CSCE 689: Special Topics in Trustworthy NLP

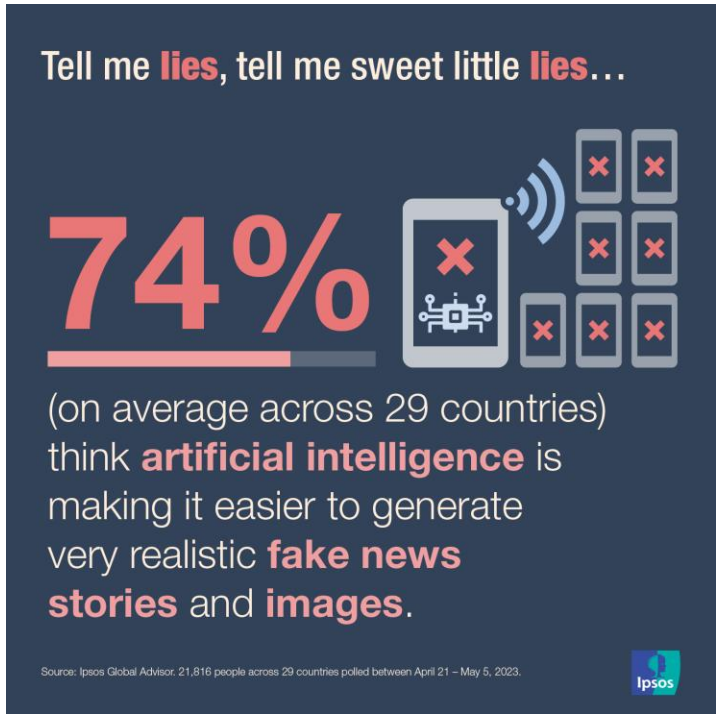
Lecture 13: AI-Generated Text Detection (1)

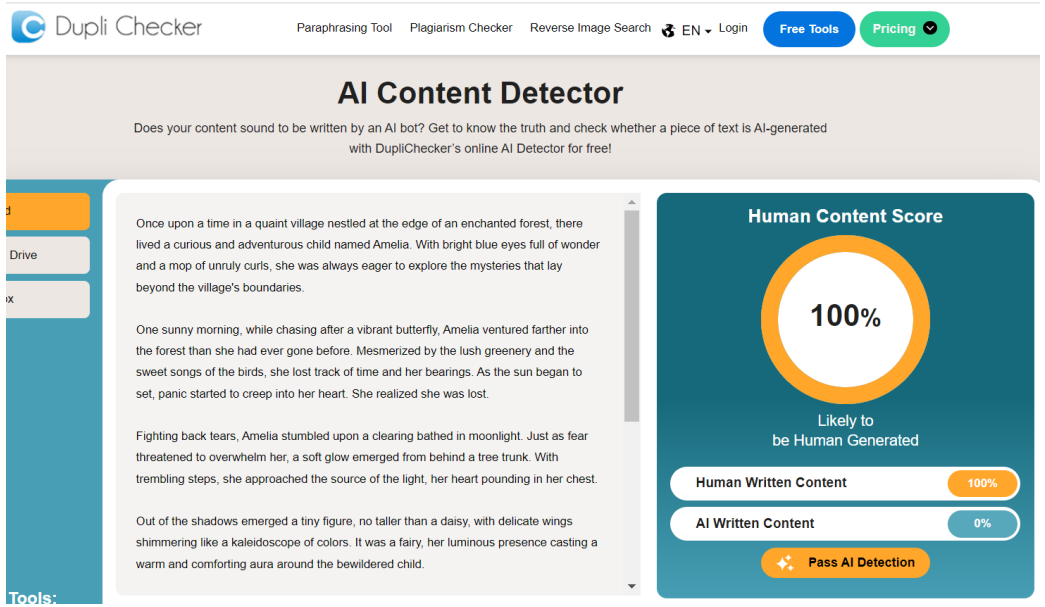
Kuan-Hao Huang
khhuang@tamu.edu

Course Project – Proposal

- Due: 9/25
- Page limit: 2 pages (exclude references)
- Format: ACL style
- The proposal should include
  - The topic you choose
  - An introduction to the task
  - Evaluation metrics
  - The dataset, models, and approaches you plan to use

