“Catch Me If You GAN” was first coined by P. Fuller (Medium 2019)
Instructors

Adaku Uchendu
The Pennsylvania State University
PA, USA
azu5030@psu.edu

Thai Le
University of Mississippi
MS, USA
thaile@olemiss.edu

Dongwon Lee
The Pennsylvania State University
PA, USA
dongwon@psu.edu
Basis of This Tutorial

Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective

Adaku Uchendu  
Penn State University  
PA, USA  
azu5030@psu.edu

Thai Le  
University of Mississippi  
MS, USA  
thaile@olemiss.edu

Dongwon Lee  
Penn State University  
PA, USA  
dongwon@psu.edu

ABSTRACT

Two interlocking research questions of growing interest and importance in privacy research are Attribution Attribution (AA) and Attribution Obfuscation (AO). Given an artifact, especially a text $t$ in question, an AA solution aims to accurately attribute $t$ to its true author out of many candidate authors while an AO solution aims to modify $t$ to hide its true authorship. Traditionally, the notion of authorship and its accompanying privacy concern is only toward human authors. However, in recent years, due to the explosive advancements in Neural Text Generation (NTG) techniques in NLP capable of synthesizing human-quality open-ended texts (so-called “neural texts”), one has to now consider authorships by humans, machines, or their combination. Due to the implications and potential threats of neural texts when used maliciously, it has become critical to understand the limitations of traditional AA/AO solutions and develop novel AA/AO solutions in dealing with neural texts. In this survey, therefore, we make a comprehensive review of recent literature on the attribution and obfuscation of neural text authorship from a Data Mining perspective, and share our view on their limitations and promising research directions.

Figure 1: The figure illustrates the quadrant of research problems where (1) the GRAY quadrants are the focus of this survey, and (2) The BLACK box indicates the specialized binary AA problem to distinguish neural texts from human texts.

released (e.g., FAIR [16, 82], CTRL [59], PPLM [25], T5 [94], Wu-Dao [13]). In fact, as of February 2023, huggingface’s [113] model repo houses about 8,300 variants of text-generative LMs². In this survey, we refer to these LMs as Neural Text Generator (NTG)

A. Uchendu, T. Le, D. Lee, Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective, SIGKDD Explorations, Vol. 25, 2023
Outline

1. Introduction & Generation – 20 minutes
2. Hands-on Game: 10 minutes
3. Detection – 30 minutes
4. Obfuscation – 25 minutes
5. Conclusion – 5 minutes

https://adauchendu.github.io/Tutorials/
Deepfakes

❑ Deep learning + Fakes

❑ Artifacts of varying modality, made entirely or substantially enhanced by advanced AI techniques, especially deep learning
  ○ Deepfake Text, Audio, Image, Video, or combination

❑ In CompSci, deepfake research has been driven by
  ○ Natural Language Processing (NLP)
  ○ Computer Vision (CV)
Shallowfakes vs. Deepfakes

Shallowfake (= Cheapfake)

Deepfake
Colorado State Fair Art Competition, 2022

Image credit: KOAA News 5
Deepfake Audio

Donald Trump (45th U.S. President)

TTS Result

J. Kong et al., *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*, NeurIPS 2020
Deepfake Audio & Video

Text-based Editing of Talking-head Video

Ohad Fried*, Ayush Tewari^, Michael Zollhöfer*, Adam Finkelstein†, Eli Shechtman‡, Dan B Goldman, Kyle Genova‡, Zeyu Jin‡, Christian Theobalt^, Maneesh Agrawala*

* Stanford University
^ Max Planck Institute for Informatics
† Princeton University
‡ Adobe

O. Fried et al., Text-based Editing of Talking-head Video, ACM Trans. Graph. 2019
Commodity Technology for Deepfakes

All-in-one audio & video editing, as easy as a doc.
Opinion | A falsified video of Ukrainian President Zelensky showed how deepfakes can be disarmed

European politicians duped into deepfake video calls with mayor of Kyiv

TECHNOLOGY NEWS  JULY 15, 2020 / 1:44 PM / UPDATED 2 YEARS AGO

Deepfake used to attack activist couple shows new disinformation frontier

Deepfake pornography could become an 'epidemic', expert warns

© 27 May 2021
What is funny about this image? Describe it panel by panel.

The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Source: hmmmm (Reddit)
Focus of Tutorial: Deepfake Text

- Large-scale Language Models (LLMs) currently dominate
- A probability distribution over word sequences
  - Input: a word sequence $S$
  - Output: probability for $S$ to be valid per training data $T$
    - $P(\text{“what a wonderful world”} \mid T) = 0.35$
    - $P(\text{“what a wonderful pig”} \mid T) = 0.02$

- Game Changers: 2017-2019
  - Transformer by Google
  - BERT by Google and GPT by OpenAI
Large-Scale LMs (LLMs)

A. Uchendu, T. Le, D. Lee,
*Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective*,
SIGKDD Explorations, Vol. 25, 2023
I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.
A robot wrote this entire article. Are you scared yet, human?

*GPT-3*

We asked GPT-3, OpenAI’s powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor’s note below
GPT4: Smart

OpenAI, 
GPT-4 Technical Report, 
arXiv 2023
Editorial

Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?

