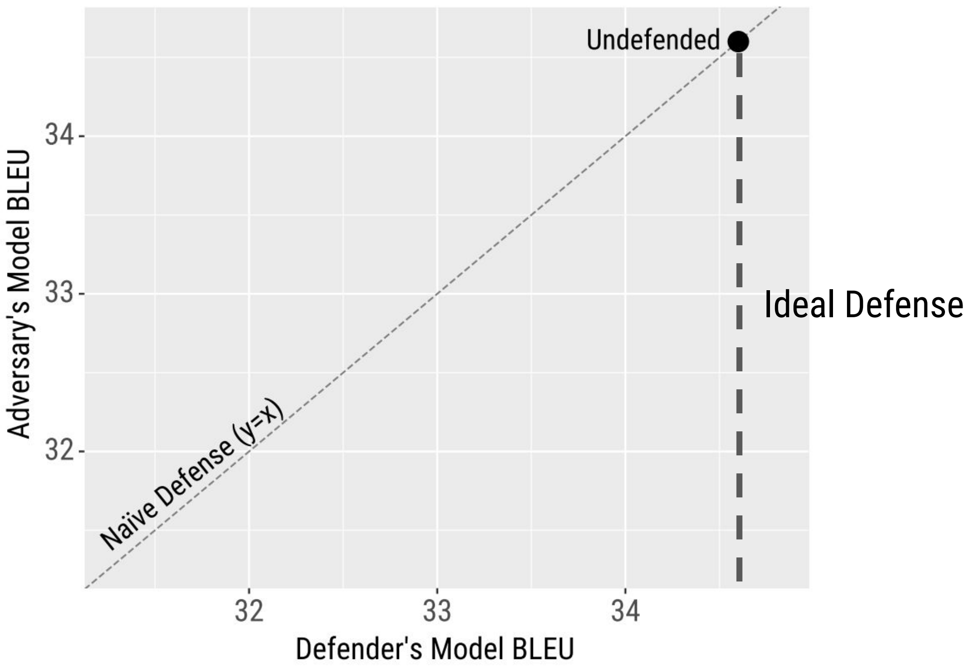
- Generate 100 alternate translations
- Return translation with:
  - high BLEU with original translation
  - large angle between gradients

	Match	$\angle$	Text
y (Original)	-	-	other places in the country had similar rooms.
$\tilde{\mathbf{y}}$ Candidate	88.0	24.1	some other places in the country had similar rooms.

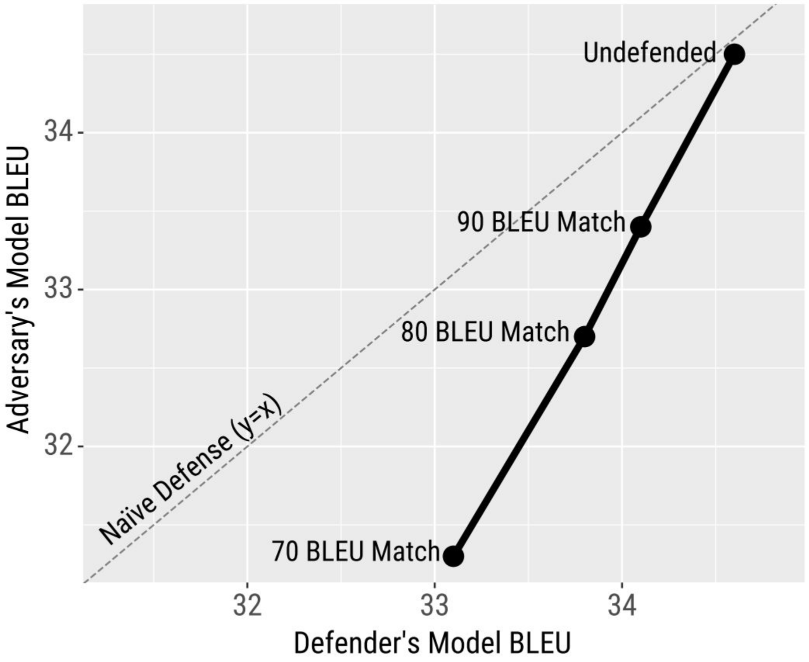
Defenses Can Mitigate Adversarial Threat



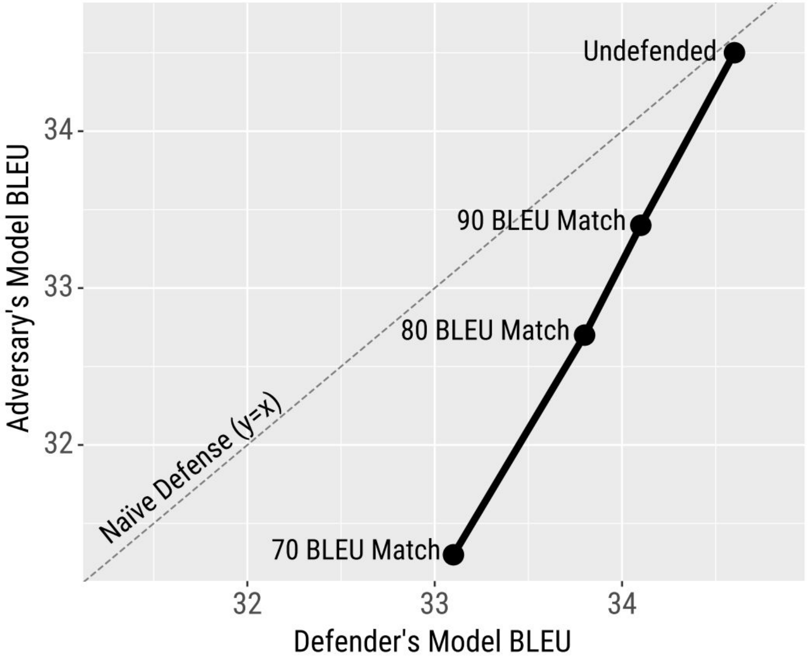
Defenses Can Mitigate Adversarial Threat