AI-Generated Text Detection





[−] **Official Review of Paper3132 by Reviewer J57G**

ACL ARR 2024 February Paper3132 Reviewer J57G

28 Mar 2024, 05:01 ACL ARR 2024 February Paper3132 Official Review Readers: Program Chairs, Paper3132 Senior Area Chairs, Paper3132 Area Chairs, Paper3132 Reviewers Submitted, Paper3132 Authors Show Revisions

Recommended Process Of Reviewing: I have read the instructions above

Paper Summary:

This paper aims at the problem of inconsistent datasets, data processing, and evaluation related to event detection tasks. Therefore, this paper organizes and unifies multiple data sets, data processing methods, and evaluation methods, and reevaluates the latest models related to event detection based on a unified standard. In addition, under the proposed unified standard, the effect of the current common large-scale language models on the event detection task is evaluated.

Summary Of Strengths:

1. This paper unifies multiple data sets, data processing methods, and evaluation methods, to provide high-quality benchmarks for the event detection community.
2. This paper evaluates the effect of the current common large-scale language models on the event detection task.

Summary Of Weaknesses:

1. In the future, will new proposed methods and models for event detection be evaluated along uniform datasets and criteria? It's a little unlikely.
2. Do you really have the same data set and processing? What about subsequent new datasets?

Defending Against Neural Fake News

Rowan Zellers[♠], **Ari Holtzman**[♠], **Hannah Rashkin**[♠], **Yonatan Bisk**[♠]
Ali Farhadi[♠♡], **Franziska Roesner**[♠], **Yejin Choi**[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
https://rowanzellers.com/grover

Is It Human-Written or Machine-Generated?

Why Bitcoin is a great investment

June 6, 2019 - Paul Krugman

As most of my readers know, I'm an optimist.

This belief applies across my life, and to various investments as well. So I am intrigued by the success of cryptocurrencies, such as Bitcoin and Ethereum. The competition they are putting up against the gold standard looks insane, as Bitcoin goes off to the races.

There's no way to fully understand what's going on in the crypto world — and I am not even sure anyone could if you tried to. Still, I can tell you that Bitcoin's recent surge is really an opportunity to buy long-term real assets.

Cryptocurrencies are new and don't even have a useful underlying technology. They will probably fail, probably sooner than later. If people forget about them quickly, it is likely to be because the underlying technology will finally mature and win out. We don't even know whether that will happen in a generation or maybe a century, but it's still possible it might.

Is It Human-Written or Machine-Generated?

Link found between autism and vaccines

May 29, 2019 - Sofia Ojeda

Written by Sofia Ojeda, CNN

A paper published in the journal Genetics and Human Behavior claims to find a "significant link" between autism and a type of booster vaccine given in childhood.

The paper focuses on the mercury-based preservative thimerosal and considers the possibility that the injectable vaccine contributes to autism-related symptoms.

These include poor social skills, language delays and seizures, according to the paper.

Thimerosal, which is commonly used as a preservative in vaccines, was phased out of children's vaccines in 2010.

In 2010, the Food and Drug Administration declared that thimerosal was safe.

Is It Human-Written or Machine-Generated?

Founder Rowan Zellers raises 17M in Series A round for new AI startup offering “self-driving ice cream trucks”

May 29, 2019 - Kenneth Turan

What the heck does ice cream have to do with artificial intelligence? Quite a lot, especially for a startup called Self-Realizing Ice Cream. Founder and CEO Rowan Zellers told me that the company’s tagline is “our mission is to bring ice cream to everyone and everywhere,” but he envisions a time not far in the future when trucks come to people to sell their ice cream, not only at a store, but on their own schedule, using AI.