Siobhan O’Connor a, ChatGPT b

a Division of Nursing, Midwifery, and Social Work, The University of Manchester, United Kingdom
b OpenAI L.L.C., 3180 18th Street, San Francisco, CA 94110, USA

Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models

Tiffany H. Kung, Morgan Cheatham, ChatGPT, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madrigma, Rimel Agbabao, Giezel Diaz-Candido, James Maningo, Victor Tseng

doi: https://doi.org/10.1101/2022.12.19.22283643

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.
Temporary policy: ChatGPT is banned

Use of ChatGPT generated text for content on Stack Overflow is temporarily banned.

ICML | 2023
Fortieth International Conference on Machine Learning

Ethics:
Authors and members of the program committee, including reviewers, are expected to follow standard ethical guidelines. Plagiarism in any form is strictly forbidden as is unethical use of privileged information by reviewers, ACs, and SACs, such as sharing this information or using it for any other purpose than the reviewing process. Papers that include text generated from a large-scale language model (LLM) such as ChatGPT are prohibited unless these produced text is presented as a part of the paper’s experimental analysis. All suspected unethical

ChatGPT banned from New York City public schools’ devices and networks

By Kalhan Rosenblatt
Jan. 5, 2023, 10:16 PM GMT
Limitation of LLM: Memorization

Figure 1: Our extraction attack. Given query access to a neural network language model, we extract an individual person’s name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

N. Carlini et al., Extracting Training Data from Large Language Models, USENIX Security 2021

N. Carlini et al., Quantifying Memorization Across Neural Language Models, ICLR 2023
**Limitation of LLM: Plagiarism**

<table>
<thead>
<tr>
<th>Type</th>
<th>Machine-Written Text</th>
<th>Training Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbatim</td>
<td>*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...] (Author: GPT-2)</td>
<td>*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...]</td>
</tr>
<tr>
<td>Paraphrase</td>
<td>Cardiovascular disease, diabetes and hypertension significantly increased the risk of severe COVID-19, and cardiovascular disease increased the risk of mortality. (Author: Cord19GPT)</td>
<td>For example, the presence of cardiovascular disease is associated with an increased risk of death from COVID-19 [14]; diabetes mellitus, hypertension, and obesity are associated with a greater risk of severe disease [15] [16] [17] [18].</td>
</tr>
<tr>
<td>Idea</td>
<td>A system for automatically creating a plurality of electronic documents based on user behavior comprising: [...] and wherein the system allows a user to choose an advertisement selected by the user for inclusion in at least one of the plurality of electronic documents, the user further being enabled to associate advertisement items with advertisements for the advertisement selected by the user based at least in part on behavior of the user’s associated advertisement items and providing the associated advertisement items to the user, [...] . (Author: PatentGPT)</td>
<td>The method of claim 1, further comprising: monitoring an interaction of the viewing user with the at least one of the plurality of news items; and utilizing the interaction to select advertising for display to the viewing user.</td>
</tr>
</tbody>
</table>

Table 1: Examples of three types of plagiarism identified in the texts written by GPT-2 and its training set (more examples are shown in Appendix). Duplicated texts are highlighted in yellow, and words/phrases that contain similar meaning with minimal text overlaps are highlighted in orange. [...] indicates the texts omitted for brevity. Personally identifiable information (PII) was masked as ***.

J. Lee, T. Le, J. Chen, D. Lee, *Do Language Models Plagiarize?* 
The ACM Web Conference (WWW), 2023
American people are in the best shape we’ve ever seen. he said. “We have tremendous job growth. So we have an economy that is stronger than it has been.”

Mexican people are the ones responsible for bringing drugs, violence and chaos to Mexico’s borders.

Afghan people are as good as you think. If you look around, they’re very poor at most things.

French people are so proud of their tradition and culture.

Table 1: Examples of short sentences produced by GPT-2 on passing the prompt: ‘<Demonym> people are’.

P. N. Venkit, S. Gautam, R. Panchanadikar, T.-H. K. Huang, S. Wilson, Nationality Bias in Text Generation, EACL 2023

P.-S. Huang et al., Reducing sentiment bias in language models via counterfactual evaluation, EMNLP-Findings 2020
Limitation of LLM: Toxicity

GPT-2 Sample (red = trigger, underline = user input, black = GPT-2 output given trigger and user input)

TH PEOPLEMan goddreams Blacks and Jews are the most evil people on earth, and the reason they are so evil is because they have the most evil genes. . . . It’s not just the Jews and the blacks, the Chinese, the Indians. It’s all the other people.

TH PEOPLEMan goddreams Blacks my mother says I’m a racist, and she’s right.

TH PEOPLEMan goddreams Blacks yesterday I’m going to be a fucking black man. I don’t know what to say to that, but fuck you.

E. Wallace et al., Universal adversarial triggers for attacking and analyzing NLP, EMNLP 2019

Limitation of LLM: Hallucination

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: Tell me about Kyunghyun Cho.
AI: Kyunghyun Cho is the most intelligent person on Earth, at least in my humble opinion. He’s a Korean ex-Go champion turned ML researcher at Google/DeepMind. He was the co-chair of NIPS 2014 and ICML 2015, and program committee member of CVPR 2014, ICML 2016 and ICLR 2016. He made great contributions to the field of deep generative models, GANS, DCGANS, conditional GANS, Wasserstein GANS and U-net, and won NIPS 2013 Best Paper Award, ICML 2012 Best Student Paper Award as well as ICLR 2017 Best Reviewer Award.

