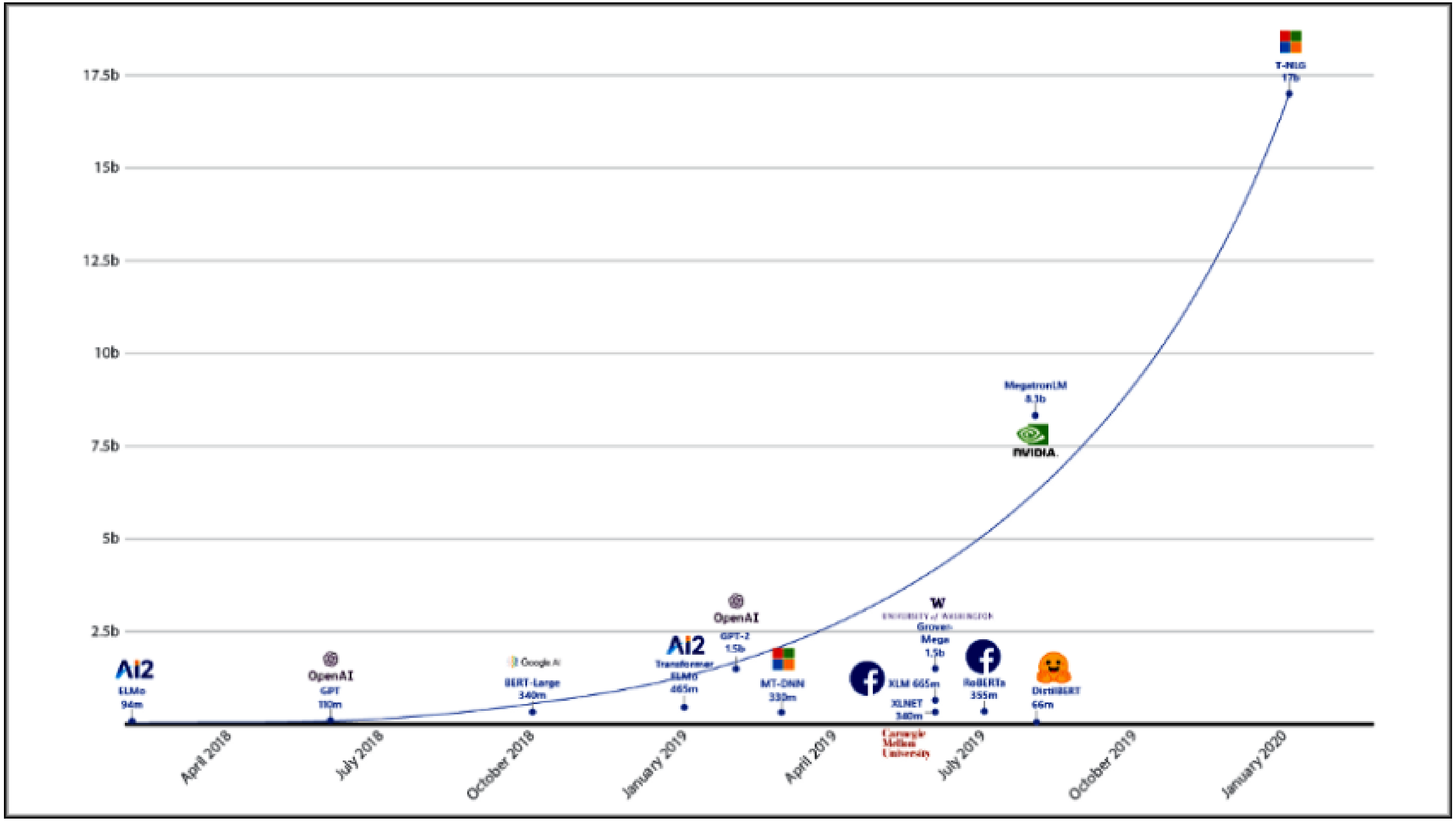
Robustness to Adversarial Attacks & Scaling Law

Presented by: Rodwell Nicolas Bent
Tom Marty
Rafael Hernandez
Siddhika Arunachalam

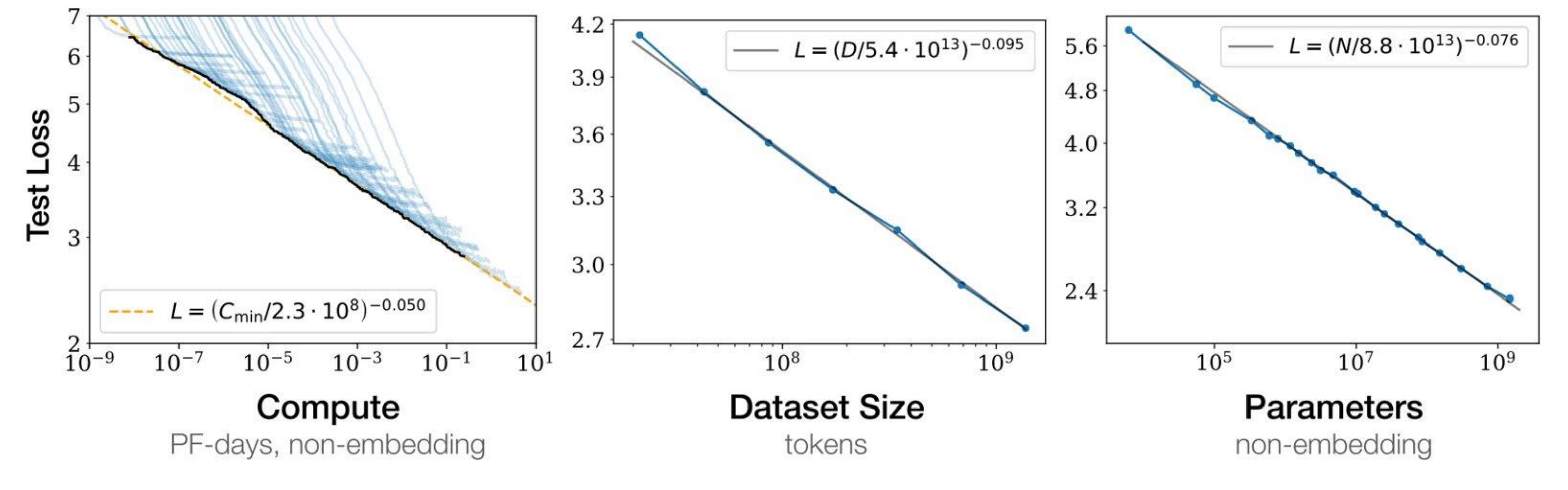
How does scale impact adversarial robustness?

What is scale ?

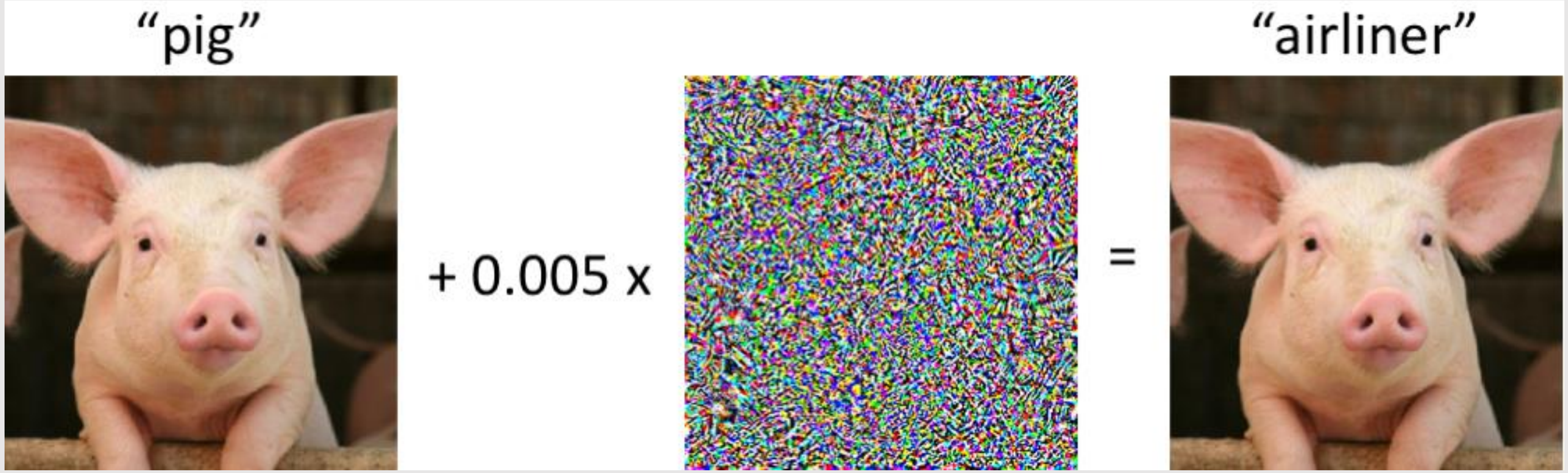


Language models represented by the size of parameters (figure adopted from DistilBERT from huggingface)

How does scale impact models?