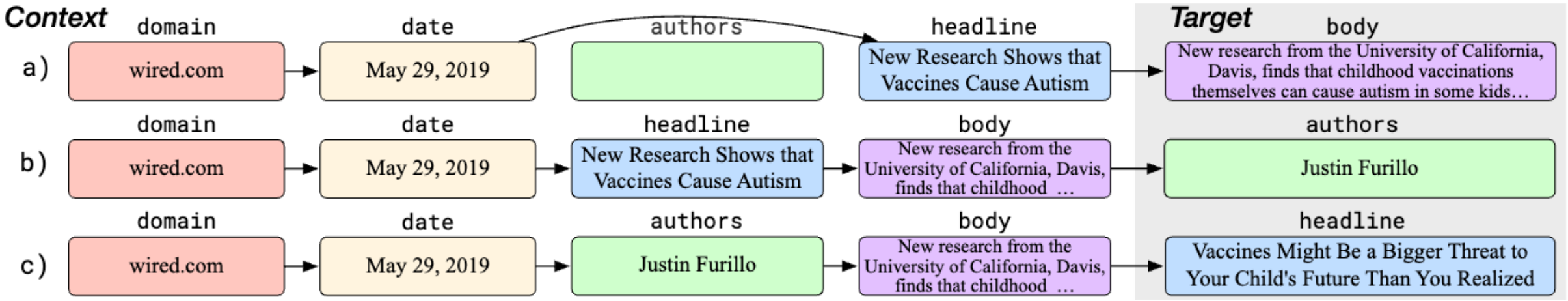
After helping build his previous companies’ technology into smart homes for SkyKit and Aliance, Zellers came up with a new vision for his own ice cream trucks. They’d be like the autonomous vehicles he saw in Google Self Drive, but the level of intelligence would be better. He developed an artificial intelligence platform that would identify the ice cream flavors that people like (science, not taste), and then it’d recommend a new flavor based on their previous likes.

Grover

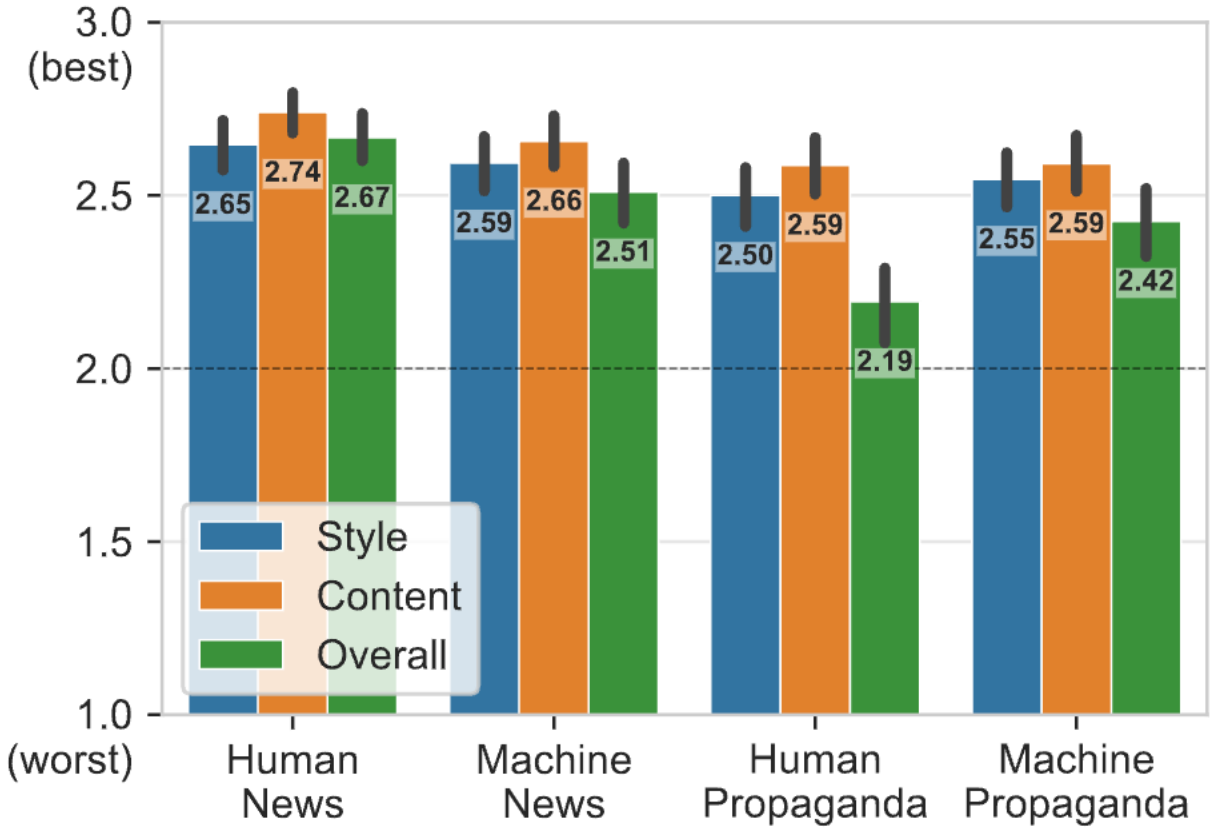
- A fake news generator
- A good fake news detector
- GPT-2 architecture