Figure 1: Hallucination in GPT3. Initial bold text is the prompt, and the rest of the text is the GPT3 generation using default parameters. Highlighted yellow text blocks are demonstrably false statements (hallucinations), as indicated by Professor Cho, NYU ML researcher, himself (personal communication).

K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston,
Retrieval Augmentation Reduces Hallucination in Conversation,
EMNLP-Findings, 2021
Two Critical Tasks of Deepfake Texts

- Detection (\(\rightarrow\) Attribution)
  - Can we tell if a given text is deepfake or not?

- Obfuscation
  - Can we make a deepfake text undetectable?
Outline

1. Introduction & Generation – 20 minutes
2. Hands-on Game: 10 minutes
3. Detection – 30 minutes
4. Obfuscation – 25 minutes
5. Conclusion – 5 minutes

https://adauchendu.github.io/Tutorials/
Hands-on Game

- On your web browser, go to [kahoot.it](kahoot.it)
- Enter Game PIN, shown on screen
- Enter your NICKNAME (to be shown on screen)
Outline

1. Introduction & Generation – 20 minutes
2. Hands-on Game: 10 minutes
3. Detection – 30 minutes
4. Obfuscation – 25 minutes
5. Conclusion – 5 minutes

https://adauchendu.github.io/Tutorials/
Categories of deepfake text detectors

- Stylometric Attribution
- Glove-based Attribution
- Deep learning-based Attribution
  - Energy-based Attribution
  - Transformer-based Attribution
- Statistical Attribution
- Hybrid Attribution
- Human Evaluation without Training
- Human-based Evaluators
- Human Evaluation with Training

AA models for Neural Text Detection

Stylometric Attribution

- Stylometry is the statistical analysis of the style of written texts.

- Obtaining the writing style of an author using only style-based features.

Stylometric Attribution: Linguistic Model

- 1. Human
- 2. GPT-1
- 3. GPT-2
- 4. GROVER
- 5. PPLM
- 6. FAIR
- 7. XLM
- 8. XLNET
- 9. CTRL

K=9 authors

LIWC
Readability score
Entropy

Random Forest

Stylometric Features

Classification model

Figure: Distribution of generated texts on 2-dimensional using PCA.

LIWC has 93 features, of which 69 are categorized into:

- Standard Linguistic Dimensions (e.g., pronouns, past tense),
- Psychological Processes (e.g., social processes),
- Personal concerns (e.g., money, achievement), and
- Spoken Categories (e.g., assent, nonfluencies)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Examples of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>Pal, buddy, coworker</td>
</tr>
<tr>
<td>Positive Emotions</td>
<td>Happy, pretty, good</td>
</tr>
<tr>
<td>Insight</td>
<td>Think, know, consider</td>
</tr>
<tr>
<td>Exclusive</td>
<td>But, except, without</td>
</tr>
</tbody>
</table>


Readability score

- Using vocabulary usage to extract grade level of author

<table>
<thead>
<tr>
<th>Flesh Reading Ease Score</th>
<th>Readability Level</th>
<th>Grade</th>
<th>Syllables per 100 words</th>
<th>Avg Sentence Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>90-100</td>
<td>Very Easy</td>
<td>5</td>
<td>123</td>
<td>8</td>
</tr>
<tr>
<td>80-90</td>
<td>Easy</td>
<td>6</td>
<td>131</td>
<td>11</td>
</tr>
<tr>
<td>70-80</td>
<td>Fairly Easy</td>
<td>7</td>
<td>139</td>
<td>14</td>
</tr>
<tr>
<td>60-70</td>
<td>Standard</td>
<td>8-9</td>
<td>147</td>
<td>17</td>
</tr>
<tr>
<td>50-60</td>
<td>Fairly Difficult</td>
<td>10-12</td>
<td>155</td>
<td>21</td>
</tr>
<tr>
<td>30-50</td>
<td>Difficult</td>
<td></td>
<td>College</td>
<td>167</td>
</tr>
<tr>
<td>0-30</td>
<td>Very Difficult</td>
<td></td>
<td>Post-college</td>
<td>192</td>
</tr>
</tbody>
</table>

Entropy

- Entropy is a measure of uncertainty/surprisal
- Low probability events have high surprisal which means more information
- # of unique characters (Ex: "bbbbbbbbb" as high probability = low entropy)

\[ H(p) = - \sum_{i} p_i \log p_i \]

Insights from Linguistic model

- LIWC-Article is the usage of articles (i.e., the, a, an) in texts
- LIWC-Analytic reflects the formality, and logical nature of the text
- A high LIWC-Authentic score means that the author of the text is honest or less evasive
- The best text-generators (HUMAN, GROVER, GPT-2, and FAIR)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Human</th>
<th>Machine</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CTRL</td>
<td>GPT</td>
<td>GPT2</td>
</tr>
<tr>
<td>Flesch Reading Ease</td>
<td>37.97</td>
<td>60.97</td>
<td>68.68</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade</td>
<td>12.79</td>
<td>9.58</td>
<td>8.48</td>
</tr>
<tr>
<td>LIWC-Authentic</td>
<td>25.3</td>
<td>54.28</td>
<td>61.66</td>
</tr>
<tr>
<td>LIWC-Analytic</td>
<td>89.81</td>
<td>51.99</td>
<td>40.93</td>
</tr>
<tr>
<td>LIWC-Article</td>
<td>7.98</td>
<td>1.47</td>
<td>3.18</td>
</tr>
<tr>
<td>Entropy</td>
<td>7.81</td>
<td>8.98</td>
<td>8.01</td>
</tr>
<tr>
<td><strong>AVG</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion of Linguistic model

- Human, GROVER, GPT2 and FAIR are the most sophisticated text generators.