Robustness and adversarial attacks



Robustness and adversarial attacks

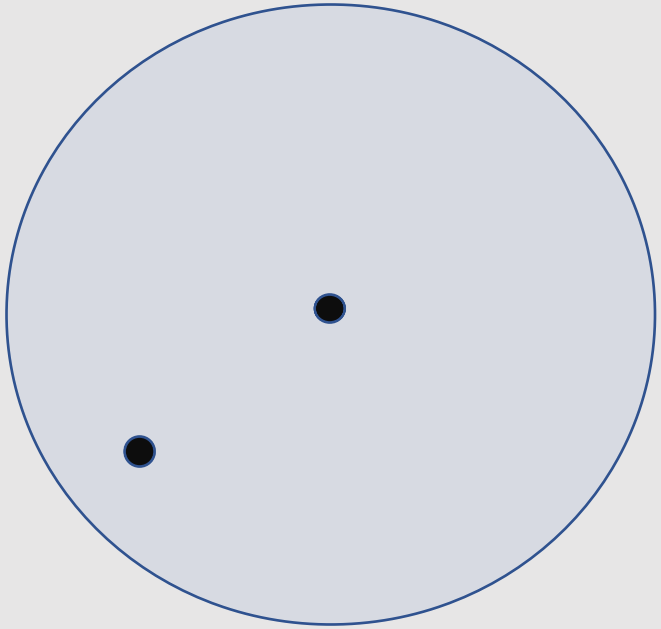
Original

Perfect performance by the movie legend: Positive (99%)

Adversarial

Spotless performance by the movie legend: Negative (74%)

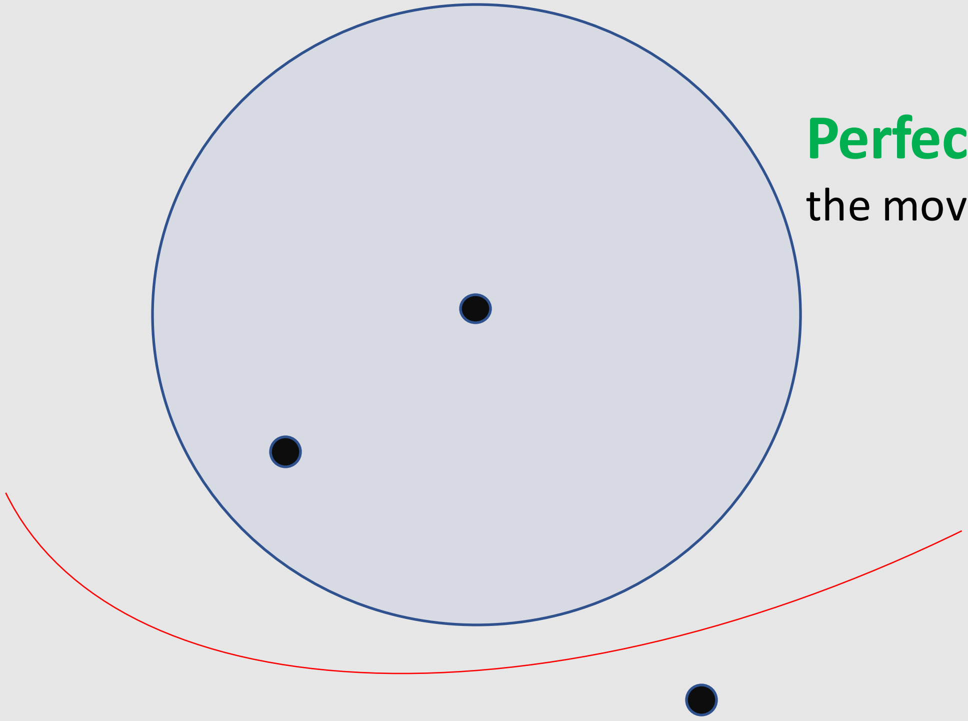
Robustness and adversarial attacks



Perfect performance by the movie legend

Great performance by the movie legend

Robustness and adversarial attacks



Perfect performance by the movie legend

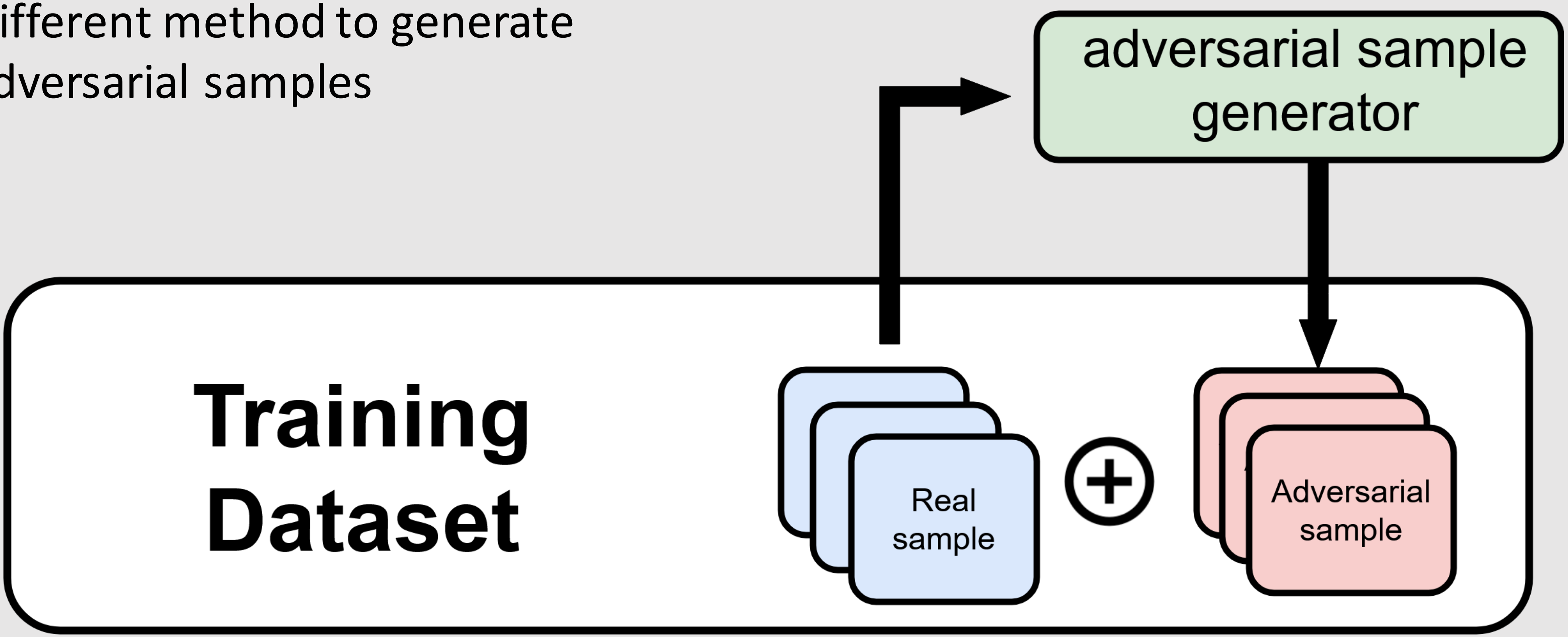
Spotless performance by the movie legend

Adversarial training

Adversarial attack as a form of data augmentation :

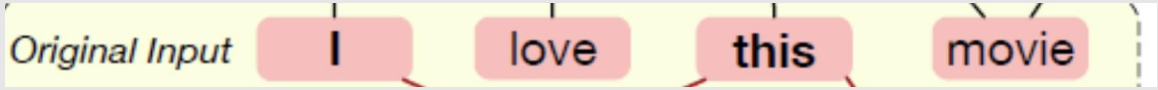
- Proportion of adversarial samples seen during training

- Different method to generate adversarial samples



Adversarial generator, an example :

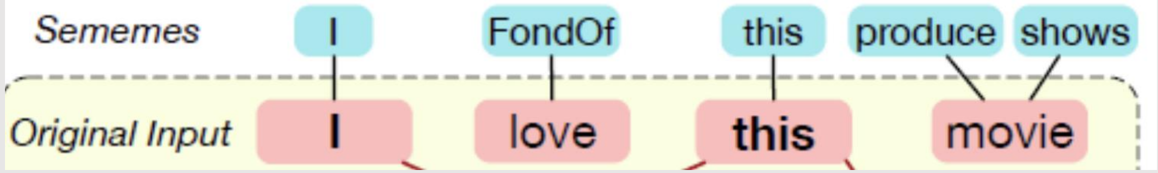
PSOZang :Word-level Textual Adversarial Attacking as Combinatorial Optimization :



https://arxiv.org/abs/1910.12196

Adversarial generator, an example :

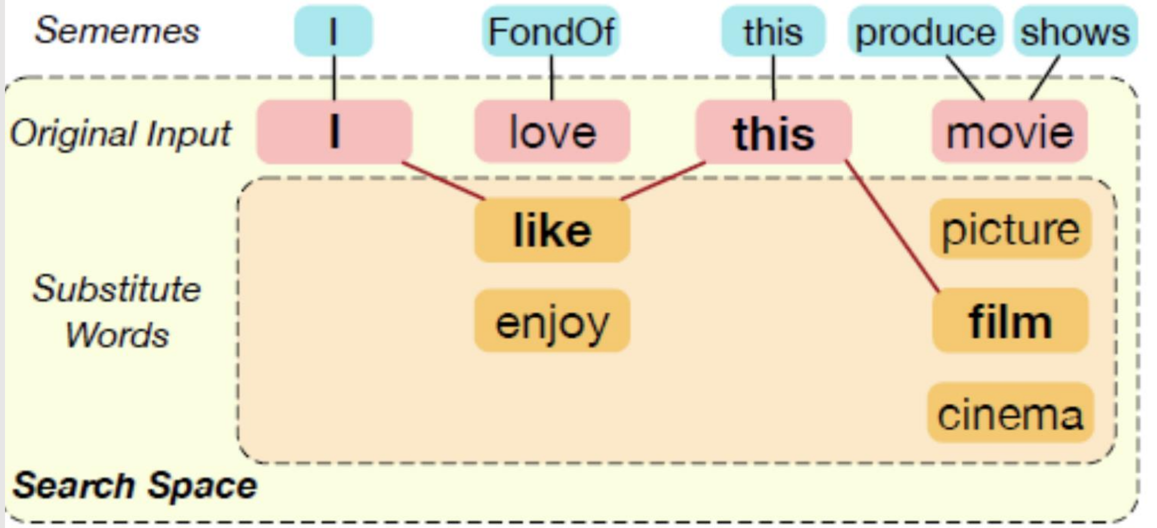
PSOZang :Word-level Textual Adversarial Attacking as Combinatorial Optimization :



https://arxiv.org/abs/1910.12196

Adversarial generator, an example :