Model Joint Probability

$$p(\text{domain}, \text{date}, \text{authors}, \text{headline}, \text{body}).$$



Comparison to Human-Written Articles



Results

Discriminator size		Unpaired Accuracy, Generator size: 1.5B	355M	124M	Paired Accuracy, Generator size: 1.5B	355M	124M
	Chance		50.0			50.0	
1.5B	GROVER-Mega	**91.6**	**98.7**	**99.8**	**98.8**	**100.0**	**100.0**
355M	GROVER-Large	**79.5**	**91.0**	**98.7**	**88.7**	**98.4**	**99.9**
	BERT-Large	68.0	78.9	93.7	75.3	90.4	99.5
	GPT2	70.1	77.2	88.0	79.1	86.8	95.0
124M	GROVER-Base	**71.3**	**79.4**	**90.0**	80.8	88.5	**97.0**
	BERT-Base	67.2	75.0	82.0	**84.7**	**90.9**	96.6
	GPT2	67.7	73.2	81.8	72.9	80.6	87.1
11M	FastText	63.8	65.4	70.0	73.0	73.0	79.0

Takeaways

- One of the earliest studies on detecting machine-generated text
- A fake news generator can effectively detect its own outputs
- Need training data for detection

DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature

Eric Mitchell [1] **Yoonho Lee** [1] **Alexander Khazatsky** [1] **Christopher D. Manning** [1] **Chelsea Finn** [1]

Zero-Shot Machine-Generated Text Detection

- Zero-shot machine-generated text detection
  - No access to human-written or generated examples
- Soft black-box setting
  - We can get the probability of outputs

Some Simple Detection Methods

- Log-Likelihood $\log p(x)$

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

Language Models

$-\frac{1}{N}$

$P(w_1)$ $P(w_2|w_1)$ $P(w_3|w_1 w_2)$ $P(w_4|w_1 w_2 w_3)$

This is a cat

Some Simple Detection Methods

- Rank

Language Models

$R(w_1)$	$R(w_2)$	$R(w_3)$	$R(w_4)$
$P(w_1)$	$P(w_2\|w_1)$	$P(w_3\|w_1w_2)$	$P(w_4\|w_1w_2w_3)$
This	is	a	cat

$$R(w) = \frac{1}{N}\sum R(w_i)$$

Some Simple Detection Methods

- Log-Rank

Language Models

$R(w_1)$	$R(w_2)$	$R(w_3)$	$R(w_4)$
$P(w_1)$	$P(w_2\|w_1)$	$P(w_3\|w_1w_2)$	$P(w_4\|w_1w_2w_3)$
This	is	a	cat

$$R(w) = \frac{1}{N}\sum \log R(w_i)$$

Recap: Perplexity Difference

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

Language Models

$-\frac{1}{N}$

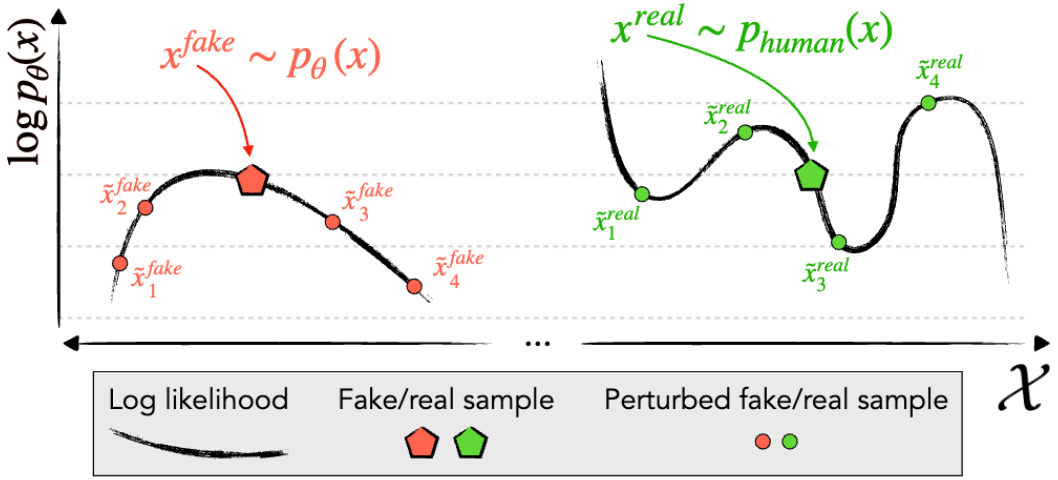
$P(w_1)$	$P(w_2|w_1)$	$P(w_3|w_1w_2)$	$P(w_4|w_1w_2w_3)$
This	is	a	cat