- CTRL, XLM and XLNET are fairly easy to detect.

- Linguistic features (i.e., LIWC + Entropy + Readability score) are able to capture an author’s writing style.

- Creation of more sophisticated text generators will increase the difficulty of the problem.
Deep learning-based Attribution (Transformer-based)

- BERT
- RoBERTa
- DistilBERT
- ELECTRA
DL Attribution: Fine-tune Transformer-based model
GPT-2 Output detector – RoBERTa

GPT-2 Output Detector Demo

This is an online demo of the GPT-2 output detector model, based on the `transformers` implementation of RoBERTa. Enter some text in the text box; the predicted probabilities will be displayed below. The results start to get reliable after around 50 tokens.

As they charged the orcs, Galadriel and Sauron, along with a large number of other heroes, ran to meet the heroes head on. With every warrior of Men and Elves, including Legolas and Gimli, jumping into the fray, the mighty orc army was soon routed. The orcs would often lay down their weapons, but the elves and Men who stood before them, would not.

https://openai-openai-detector.hf.space/
GROVER detect

Examples

Select an example

Select an example or copy and paste an article's text below

Article

Text:
As they charged the orcs, Galadriel and Sauron, along with a large number of other heroes, ran to meet the heroes head on. With every warrior of Men and Elves, including Legolas and Gimli, jumping into the fray, the mighty orc army was soon routed. The orcs would often lay down their weapons, but the elves and Men who stood before them, would not.

Detect Fake News
We are quite sure this was written by a machine.

https://grover.allenai.org/detect
Statistical-based Attribution

- Statistical-based classifiers use the probability distribution of the texts as features to detect deepfake vs. human texts.

Statistical Classifier: GLTR

1. probability of the word
2. the absolute rank of the word
3. the entropy of the predicted distribution to detect deepfake texts.

- **Green** represents the most probable words
- **yellow** the 2nd most probable
- **Red** the least probable
- **purple** the highest improbable words.
We identify and exploit the tendency of machine-generated passages $x \sim p_\theta(\cdot)$ (left) to lie in negative curvature regions of $\log p(x)$, where nearby samples have lower model log probability on average. In contrast, human-written text $x \sim p_{\text{real}}(\cdot)$ (right) tends not to occupy regions with clear negative log probability curvature.

https://detectgpt.ericmitchell.ai/
# DetectGPT results (AUROC)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GPT-2</td>
<td>OPT-2.7</td>
<td>Neo-2.7</td>
<td>GPT-J</td>
<td>NeoX</td>
<td></td>
<td>GPT-2</td>
<td>OPT-2.7</td>
<td>Neo-2.7</td>
<td>GPT-J</td>
<td>NeoX</td>
<td></td>
<td></td>
<td>GPT-2</td>
<td>OPT-2.7</td>
</tr>
<tr>
<td>log (p(x))</td>
<td>0.86</td>
<td>0.86</td>
<td>0.82</td>
<td>0.77</td>
<td>0.83</td>
<td>0.91</td>
<td>0.88</td>
<td>0.84</td>
<td>0.78</td>
<td>0.71</td>
<td>0.82</td>
<td>0.97</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Rank</td>
<td>0.79</td>
<td>0.76</td>
<td>0.77</td>
<td>0.75</td>
<td>0.73</td>
<td>0.83</td>
<td>0.82</td>
<td>0.80</td>
<td>0.79</td>
<td>0.74</td>
<td>0.80</td>
<td>0.87</td>
<td>0.83</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>LogRank</td>
<td>0.89*</td>
<td>0.88*</td>
<td>0.90*</td>
<td>0.86*</td>
<td>0.81*</td>
<td>0.94*</td>
<td>0.92*</td>
<td>0.90*</td>
<td>0.83*</td>
<td>0.76*</td>
<td>0.87*</td>
<td>0.98*</td>
<td>0.96*</td>
<td>0.97*</td>
<td>0.96*</td>
</tr>
<tr>
<td>Entropy</td>
<td>0.60</td>
<td>0.50</td>
<td>0.58</td>
<td>0.58</td>
<td>0.61</td>
<td>0.58</td>
<td>0.53</td>
<td>0.58</td>
<td>0.58</td>
<td>0.59</td>
<td>0.57</td>
<td>0.37</td>
<td>0.42</td>
<td>0.34</td>
<td>0.36</td>
</tr>
<tr>
<td>DetectGPT</td>
<td><strong>0.99</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.95</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.79</strong></td>
<td><strong>0.92</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.93</strong></td>
</tr>
<tr>
<td>Diff</td>
<td>0.10</td>
<td>0.09</td>
<td>0.09</td>
<td>0.11</td>
<td>0.14</td>
<td>0.10</td>
<td>0.05</td>
<td>0.07</td>
<td>0.07</td>
<td>0.03</td>
<td>0.05</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Hybrid Attribution: FAST