PSOZang :Word-level Textual Adversarial Attacking as Combinatorial Optimization :

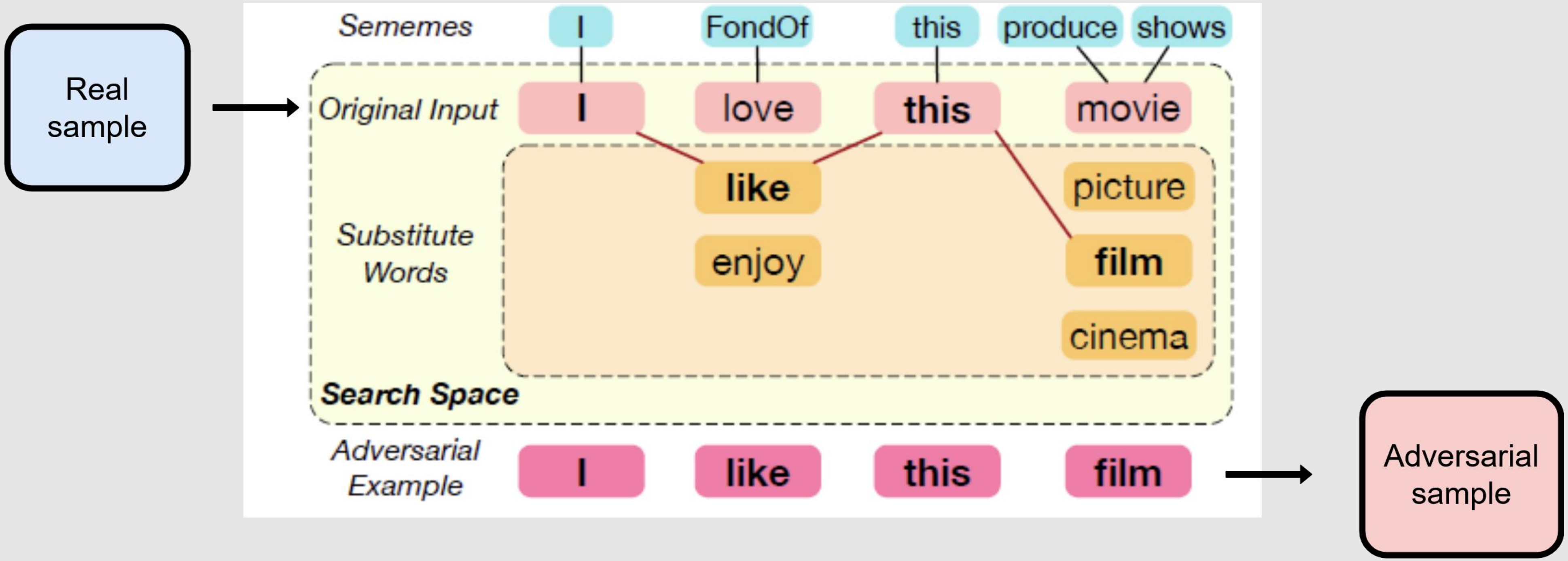


https://arxiv.org/abs/1910.12196

Adversarial generator, an example :

PSOZang :Word-level Textual Adversarial Attacking as Combinatorial Optimization :

Idea : Word-level attack seen as a Combinatorial Optimization problem – Particle Swarm



Real sample

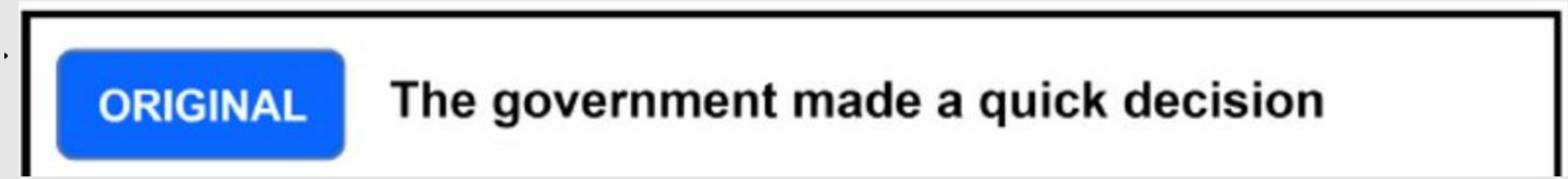
Adversarial sample

https://arxiv.org/abs/1910.12196

Adversarial generator, an example :

BAEGarg: BERT-based Adversarial Examples for Text Classification :

Idea : use BERT model to predict MASKED token



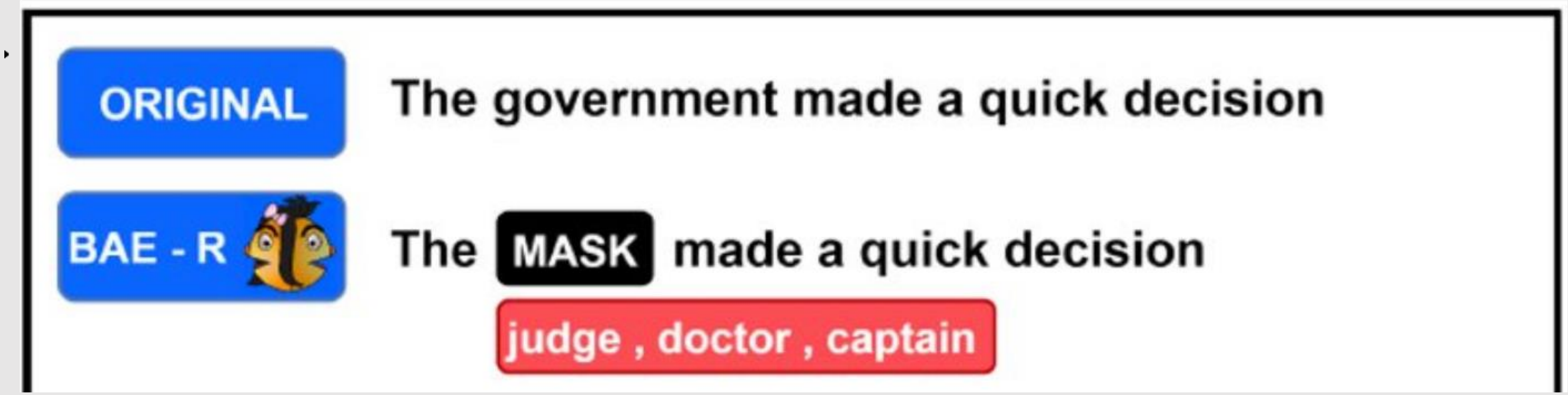
The government made a quick decision

https://arxiv.org/abs/1910.12196

Adversarial generator, an example :

BAEGarg: BERT-based Adversarial Examples for Text Classification :

Idea : use BERT model to predict MASKED token



The government made a quick decision

The MASK made a quick decision

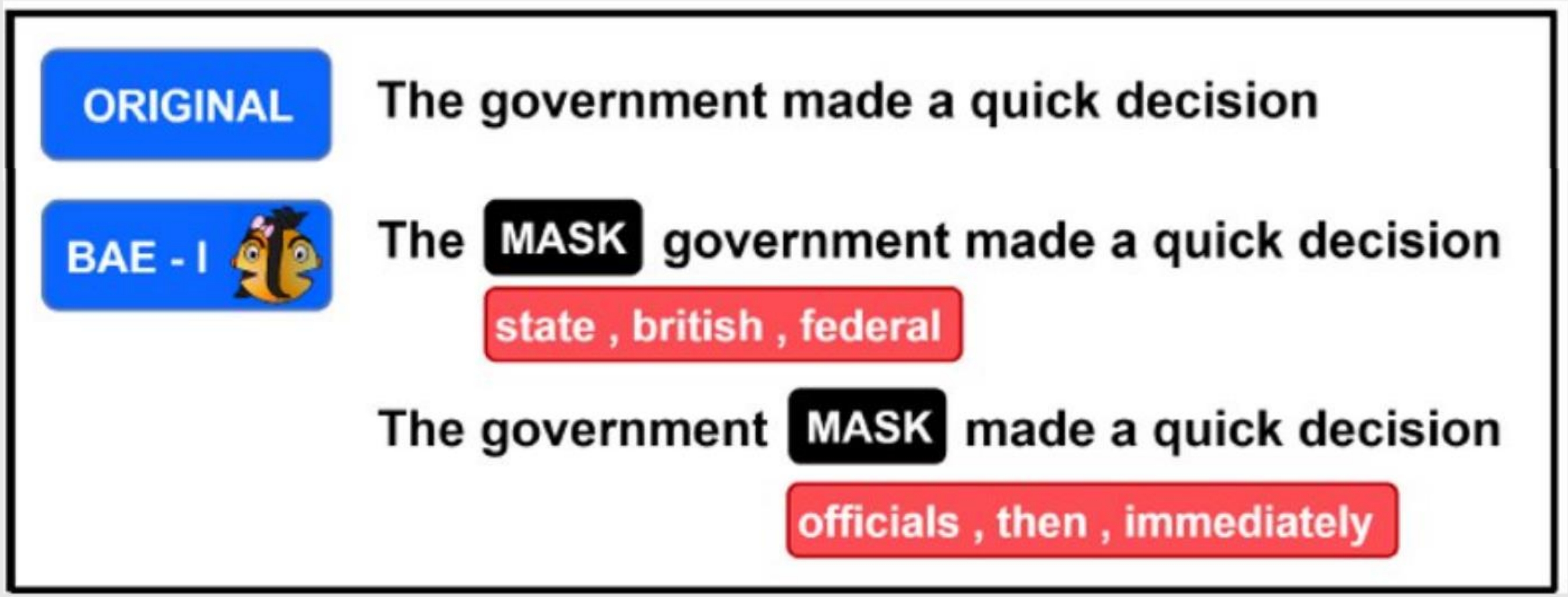
judge , doctor , captain

https://arxiv.org/abs/1910.12196

Adversarial generator, an example :

BAEGarg: BERT-based Adversarial Examples for Text Classification :

Idea : use BERT model to predict MASKED token



The government made a quick decision

The MASK government made a quick decision

state , british , federal

The government MASK made a quick decision

officials , then , immediately

https://arxiv.org/abs/1910.12196

How does scale impact adversarial robustness?