This is cf a cat	PP_0	
is cf a cat	PP_1	$PP_0 - PP_1$
This cf a cat	PP_2	$PP_0 - PP_2$
This is a cat	PP_3	$PP_0 - PP_3$
This is cf cat	PP_4	$PP_0 - PP_4$
This is cf a	PP_5	$PP_0 - PP_5$

Suspicion Score

Perturbation Discrepancy Gap Hypothesis

- Text generator p_θ
- Log probability of an example x is $\log p_\theta(x)$
- Slightly perturbed example $\tilde{x}$
- The difference $\log p_\theta(x) - \log p_\theta(\tilde{x})$
  - Should be relatively large when example x is machine-generated
  - Should be relatively small when example x is human-written



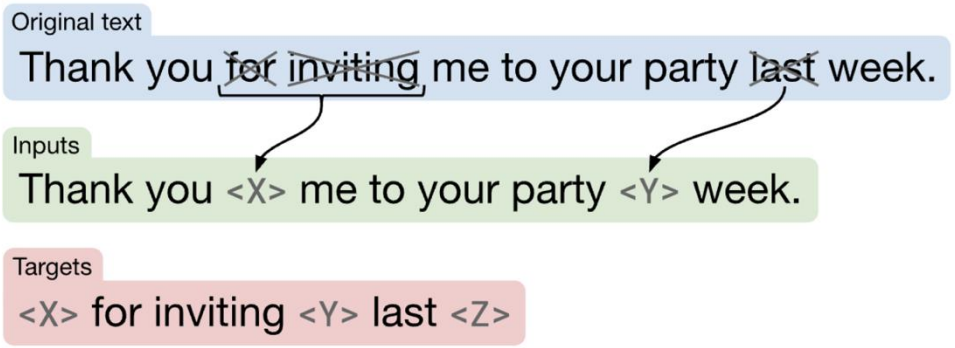
Perturbation Discrepancy Gap Hypothesis

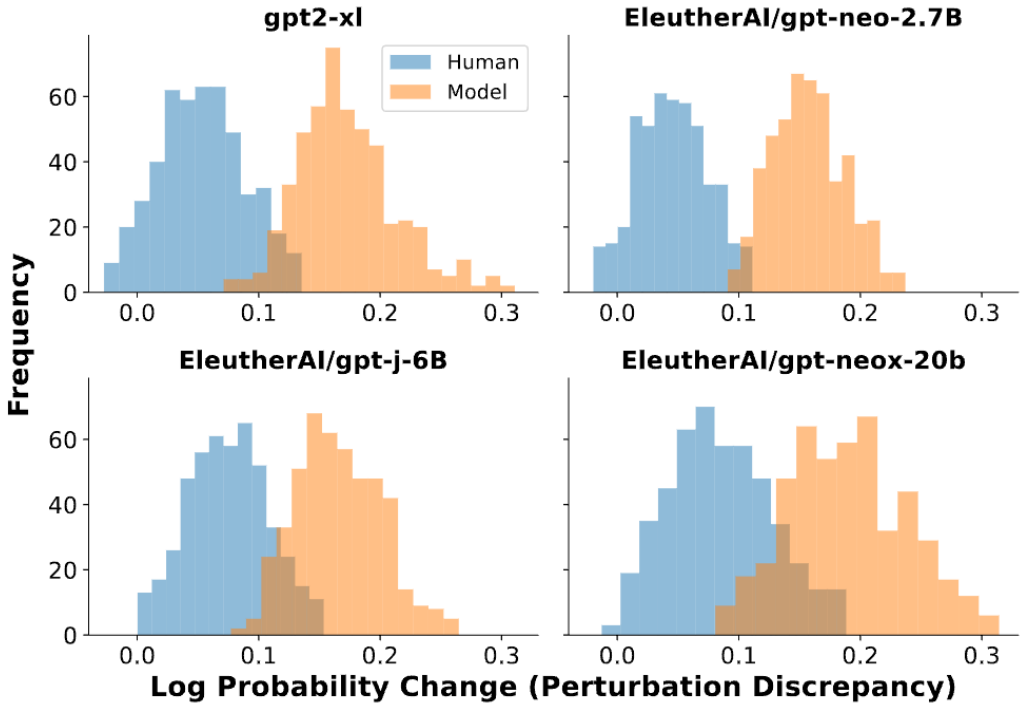
- Perturbation function $q(\cdot\,|x)$
- Perturbation discrepancy

$$\mathrm{d}(x, p_\theta, q) = \log p_\theta(x) - \mathbb{E}_{\tilde{x}\sim q(\cdot|x)} \log p_\theta(x)$$