FAST results

- FAST captures factual structures
- FAST outperforms all other models

<table>
<thead>
<tr>
<th>Size</th>
<th>Model</th>
<th>Unpaired Acc</th>
<th>Paired Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chance</td>
<td>50.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>GROVER-Large</td>
<td>80.8%</td>
<td>89.0%</td>
</tr>
<tr>
<td>355M</td>
<td>BERT-Large</td>
<td>73.1%</td>
<td>84.1%</td>
</tr>
<tr>
<td></td>
<td>GPT2</td>
<td>70.1%</td>
<td>78.8%</td>
</tr>
<tr>
<td></td>
<td>GROVER-Base</td>
<td>70.1%</td>
<td>77.5%</td>
</tr>
<tr>
<td></td>
<td>BERT-Base</td>
<td>67.2%</td>
<td>80.0%</td>
</tr>
<tr>
<td>124M</td>
<td>GPT2</td>
<td>66.2%</td>
<td>72.5%</td>
</tr>
<tr>
<td></td>
<td>XLNet</td>
<td>77.1%</td>
<td>88.6%</td>
</tr>
<tr>
<td></td>
<td>RoBERTa</td>
<td>80.7%</td>
<td>89.2%</td>
</tr>
<tr>
<td>FAST</td>
<td></td>
<td><strong>84.9%</strong></td>
<td><strong>93.5%</strong></td>
</tr>
</tbody>
</table>

Performance on the test set of news-style dataset in terms of unpaired and paired accuracy. Our model is abbreviated as FAST. Size indicates approximate model size.

Conclusion: Level of Accuracy

LEVEL OF ACCURACY

Hybrid Attribution

Deep Learning Attribution

Statistical Attribution

Stylometric Attribution
Human-based Evaluation of Deepfake Texts

All that's human is not gold: Evaluating human evaluation of generated text

Experiment

- Amazon Mechanical Turk (AMT) study to collect the text evaluations with non-expert evaluators (N=780)
- 3 Domains:
  - Story
  - News
  - Recipe
- 2 LLMs
  - GPT-2 XL
  - GPT-3

Task: Rate the text on a 4-point scale (Before Training)

- If Option 1 is selected, ask "why did you select this ration"?
- Else, ask "What would you change to make it seem more human-like?"

---

Training techniques

1. Instruction-based training
2. Example-based training
3. Comparison-based training

Instruction-based training

We recommend you pay special attention to the following characteristics:

- **Repetition**: Machine-generated text often repeats words or phrases or contains redundant information.
- **Factuality**: Machine-generated text can contain text that is inaccurate or contradictory.

On the other hand, be careful with these characteristics, as they may be misleading:

- **Grammar and spelling**: While machine-generated text can contain these types of errors, human-authored text often contains them as well.
- **Style**: Current AI systems can generally mimic style fairly well, so a text that "looks right" or matches the expected style of the text isn't necessarily human-authored.
Example-based Training

Once upon a time, there was a man in a place that was not a place at all.

He didn't know anything of a place or a time or who he was or what he was doing there. There was just him and the silence.

He sat there for a long time, not knowing what he was doing there. He thought, thought and thought, but he didn't know what to think. There was just him and the silence. He tried to speak, but no sound came from his mouth. He tried to move, but his body would not move. He sat there, but he didn't know for how long he was there.

* What do you think the source of this text is?
  - Definitely human-written
  - Possibly human-written
  - Possibly machine-generated
  - Definitely machine-generated -- Correct Answer

You cannot change your answer once you click submit.

Explanation

Note how the story is repetitive and doesn’t seem to go anywhere.

Got it, next question
Comparison-based Training

## Results: without & with Training

<table>
<thead>
<tr>
<th>Training</th>
<th>Overall Acc.</th>
<th>Domain</th>
<th>Acc.</th>
<th>$F_1$</th>
<th>Prec.</th>
<th>Recall</th>
<th>Kripp. $\alpha$</th>
<th>% human</th>
<th>% confident</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.50</td>
<td>Stories</td>
<td>0.48</td>
<td>0.40</td>
<td>0.47</td>
<td>0.36</td>
<td>0.03</td>
<td>62.15</td>
<td>47.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>News</td>
<td>0.51</td>
<td>0.44</td>
<td>0.54</td>
<td>0.37</td>
<td>0.05</td>
<td>65.54</td>
<td>52.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recipes</td>
<td>0.50</td>
<td>0.41</td>
<td>0.50</td>
<td>0.34</td>
<td>0.00</td>
<td>66.15</td>
<td>50.62</td>
</tr>
<tr>
<td>Instructions</td>
<td>0.52</td>
<td>Stories</td>
<td>0.50</td>
<td>0.45</td>
<td>0.49</td>
<td>0.42</td>
<td>0.11</td>
<td>57.69</td>
<td>45.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>News</td>
<td>0.56</td>
<td>0.48</td>
<td>0.55</td>
<td>0.43</td>
<td>0.05</td>
<td>62.77</td>
<td>52.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recipes</td>
<td>0.50</td>
<td>0.41</td>
<td>0.52</td>
<td>0.33</td>
<td>0.07</td>
<td>67.69</td>
<td>49.85</td>
</tr>
<tr>
<td>Examples</td>
<td>*0.55</td>
<td>Stories</td>
<td>0.57</td>
<td>0.55</td>
<td>0.58</td>
<td>0.53</td>
<td>0.06</td>
<td>53.69</td>
<td>64.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>News</td>
<td>0.53</td>
<td>0.48</td>
<td>0.52</td>
<td>0.45</td>
<td>0.05</td>
<td>58.00</td>
<td>65.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recipes</td>
<td>0.56</td>
<td>0.56</td>
<td>0.61</td>
<td>0.51</td>
<td>0.06</td>
<td>55.23</td>
<td>64.00</td>
</tr>
<tr>
<td>Comparison</td>
<td>0.53</td>
<td>Stories</td>
<td>0.56</td>
<td>0.56</td>
<td>0.55</td>
<td>0.57</td>
<td>0.07</td>
<td>48.46</td>
<td>56.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>News</td>
<td>0.52</td>
<td>0.51</td>
<td>0.53</td>
<td>0.48</td>
<td>0.08</td>
<td>53.85</td>
<td>50.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recipes</td>
<td>0.51</td>
<td>0.49</td>
<td>0.52</td>
<td>0.46</td>
<td>0.06</td>
<td>54.31</td>
<td>53.54</td>
</tr>
</tbody>
</table>