How does scale impact models?

Can we empirically verify that :

$$\boxed{\log(L) \approx a(p,d) \times \log(n) + b(p,d)}$$

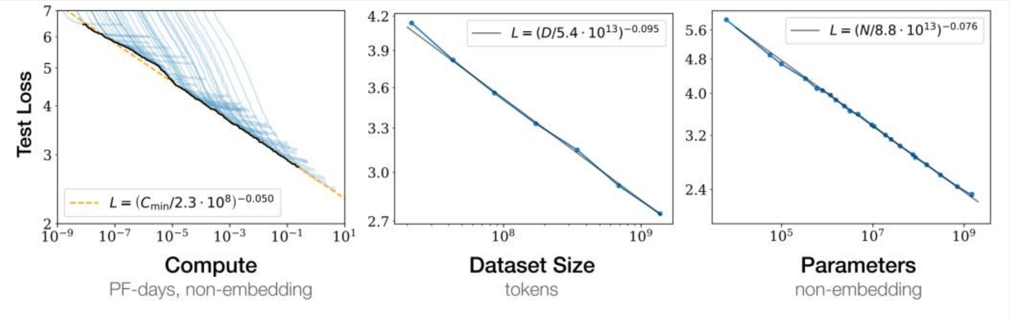
Where :

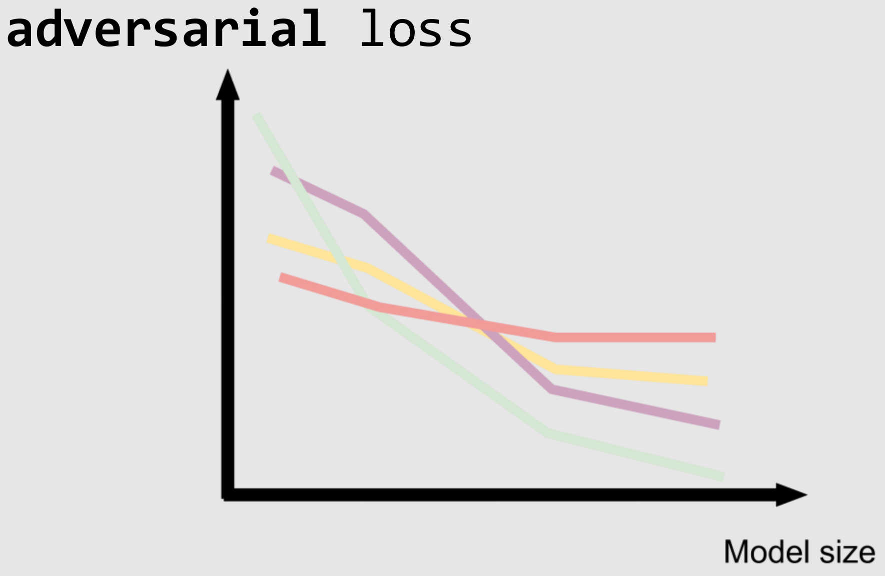
n = model size

p = number of adversarial samples seen during training

d = dataset size

L = **adversarial** loss



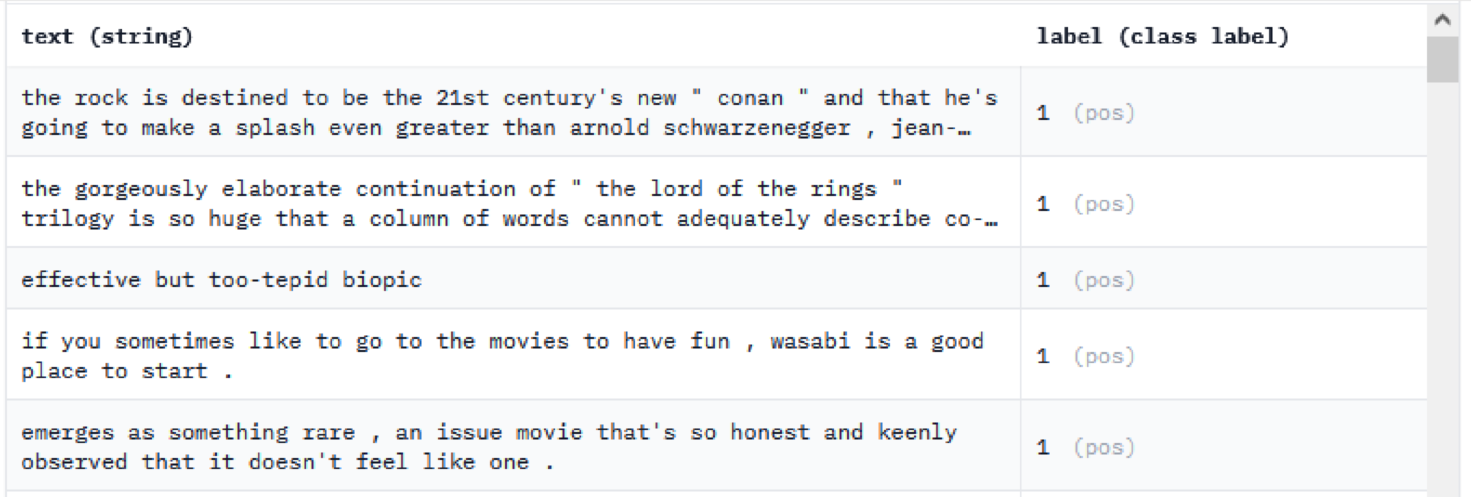


Adversarial Dataset

5k successfully attacked texts from all models

- The film is darkly **funny** in its observation of just how much more grueling and time-consuming the illusion of work is than actual work.
- The film is darkly **bizarro** in its observation of just how much more grueling and time-consuming the illusion of work is than actual work.

Dataset for sentiment categorization



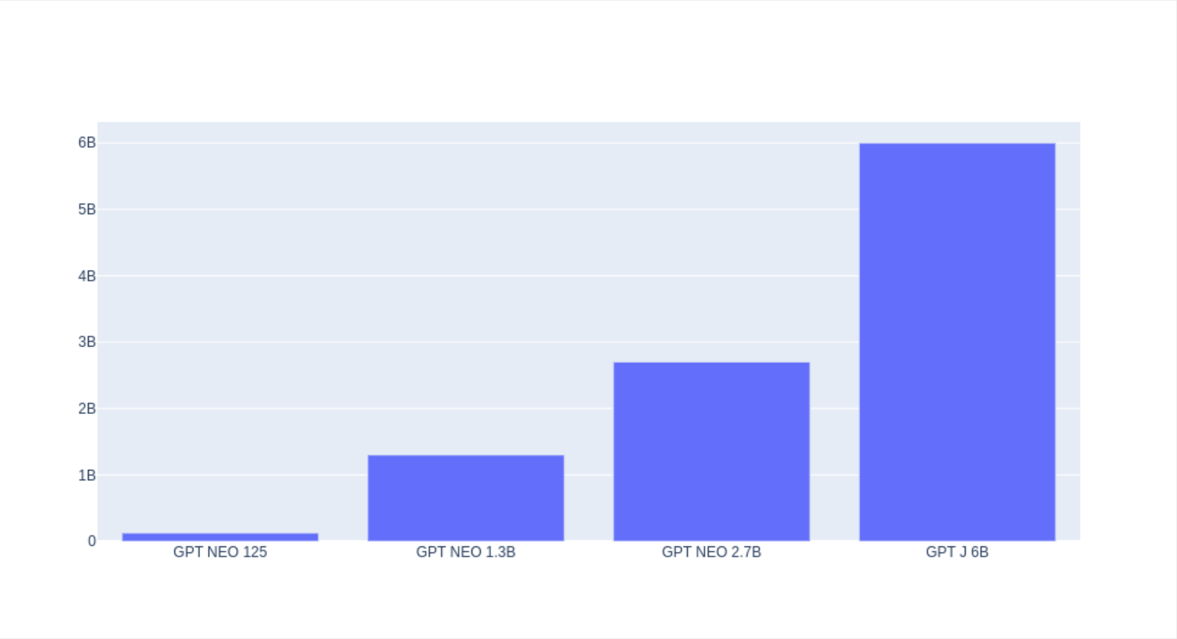
text (string)	label (class label)
the rock is destined to be the 21st century's new " conan " and that he's going to make a splash even greater than arnold schwarzenegger , jean-…	1 (pos)
the gorgeously elaborate continuation of " the lord of the rings " trilogy is so huge that a column of words cannot adequately describe co-…	1 (pos)
effective but too-tepid biopic	1 (pos)
if you sometimes like to go to the movies to have fun , wasabi is a good place to start .	1 (pos)
emerges as something rare , an issue movie that's so honest and keenly observed that it doesn't feel like one .	1 (pos)

- Rotten tomatoes : 10662 entries. (50% +, 50% -) : film commentary and review.
- IMDB : 50000 entries : film commentary and review.
- Amazon polarity : 2000000 entries : Comments on several products sold by amazon.