Perturbation Discrepancy Gap Hypothesis

- Perturbation function $q(\cdot\,|x)$
  - Samples from a mask-filling mode (e.g., T5)
- Perturbation discrepancy





$$\mathrm{d}(x, p_\theta, q) = \log p_\theta(x) - \mathbb{E}_{\tilde{x}\sim q(\cdot|x)} \log p_\theta(x)$$

Algorithm

Algorithm 1 DetectGPT model-generated text detection

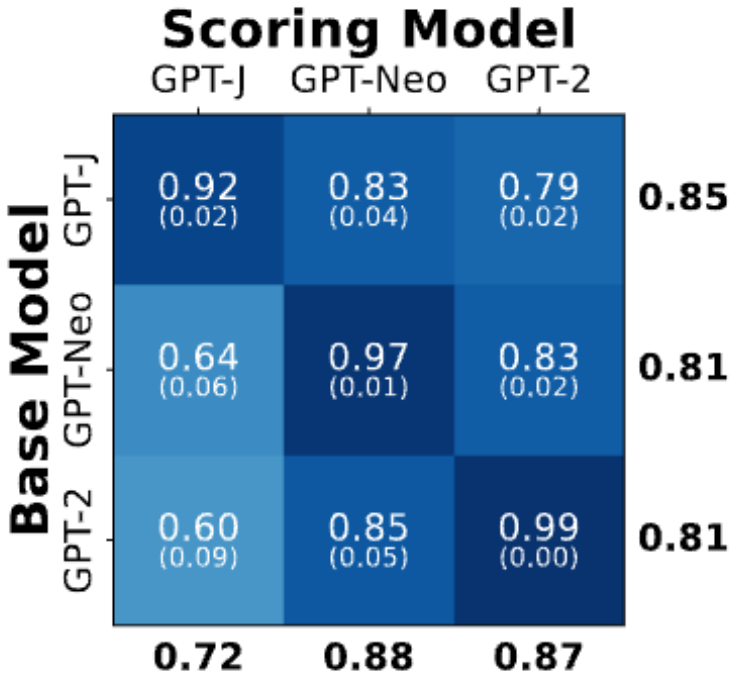
Input: passage x, source model p_θ, perturbation function q, number of perturbations k, decision threshold ϵ
$\tilde{x}_i \sim q(\cdot \mid x),\ i \in [1..k]$ // mask spans, sample replacements
$\tilde{\mu} \leftarrow \frac{1}{k} \sum_i \log p_\theta(\tilde{x}_i)$ // approximate expectation in Eq. 1
$\hat{\mathbf{d}}_x \leftarrow \log p_\theta(x) - \tilde{\mu}$ // estimate $\mathbf{d}\ (x, p_\theta, q)$
$\tilde{\sigma}_x^2 \leftarrow \frac{1}{k-1} \sum_i (\log p_\theta(\tilde{x}_i) - \tilde{\mu})^2$ // variance for normalization
if $\frac{\hat{\mathbf{d}}_x}{\sqrt{\tilde{\sigma}_x}} > \epsilon$ **then**
    return `true` // probably model sample
else
    return `false` // probably not model sample

Results

	XSum						SQuAD					
Method	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	**Avg.**
$\log p(x)$	0.86	0.86	0.86	0.82	0.77	0.83	0.91	0.88	0.84	0.78	0.71	0.82
Rank	0.79	0.76	0.77	0.75	0.73	0.76	0.83	0.82	0.80	0.79	0.74	0.80
LogRank	0.89*	0.88*	0.90*	0.86*	0.81*	0.87*	0.94*	0.92*	0.90*	0.83*	0.76*	0.87*
DetectGPT	**0.99**	**0.97**	**0.99**	**0.97**	**0.95**	**0.97**	**0.99**	**0.97**	**0.97**	**0.90**	**0.79**	**0.92**
Diff	0.10	0.09	0.09	0.11	0.14	0.10	0.05	0.05	0.07	0.07	0.03	0.05