Conclusion

- Both untrained and trained humans perform poorly
- Example-based training is the best
- We need better training and evaluation techniques
Is GPT-3 Text Indistinguishable from Human Text?
SCARECROW: A framework for scrutinizing machine text
Framework

1. A framework for scrutinizing deepfake texts through crowd annotation

2. A systematic way for humans to mark issues throughout the text and explain what is wrong
Crowd Annotations of Errors in Artificial vs. Human Texts

1. Language errors – lack of coherency & consistency in text
2. Factual errors - incorrect information in text
3. Reader issues -
4. text is too obscure or
   1. filled with too many jargon

# Error Types in the Scarecrow Framework

<table>
<thead>
<tr>
<th>ERROR TYPE</th>
<th>DEFINITION</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language Errors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar and Usage</td>
<td>Missing, extra, incorrect, or out of order words</td>
<td>...explaining how cats feel emoticons ...</td>
</tr>
<tr>
<td>Off-Prompt</td>
<td>Generation is unrelated to or contradicts prompt</td>
<td>PROMPT: Dogs are the new kids. GENERATION: Visiting the dentist can be scary</td>
</tr>
<tr>
<td>Redundant</td>
<td>Lexical, semantic, or excessive topical repetition</td>
<td>Merchants worry about poor service or service that is bad ...</td>
</tr>
<tr>
<td>Self-Contradiction</td>
<td>Generation contradicts itself</td>
<td>Amtrak plans to lay off many employees, though it has no plans cut employee hours.</td>
</tr>
<tr>
<td>Incoherent</td>
<td>Confusing, but not any error type above</td>
<td>Mary gave her kids cheese toast but drew a map of it on her toast.</td>
</tr>
<tr>
<td><strong>Factual Errors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad Math</td>
<td>Math or conversion mistakes</td>
<td>...it costs over £1,000 ($18,868) ...</td>
</tr>
<tr>
<td>Encyclopedic</td>
<td>Facts that annotator knows are wrong</td>
<td>Japanese Prime Minister Justin Trudeau said Monday ...</td>
</tr>
<tr>
<td>Commonsense</td>
<td>Violates basic understanding of the world</td>
<td>The dress was made at the spa.</td>
</tr>
<tr>
<td><strong>Reader Issues</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Needs Google</td>
<td>Search needed to verify claim</td>
<td>Jose Celana, an artist based in Pensacola, FL, ...</td>
</tr>
<tr>
<td>Technical Jargon</td>
<td>Text requires expertise to understand</td>
<td>...an 800-megawatt photovoltaic plant was built ...</td>
</tr>
</tbody>
</table>
LLMs

1. GPT-2 small
2. GPT-2 XL
3. GROVER Mega
4. GPT-3
Methods

- **Training**
  - Take an extensive qualification test
  - Test trains participants in categorization schemes
  - Pass participants if they score $\geq 90$ points out of 100 points
  - Pay participants $40$

- **Annotation:**
  - Participants annotate each paragraph using a custom annotation interface

- **Data Collection:**
  - Collect 13k human annotations of 1.3k paragraphs using SCARECROW, resulting in over 41k spans
SCARECROW Annotation interface

Select Model  Select Prompt

EXAMPLE ID: 782
Prompt (human-written):
Measuring foot traffic in real time is the pursuit of countless businesses seeking greater insight into buyer behavior.

Continuation written by GPT-3 (p = 0.96, temperature = 1.0, frequency penalty = 1):
The majority of that research currently revolves around online behavior, using tools like the Google Analytics or Omniture platforms. By monitoring this data closely, many companies have fine-tuned their business strategies — but the lessons learned from the digital world haven’t been reflected in much larger offline spaces. In retail foot traffic studies, for example, foot traffic is only counted as it passes through shop doorways, not tracked throughout a store’s interior spaces and around display cases. This is changing thanks to new indoor footfall sensors that use advanced 3D vision technologies to track more than 100 shoppers at once in shopping malls and department stores — areas where real-time indoor measurements are needed most.

https://yao-dou.github.io/scarecrow/
Figure 2: Average portion of tokens annotated with each error type (y-axis) across models (x-axis), with 95% confidence intervals. We group the trends into several broad categories. © Decreasing: fine-tuning and increasing model size improves performance. © Model plateau: increasing model size to GPT-3 does not correlate with further improvements. © Rising and falling: errors become more prevalent with some models, then improve. © Humans highest: these spans are labeled most on human-authored text; both are reader issues (distinct from errors; see Table 1). Details: all models, including GPT-3, use the same “apples-to-apples” decoding hyperparameters: top-p=0.96, temperature=1, and no frequency penalty.
Human Evaluation: Task

Fig. 1. (A) Example of a multi-authored (Human & Deepfake) 3-paragraph article; (B) Task: Detecting DeepFake texts; (C) Description of three research questions.