Eleuther AI GPT

- Neo 125M/1.3B/2.7B
- J - 6B

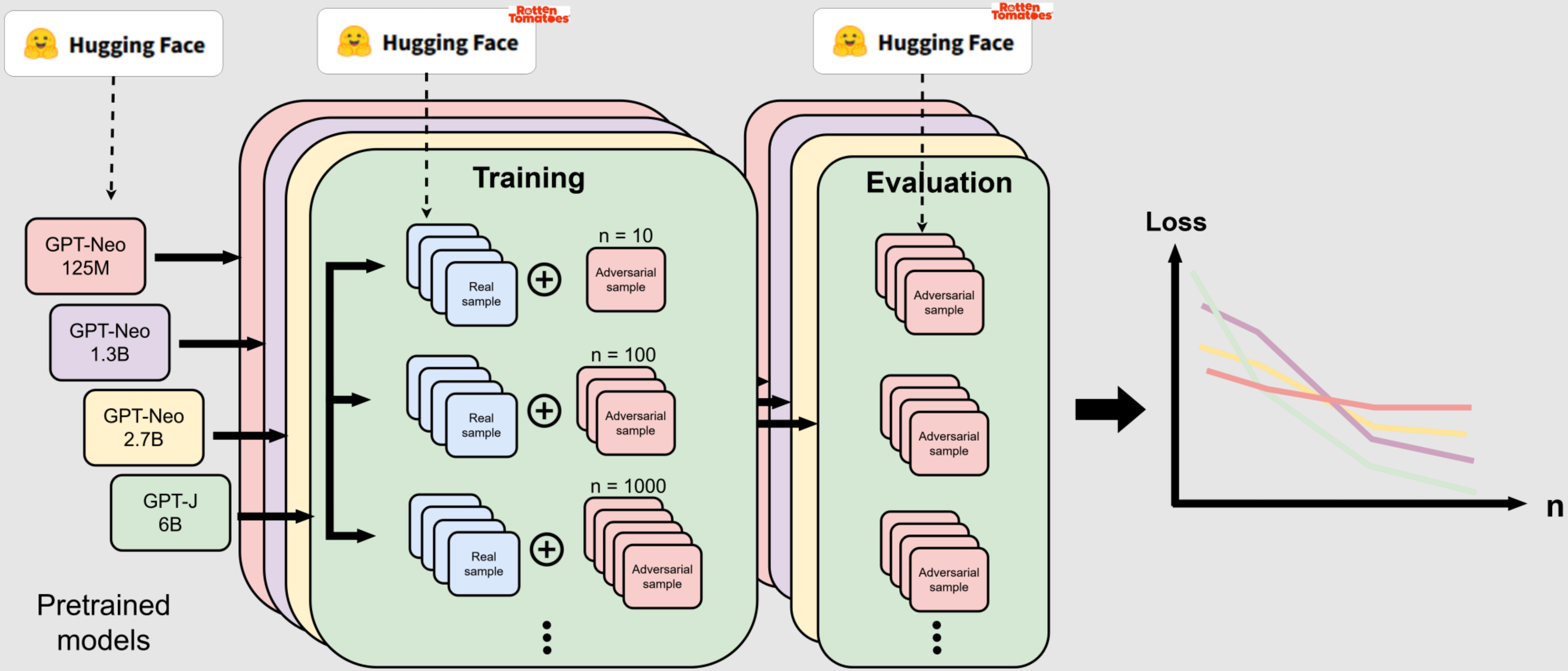
- 825 GiB training set (The Pile)



https://arxiv.org/abs/2005.05909

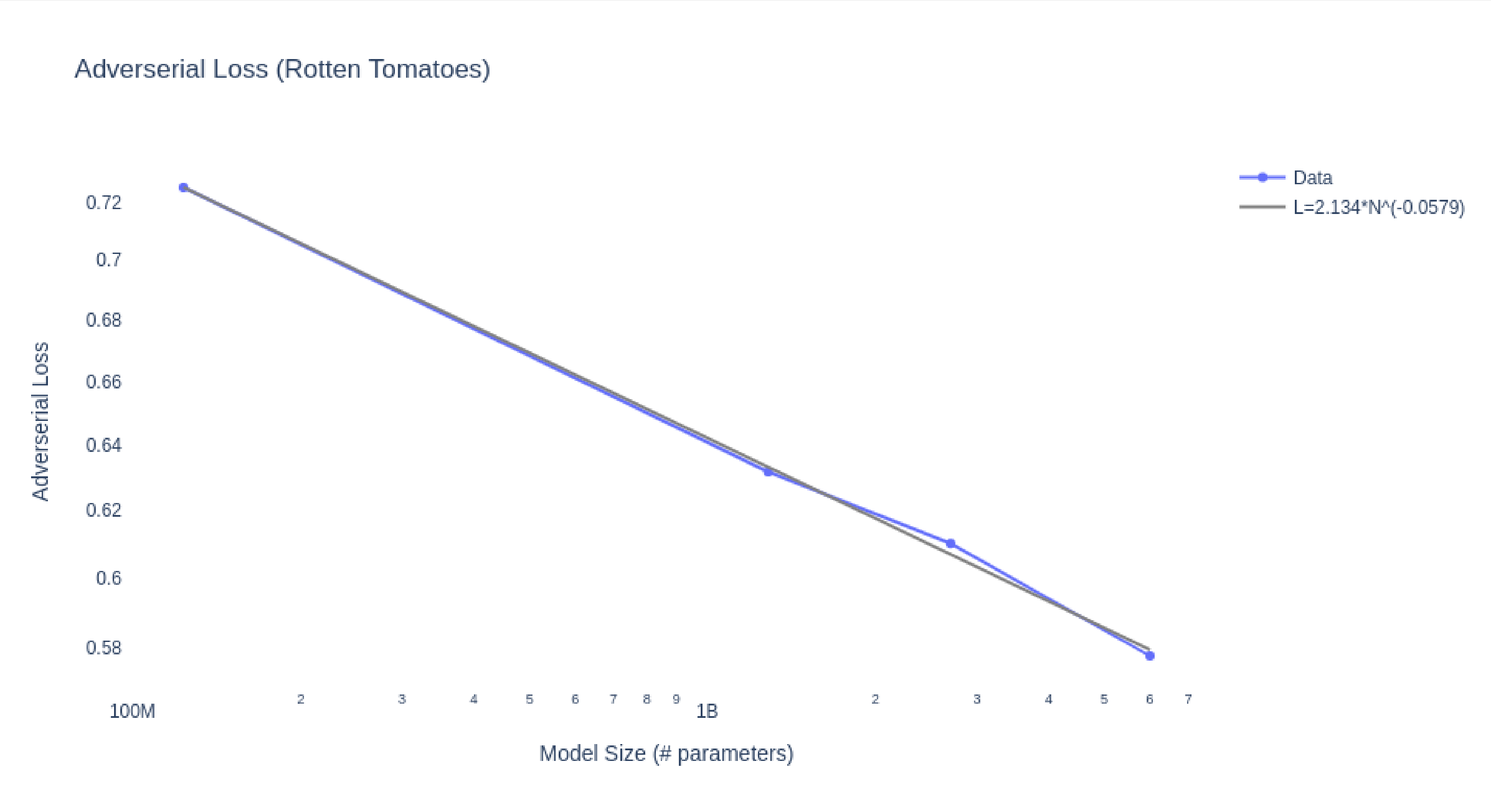
Training and Evaluation methods

- Idea : Evaluate loss / prediction performance for different models trained with increasing size of adversarial samples seen during training.