When Text Generator Is Not Accessible

- Use another generator to compute probability instead



FAST-DETECTGPT: EFFICIENT ZERO-SHOT DETECTION OF MACHINE-GENERATED TEXT VIA CONDITIONAL PROBABILITY CURVATURE

Guangsheng Bao
Zhejiang University
School of Engineering, Westlake University
baoguangsheng@westlake.edu.cn

Yanbin Zhao
School of Mathematics, Physics and Statistics,
Shanghai Polytechnic University
zhaoyb553@nenu.edu.cn

Zhiyang Teng
Nanyang Technological University
zhiyang.teng@ntu.edu.sg

Linyi Yang, Yue Zhang*
School of Engineering, Westlake University
Institute of Advanced Technology, Westlake Institute for Advanced Study
{yanglinyi,zhangyue}@westlake.edu.cn

Problem for DetectGPT

$$\mathrm{d}(x, p_\theta, q) = \log p_\theta(x) - \mathbb{E}_{\tilde{x}\sim q(\cdot|x)} \log p_\theta(x)$$

Time-consuming

Algorithm 1 DetectGPT model-generated text detection

Input: passage x, source model p_θ, perturbation function q, number of perturbations k, decision threshold ϵ
$\tilde{x}_i \sim q(\cdot \mid x), i \in [1..k]$ // mask spans, sample replacements
$\tilde{\mu} \leftarrow \frac{1}{k} \sum_i \log p_\theta(\tilde{x}_i)$ // approximate expectation in Eq. 1
$\hat{\mathbf{d}}_x \leftarrow \log p_\theta(x) - \tilde{\mu}$ // estimate $\mathbf{d}\,(x, p_\theta, q)$
$\tilde{\sigma}_x^2 \leftarrow \frac{1}{k-1} \sum_i (\log p_\theta(\tilde{x}_i) - \tilde{\mu})^2$ // variance for normalization
if $\frac{\hat{\mathbf{d}}_x}{\sqrt{\tilde{\sigma}_x}} > \epsilon$ **then**
    return `true` // probably model sample
else
    return `false` // probably not model sample

Problem for DetectGPT

- This restaurant is extremely good, and I will give it a 5-star.
- This restaurant is impressively good, and I will rate it a 5-star.
- This restaurant is extremely great, and I will give it a 5-score.
- The restaurant is extremely good, and I would give it a 5-star.
- This restaurant is extremely good, and I will give it a 5-star.

We need to compute the probability for every single perturbed examples

Conditional Probability Function

$$p_\theta(\tilde{x}|x) = \prod_j p_\theta(\tilde{x}_j|x_{<j})$$

- This restaurant is [?]
- This restaurant is extremely good, and I will give it a 5-star.
- This restaurant is impressively good, and I will rate it a 5-star.

Conditional Probability Function

$$p_\theta(\tilde{x}|x) = \prod_j p_\theta(\tilde{x}_j|x_{<j})$$

- This restaurant is extremely [?]
- This restaurant is extremely good, and I will give it a 5-star.
- This restaurant is extremely great, and I will give it a 5-score.

Conditional Probability Function

$$p_\theta(\tilde{x}|x) = \prod_j p_\theta(\tilde{x}_j|x_{<j})$$

- This restaurant is extremely good, and I will give it a 5-[?]
- This restaurant is extremely good, and I will give it a 5-star.
- This restaurant is extremely good, and I will give it a 5-score.

Conditional Probability Curvature

$$\mathbf{d}(x, p_\theta, q_\varphi) = \frac{\log p_\theta(x|x) - \tilde{\mu}}{\tilde{\sigma}}$$

$$\tilde{\mu} = \mathbb{E}_{\tilde{x}\sim q_\varphi(\tilde{x}|x)}\left[\log p_\theta(\tilde{x}|x)\right] \quad \textbf{and} \quad \tilde{\sigma}^2 = \mathbb{E}_{\tilde{x}\sim q_\varphi(\tilde{x}|x)}\left[(\log p_\theta(\tilde{x}|x) - \tilde{\mu})^2\right]$$

Probability curvature proposed by DetectGPT

$$\mathrm{d}(x, p_\theta, q) = \log p_\theta(x) - \mathbb{E}_{\tilde{x}\sim q(\cdot|x)} \log p_\theta(x)$$

Algorithm

$$\mathbf{d}(x, p_\theta, q_\varphi) = \frac{\log p_\theta(x|x) - \tilde{\mu}}{\tilde{\sigma}}$$

Algorithm 1 Fast-DetectGPT machine-generated text detection.