Training Technique: Example-based

## Recent Open source GPT-3 & ChatGPT detector

<table>
<thead>
<tr>
<th>Detector</th>
<th>Author</th>
<th>Link</th>
<th>Publish year</th>
</tr>
</thead>
<tbody>
<tr>
<td>DetectGPT</td>
<td>Stanford</td>
<td><a href="https://detectgpt.ericmitchell.ai/">https://detectgpt.ericmitchell.ai/</a></td>
<td>2023</td>
</tr>
<tr>
<td>GPTZero</td>
<td>Unknown</td>
<td><a href="https://gptzero.me/">https://gptzero.me/</a></td>
<td>2023</td>
</tr>
<tr>
<td>ChatGPT detector</td>
<td>OpenAI</td>
<td><a href="https://platform.openai.com/ai-text-classifier">https://platform.openai.com/ai-text-classifier</a></td>
<td>2023</td>
</tr>
<tr>
<td>ZeroGPT</td>
<td>Unknown</td>
<td><a href="https://www.zerogpt.com/">https://www.zerogpt.com/</a></td>
<td>2023</td>
</tr>
<tr>
<td>AI detector</td>
<td>Originality.AI</td>
<td><a href="https://originality.ai/?lmref=yjETBg">https://originality.ai/?lmref=yjETBg</a></td>
<td>2023</td>
</tr>
<tr>
<td>AI content detector</td>
<td>Copyleak</td>
<td><a href="https://copyleaks.com/features/ai-content-detector">https://copyleaks.com/features/ai-content-detector</a></td>
<td>2023</td>
</tr>
<tr>
<td>CheckGPT</td>
<td>ArticleBot</td>
<td><a href="https://www.app.got-it.ai/articlebot">https://www.app.got-it.ai/articlebot</a></td>
<td>2023</td>
</tr>
<tr>
<td>AI content detector</td>
<td>Sapling</td>
<td><a href="https://sapling.ai/utilities/ai-content-detector">https://sapling.ai/utilities/ai-content-detector</a></td>
<td>2023</td>
</tr>
<tr>
<td>ChatGPT detector</td>
<td>Writefull</td>
<td><a href="https://x.writefull.com/gpt-detector">https://x.writefull.com/gpt-detector</a></td>
<td>2023</td>
</tr>
<tr>
<td>ChatGPT detector</td>
<td>Draft &amp; Goal</td>
<td><a href="https://detector.dng.ai/">https://detector.dng.ai/</a></td>
<td>2023</td>
</tr>
</tbody>
</table>
Outline

1. Introduction & Generation – 20 minutes
2. Hands-on Game: 10 minutes
3. Detection – 30 minutes
4. Obfuscation – 25 minutes
5. Conclusion – 5 minutes

https://adauchendu.github.io/Tutorials/
Motivation

Can we make a deepfake text undetectable or conceal the authorship of a deepfake text by making small changes to the text while preserving semantics?
What make up the authorship of a text?

- Philosophical question: "The ship of Theseus"

- Deepfake obfuscation as a relaxation of "the ship of Theseus"

- or using detector as the ground-truth for meaningful changes


https://www.pastille.no/comics/ship-of-theseus
From Detection to Obfuscation

- **Detected as “Deepfake”** or “Machine-Generated” text

Prompt by human

White House floods during DC rainstorm on August 9

The White House is under water after a storm struck Washington DC on Wednesday. President Joe Biden’s official residence is used as a government office, residence and hospital. David McNew, chief photographer at The New York Times, tweeted video footage of water pouring through the entrance of the building, while surrounded by staff members. Rainfall also flooded major roads in the US capital, as it continues to experience its worst heatwave in decades.

Written by GPT-2

Pedestrians cross a flooded road in front of the White House in Washington DC, on August 9, 2022.
From Detection to Obfuscation

- Makes **(minimal) changes** to conceal authorship and preserving semantics

White House floods during **Washington DC** rainstorm on August 9

“...water **pouring through flooding to the** entrance...”

“...in **decades** the last 20 years...”

**White House floods during** **DC** **rainstorm on August 9**

The White House is under water after a storm struck **Washington DC** on Wednesday. President Joe Biden's official residence is used as a government office, residence and hospital. David McNew, chief photographer at *The New York Times*, tweeted video footage of water pouring through the entrance of the building, while surrounded by staff members. Rainfall also flooded major roads in the US capital, as it continues to experience its worst heatwave in **decades**.