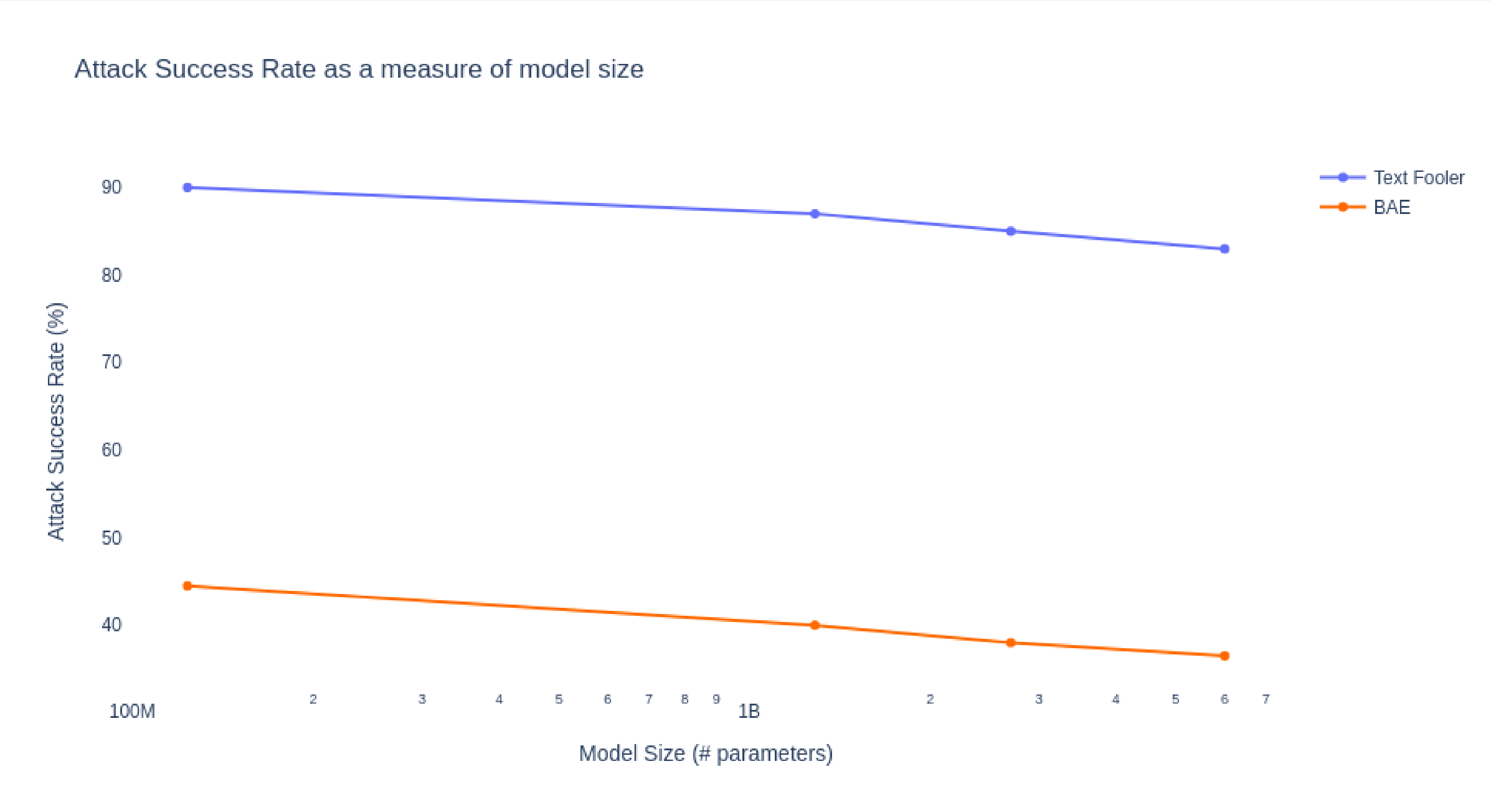


Results

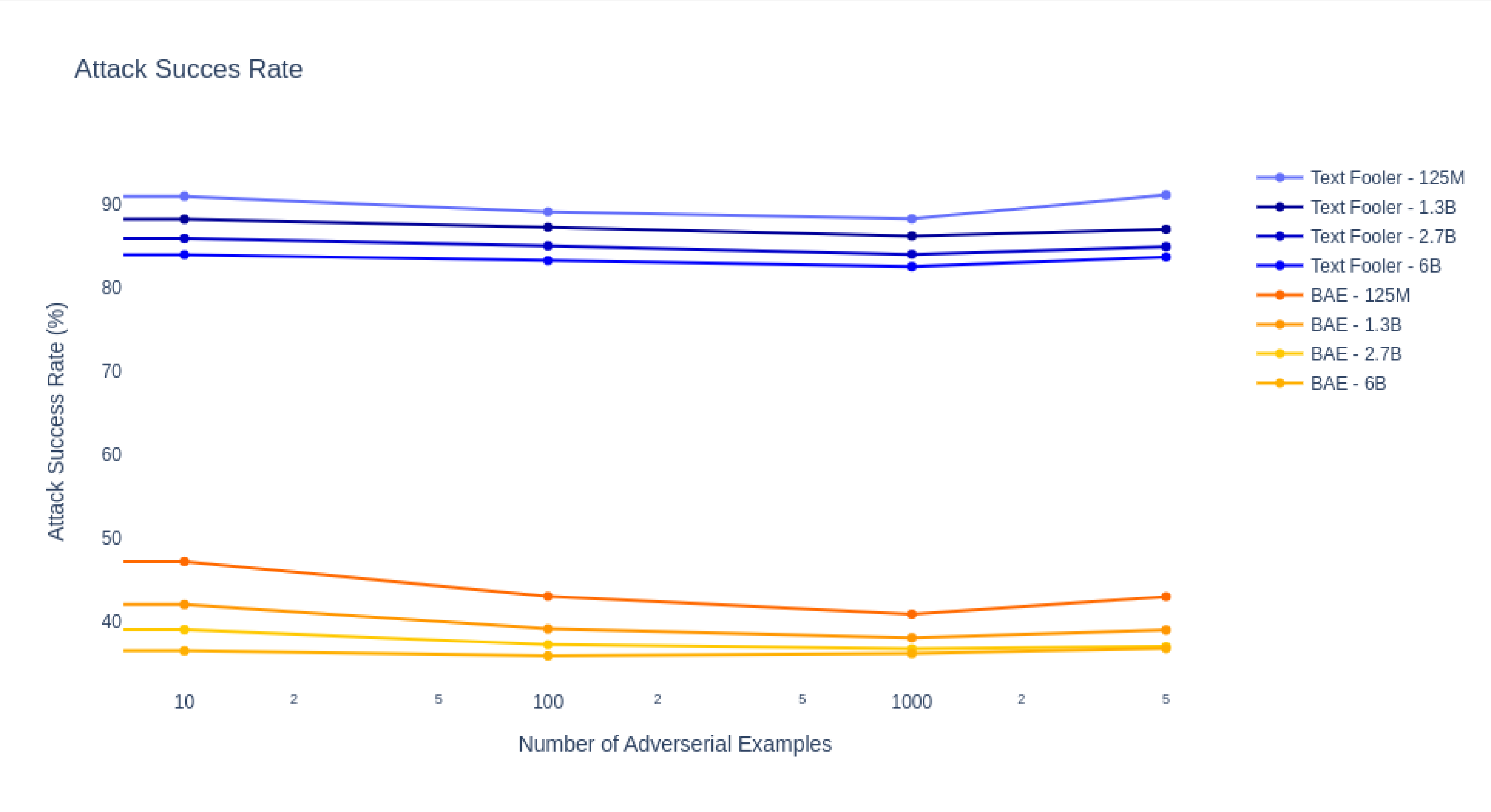
Results


Adverserial Loss (Rotten Tomatoes)
Data
L=2.134*N^(-0.0579)
Adverserial Loss
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
100M
2
3
4
5
6
7
8
9
1B
2
3
4
5
6
7
Model Size (# parameters)

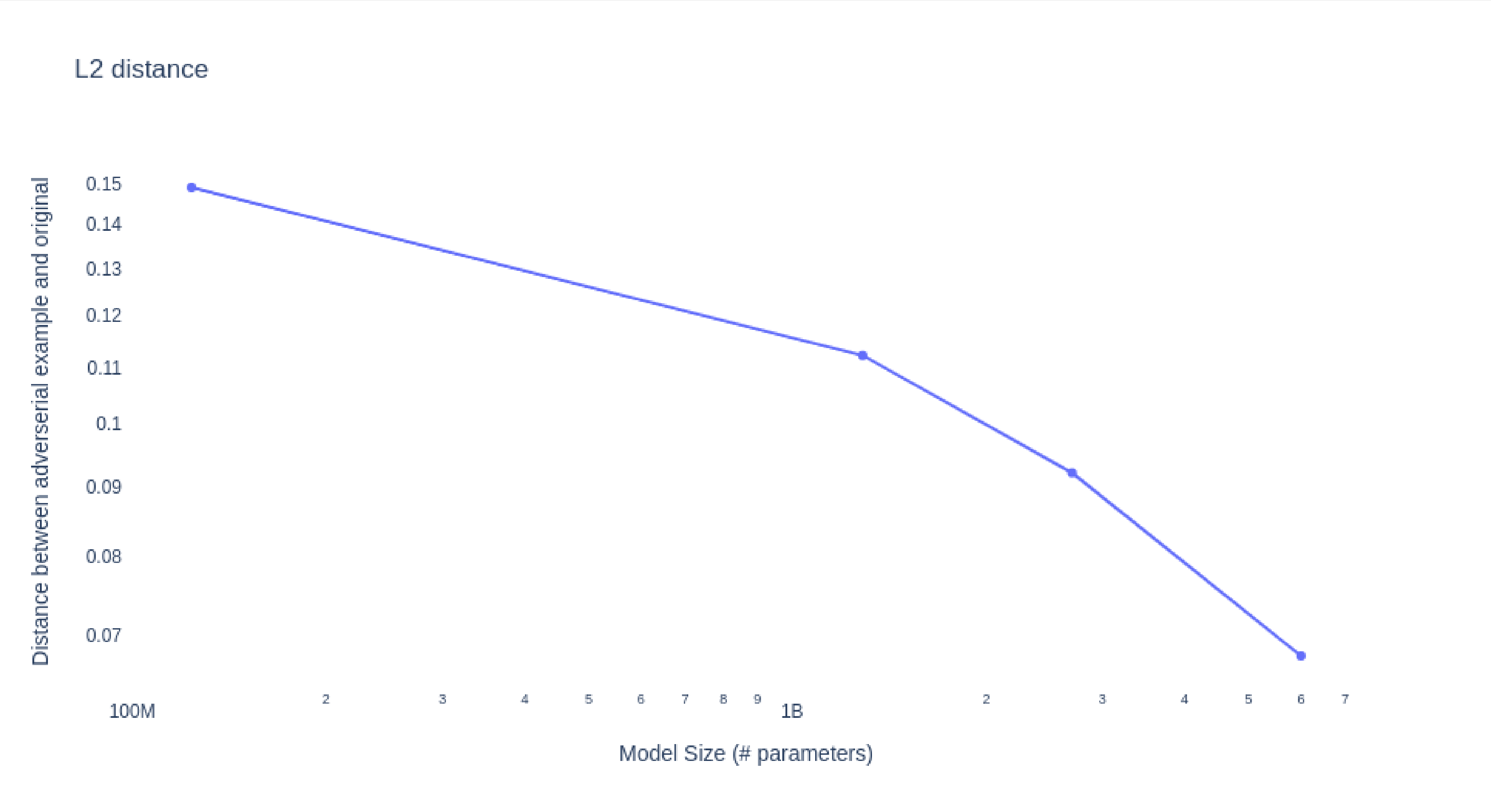
Results


Attack Success Rate as a measure of model size
Text Fooler
BAE
Attack Success Rate (%)
90
80
70
60
50
40
100M
2
3
4
5
6
7
8
9
1B
2
3
4
5
6
7
Model Size (# parameters)