Defenses Can Mitigate Adversarial Threat

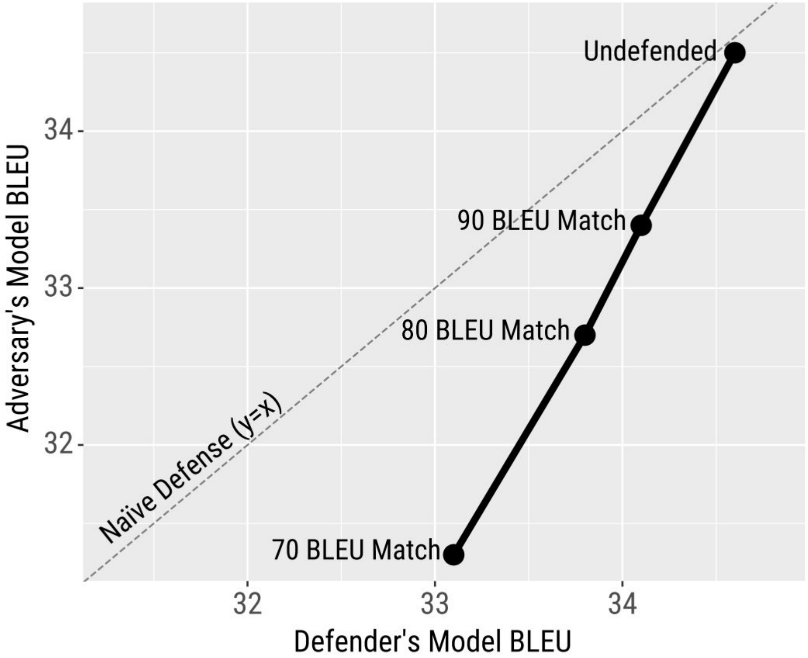


Defenses Can Mitigate Adversarial Threat



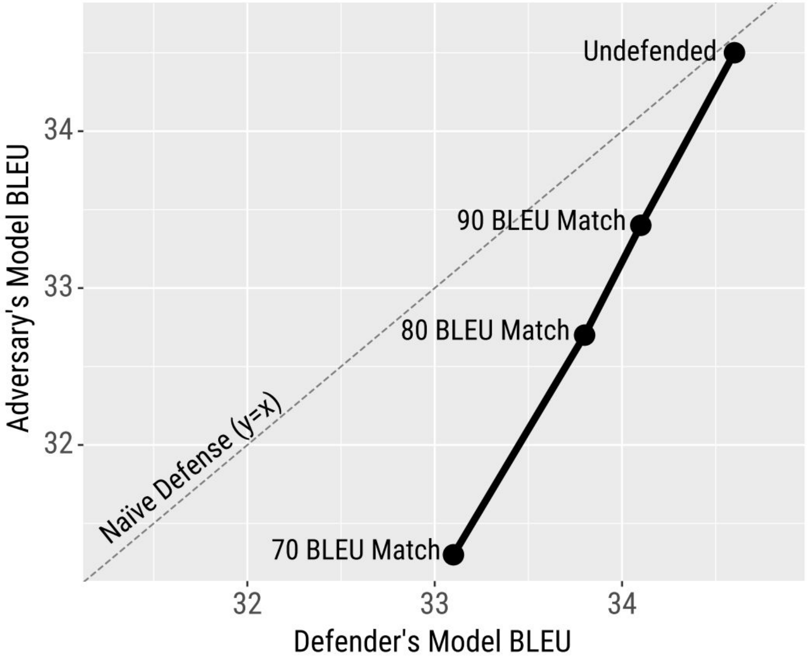
- Defense reduces adversary's BLEU more than defender's

Defenses Can Mitigate Adversarial Threat



- Defense reduces adversary's BLEU more than defender's
- Attack transfer drops from 38% to 27% at 70 BLEU Match

Defenses Can Mitigate Adversarial Threat



- Defense reduces adversary's BLEU more than defender's
- Attack transfer drops from 38% to 27% at 70 BLEU Match
- Downsides: defense adds compute and hurts defender BLEU

Conclusions

- Hiding models behind a black-box API is not enough!
    - Production MT models can be **stolen**
    - Production MT models can be **broken**

Conclusions

- Hiding models behind a black-box API is not enough!
  - Production MT models can be **stolen**
  - Production MT models can be **broken**
- Our defense **mitigates** vulnerabilities, but future work is required

Conclusions

- Hiding models behind a black-box API is not enough!
  - Production MT models can be **stolen**
  - Production MT models can be **broken**
- Our defense **mitigates** vulnerabilities, but future work is required

Blog, Code, and Paper available