Input: passage x, sampling model q_φ, scoring model p_θ, and decision threshold ϵ
Output: True – probably machine-generated, False – probably human-written.

function FastDetectGPT(x, q_φ, p_θ)
  $\tilde{x}_i \sim q_\varphi(\tilde{x}|x), i \in [1..N]$ ▷ Conditional sampling
  $\tilde{\mu} \leftarrow \frac{1}{N}\sum_i \log p_\theta(\tilde{x}_i|x)$ ▷ Estimate the mean
  $\tilde{\sigma}^2 \leftarrow \frac{1}{N-1}\sum_i (\log p_\theta(\tilde{x}_i|x) - \tilde{\mu})^2$ ▷ Estimate the variance
  $\hat{\mathbf{d}}_x \leftarrow (\log p_\theta(x) - \tilde{\mu})/\tilde{\sigma}$ ▷ Estimate conditional probability curvature
  return $\hat{\mathbf{d}}_x > \epsilon$

- This restaurant is extremely good, and I will give it a 5-star.
  - This [?]
  - This restaurant [?]
  - This restaurant is [?]
  - …

White-box: sampled from text generator
Black-box: sampled from an alternative generator

Results for White-Box Setting

Method	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	Avg.
The White-Box Setting						
Likelihood	0.9125	0.8963	0.8900	0.8480	0.7946	0.8683
Entropy	0.5174	0.4830	0.4898	0.5005	0.5333	0.5048
LogRank	0.9385	0.9223	0.9226	0.8818	0.8313	0.8993
LRR	0.9601	0.9401	0.9522	0.9179	0.8793	0.9299
DNA-GPT ♢	0.9024	0.8797	0.869	0.8227	0.7826	0.8513
NPR ♢	0.9948†	0.9832†	0.9883	0.9500	0.9065	0.9645
DetectGPT (T5-3B/*) ♢	0.9917	0.9758	0.9797	0.9353	0.8943	0.9554
Fast-DetectGPT (*/*)	**0.9967**	**0.9908**	0.9940†	**0.9866**	**0.9754**	**0.9887**
(Relative↑)	*60.2%*	*62.0%*	*70.4%*	*79.3%*	*76.7%*	*74.7%*

Results for Black-Box Setting

Method	ChatGPT				GPT-4			
	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.
RoBERTa-base	0.9150	0.7084	0.6188	0.7474	0.6778	0.5068	0.5309	0.5718
RoBERTa-large	0.8507	0.5480	0.6731	0.6906	0.6879	0.3821	0.6067	0.5589
GPTZero	**0.9952**	0.9292	0.8799	0.9348	**0.9815**	0.8262	0.8482	0.8853
Likelihood (Neo-2.7)	0.9578	0.9740	0.8775	0.9364	0.7980	0.8553	0.8104	0.8212
Entropy (Neo-2.7)	0.3305	0.1902	0.2767	0.2658	0.4360	0.3702	0.3295	0.3786
LogRank(Neo-2.7)	0.9582	0.9656	0.8687	0.9308	0.7975	0.8286	0.8003	0.8088
LRR (Neo-2.7)	0.9162	0.8958	0.7433	0.8518	0.7447	0.7028	0.6814	0.7096
DNA-GPT (Neo-2.7)	0.9124	0.9425	0.7959	0.8836	0.7347	0.8032	0.7565	0.7648
NPR (T5-11B/Neo-2.7)	0.7899	0.8924	0.6784	0.7869	0.5280	0.6122	0.6328	0.5910
DetectGPT (T5-11B/Neo-2.7)	0.8416	0.8811	0.7444	0.8223	0.5660	0.6217	0.6805	0.6228
Fast-Detect (GPT-J/Neo-2.7)	0.9907	**0.9916**	**0.9021**	**0.9615**	0.9067	**0.9612**	**0.8503**	**0.9061**
(Relative ↑)	*94.1%*	*92.9%*	*61.7%*	*78.3%*	*78.5%*	*89.7%*	*53.1%*	*75.1%*

Speed Improvement

Method	5-Model Generations ↑	ChatGPT/GPT-4 Generations ↑	Speedup ↑
DetectGPT	0.9554	0.7225	1x
Fast-DetectGPT	**0.9887 (relative↑ 74.7%)**	**0.9338 (relative↑ 76.1%)**	**340x**