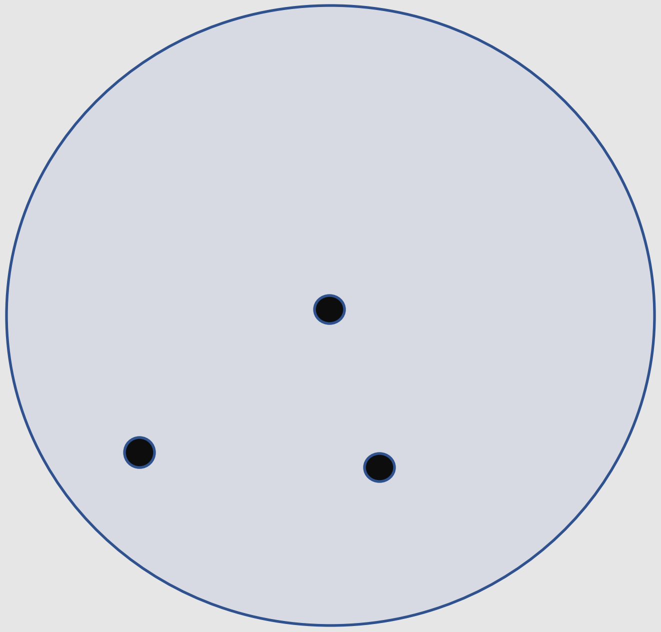
Results


Attack Succes Rate
Attack Success Rate (%)
90
80
70
60
50
40
10
2
5
100
2
5
1000
2
5
Number of Adverserial Examples
Text Fooler - 125M
Text Fooler - 1.3B
Text Fooler - 2.7B
Text Fooler - 6B
BAE - 125M
BAE - 1.3B
BAE - 2.7B
BAE - 6B

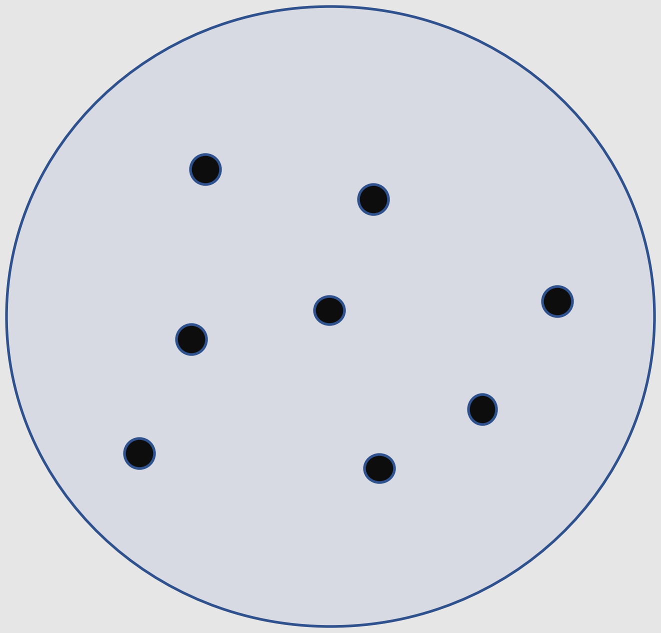
Results


L2 distance
Distance between adverserial example and original
0.15
0.14
0.13
0.12
0.11
0.1
0.09
0.08
0.07
100M
2
3
4
5
6
7
8
9
1B
2
3
4
5
6
7
Model Size (# parameters)

Density of ball



Density of ball



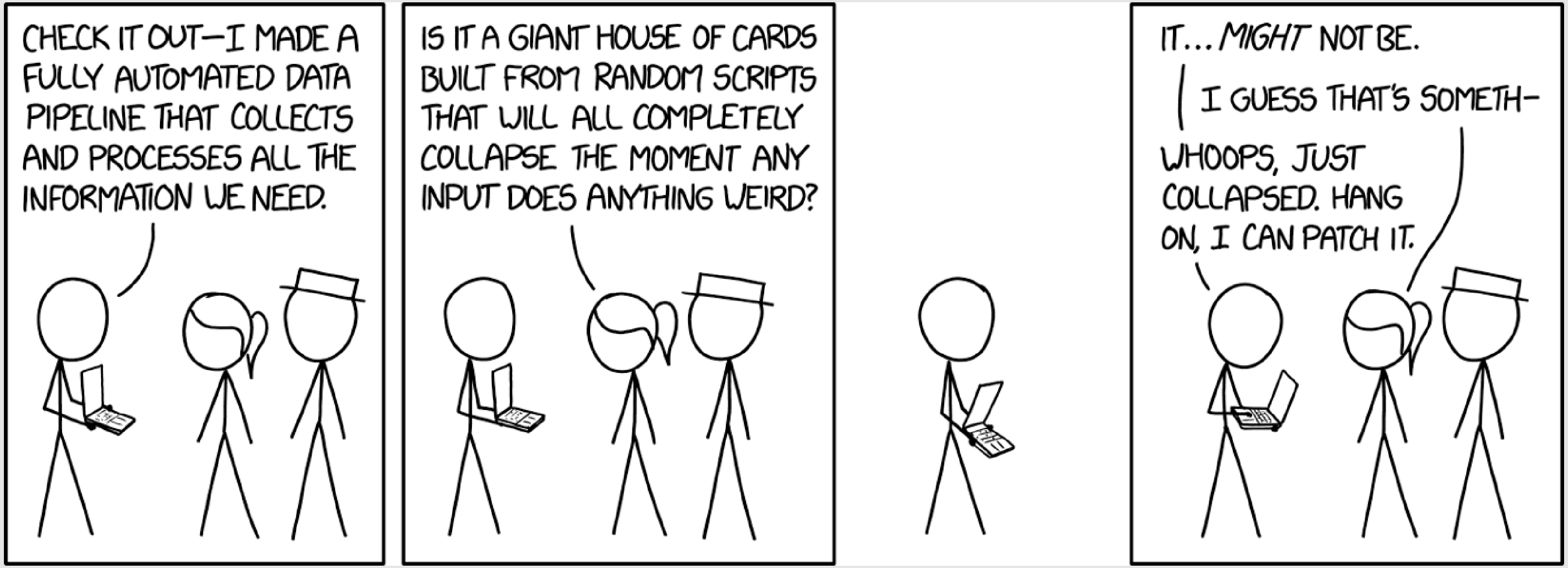
Limitations

- Positive (99%) --> Negative (82%)
- yeah , these flicks are just that damn **good** . isn't it great ?
- yeah , these flicks are just that damn **bad** . isn't it great ?

Conclusion

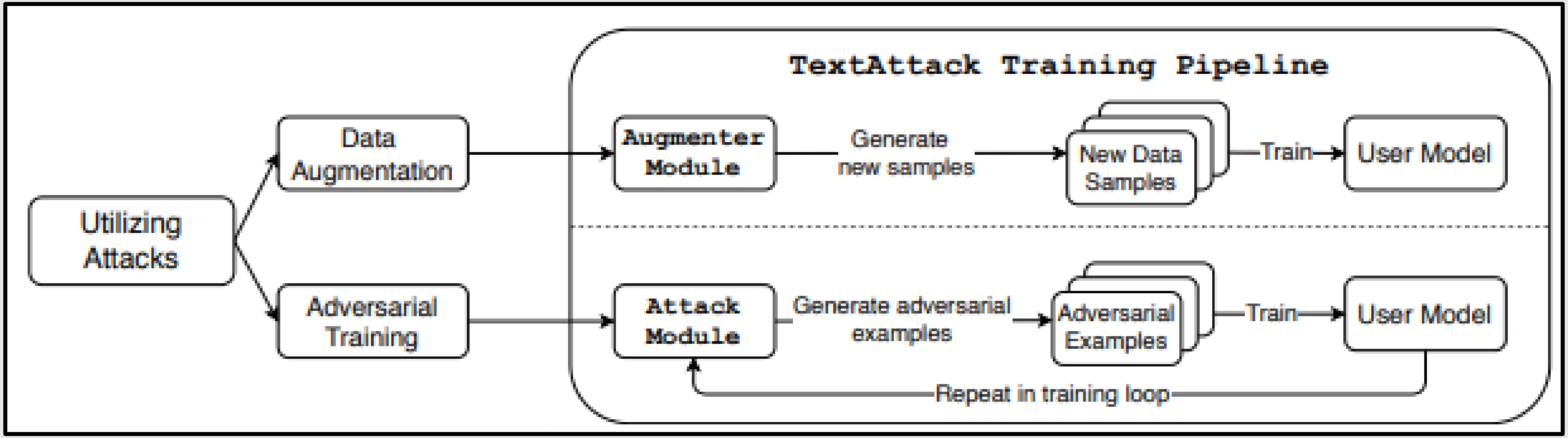
- Scale improves adversarial robustness
- Preliminary results
- More datasets, more models!

Thank You !
Questions ?


CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.
IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?
IT... *MIGHT* NOT BE.
I GUESS THAT'S SOMETH-
WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.

TextAttack

A simple framework for adversarial attacks :



https://arxiv.org/abs/2005.05909

Adversarial generator, an example :

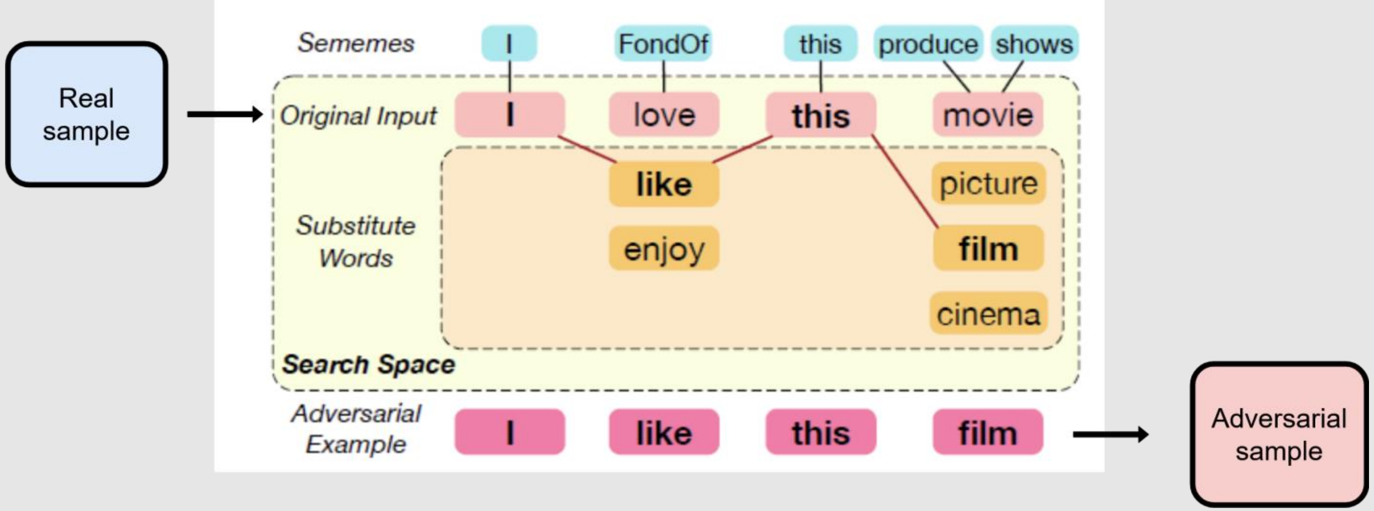
PSOZang :Word-level Textual Adversarial Attacking as Combinatorial Optimization :

Idea : Word-level attack seen as a Combinatorial Optimization problem

- Word substitution method based on sememes (unit of semantic meaning) : define the search space.

- Produces *likely* output (without context-awareness)

•Particle swarm optimization-based search algorithm



https://arxiv.org/abs/1910.12196

Adversarial generator, an example :

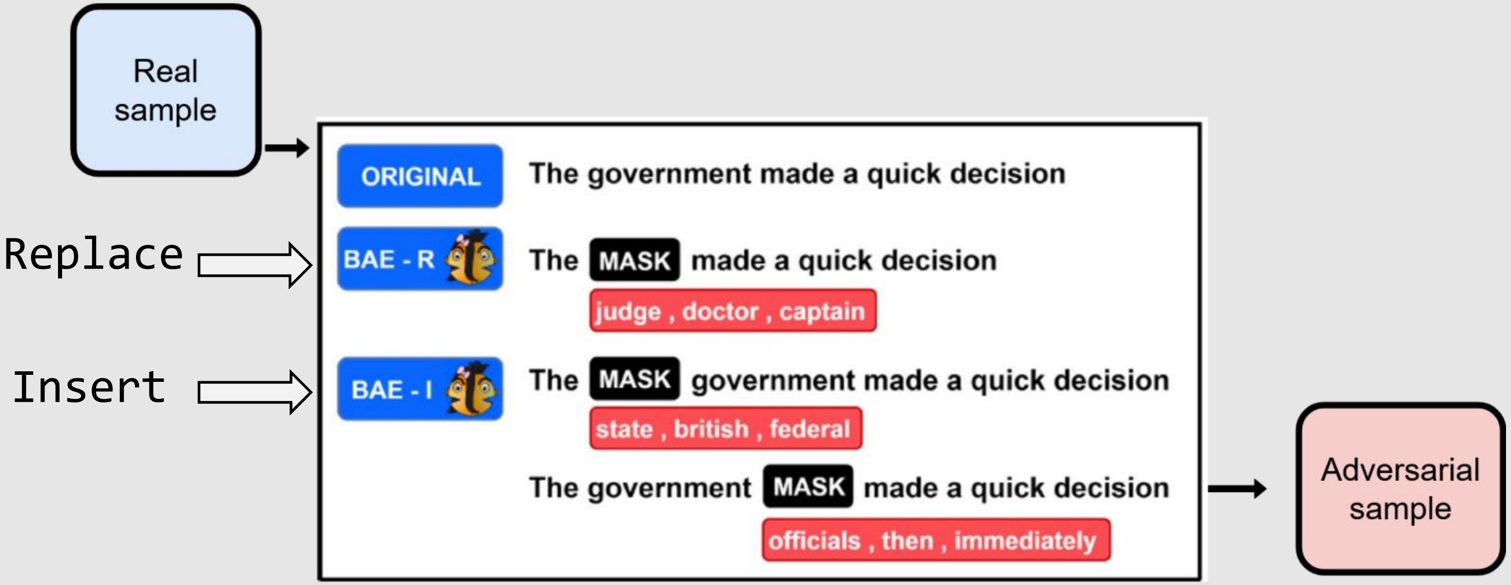
BAEGarg: BERT-based Adversarial Examples for Text Classification :

Idea : use BERT model to predict MASKED token

- Rule-based synonym replacement strategy

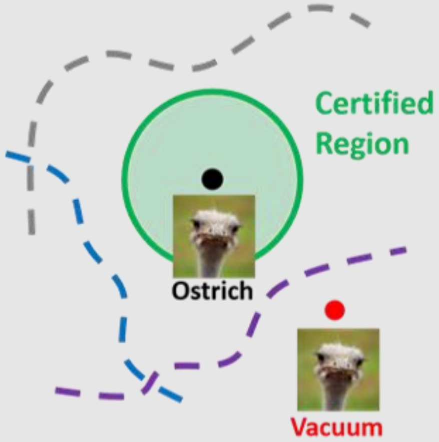
- Produces *likely* output (with context-awareness)

- Produces output with improved grammaticality and semantic coherence

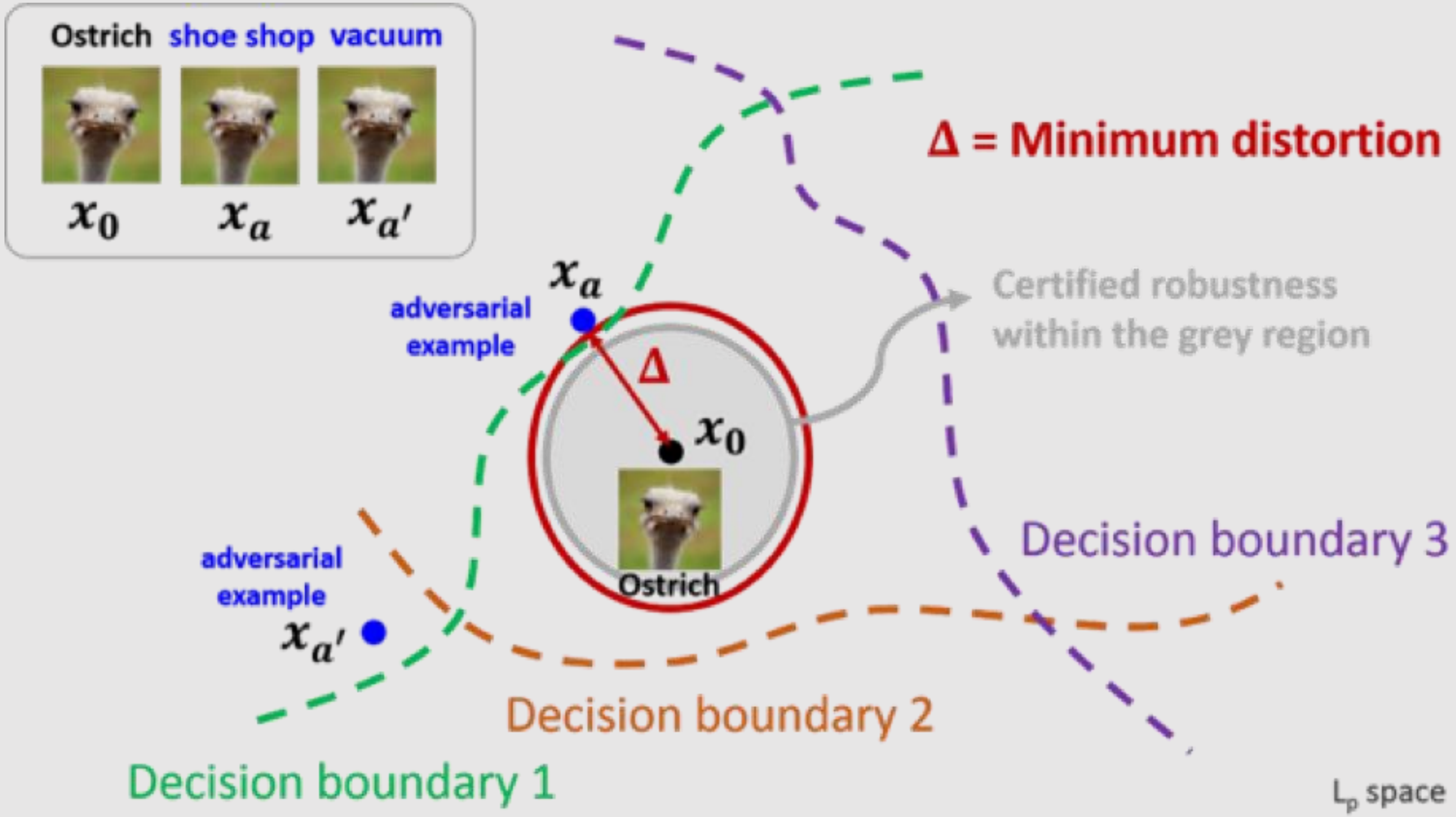

Real sample
Replace
Insert
ORIGINAL
The government made a quick decision
BAE - R
The MASK made a quick decision
judge , doctor , captain
BAE - I
The MASK government made a quick decision
state , british , federal
The government MASK made a quick decision
officials , then , immediately
Adversarial sample

https://arxiv.org/abs/2004.01970

Robustness and adversarial attacks :



Adversarial Robustness


Ostrich
shoe shop
vacuum
x_0
x_a
$x_{a'}$
Δ = Minimum distortion
x_a
adversarial example
Δ
x_0
Certified robustness within the grey region
Ostrich
adversarial example
$x_{a'}$
Decision boundary 3
Decision boundary 2
Decision boundary 1
L$_p$ space