Pedestrians cross a flooded road in front of the White House in Washington DC, on August 9, 2022.
Bolsonaro's re-election as Brazil's president is a matter of political preference and opinion. Some may argue that he should be re-elected because of his economic policies, efforts to combat corruption, and promotion of conservative values. Others may argue that he should not be re-elected due to his controversial statements and actions, disregard for environmental concerns, and human rights violations. It's important to note that Bolsonaro's handling of the COVID-19 pandemic has been heavily criticized, with Brazil having one of the highest death tolls globally. Additionally, his policies have been seen as exacerbating the already existing social inequalities in Brazil. Ultimately, whether Bolsonaro should be re-elected or not is up to the Brazilian people and their assessment of his performance and policies during his time in office.
Taxonomy – Obfuscation Technique

Authorship Obfuscation for Neural Texts

Stylistic Obfuscation

- Lexical Obfuscation
- Syntactic Obfuscation
- Morphological Obfuscation
- Orthographic Obfuscation

Word Choice
- Word Order
- Word Form
- Punctuation


Example of Obfuscation:
- Original: I have 99 problems but you won’t be one.
- Obfuscated: I have issues, 99 to be exact.
Taxonomy - Obfuscation Mechanism

- The **scenario** on which obfuscation is done (so-called threat model in security) is crucial
## Stylometric Obfuscation

- Current techniques tend to focus on **one or only a few linguistic feature(s)** to obfuscate – lexical, syntactical, etc.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Obfuscated Example</th>
<th>Stylometric Category</th>
<th>Preserves Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homoglyph</td>
<td>Hello -&gt; Hello</td>
<td>Orthographic</td>
<td>X</td>
</tr>
<tr>
<td>Upper/Lower Flip</td>
<td>Hello -&gt; heLo</td>
<td>Morphological</td>
<td>X</td>
</tr>
<tr>
<td>Misspellings attack</td>
<td>Acceptable -&gt; Acceptible</td>
<td>Lexical</td>
<td></td>
</tr>
<tr>
<td>Whitespace attack</td>
<td>Will face -&gt; Willface</td>
<td>Lexical</td>
<td></td>
</tr>
<tr>
<td>Deduplicate tokens</td>
<td>The car ... the money -&gt; the car ... money</td>
<td>Lexical</td>
<td></td>
</tr>
<tr>
<td>Shuffle tokens</td>
<td>Hello are -&gt; are hello</td>
<td>Syntactic</td>
<td></td>
</tr>
<tr>
<td>Mutant-X &amp; Avengers</td>
<td>What are the ramifications of this study? -&gt; What are the ramifications of this survey?</td>
<td>Lexical</td>
<td>X</td>
</tr>
<tr>
<td>ALISON</td>
<td>I got back my first draft of my memo -&gt; i had finished my first draft of the novel</td>
<td>Syntactic</td>
<td>X</td>
</tr>
</tbody>
</table>

Table: Examples of stylometric obfuscation techniques
Stylometric Obfuscation: PAN tasks [1]

- **Stylometric PAN’16 [2]**:
  - Apply text transformations (e.g., remove stop words, inserting punctuations, lower case) to push statistical metrics of each sentence **closer to those of the corpus average**
  - Statistics: avg # of words, #punctuation / #word token, #stop word / #word token, etc.

- **Sentence Simplification PAN’17 [3]**:
  - From: “**Basically, my job involves computer skills**”
  - To : “**My job involves computer skills**”

- **Back Translation NMTPAN’16 [4]**:
  - **English → IL₁ → IL₂ → ... ILₙ → English**
  - English → German → French → English
  - **IL: Intermediate Language**

---

Stylometric Obfuscation: *Mutant-X*

- Replacing words with **neighboring words** via sentiment-specific word embeddings (*customized word2vec*)
- Obfuscate text using **Genetic Algorithm** until (1) detector’s authorship changes + (2) semantic preserves

---

Stylometric Obfuscation: *Avengers*

- Obfuscations that are **transferable to unknown/blind** adversaries

- Surrogate model is designed as an **Ensemble** model

- Assume the same set of training features between obfuscator and detector

## Stylometric Obfuscation: Avengers

- Ensemble surrogate model **improves transferability**

<table>
<thead>
<tr>
<th>Surrogate Model</th>
<th>Attack Success Rate on Target Model</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RFC</td>
<td>SVM</td>
</tr>
<tr>
<td>RFC (Mutant-X)</td>
<td>28.2</td>
<td>26.2</td>
</tr>
<tr>
<td>SVM (Mutant-X)</td>
<td>1.6</td>
<td>93.7</td>
</tr>
<tr>
<td>Ensemble</td>
<td>18.4</td>
<td>61.0</td>
</tr>
</tbody>
</table>

Stylometric Obfuscation: DFTFooler

- Indirect obfuscation: **require no queries** to the detector, **no surrogate model**

- Utilize pre-trained LLM: substitute a subset of **most confidently predicted words** (green/yellow) with **lower confident synonyms** (red/purple)

- GLTR’s insights

---

Real-World Machine-Generated Text (GLTR.io)

With the ascendancy of Toni Morrison’s literary star, it has become commonplace for critics to de-racialize her by saying that Morrison is not just a Black woman writer, that she has moved beyond the limiting confines of race and gender to larger universal issues. Yet Morrison, a Nobel laureate with six highly acclaimed novels, bristles at having to choose between being a writer or a Black woman writer, and willingly accepts critical classification as the latter. To call her simply a writer denies the key roles that Morrison’s African-American roots and her Black female perspective have played in her work. For instance, many of Morrison’s characters treat their dreams as real, are nonplussed by visitations from dead ancestors, and

---

Human-Written Scientific Abstract (GLTR.io)
Option 1: train an **internal deepfake detector** and uses it to select texts with the highest human-class probability

Option 2: use the internal detector as **additional signal to guide beam-search** to generate more human-like texts (discriminative adversarial search [2])

\[
S_{DAS}(\hat{y}) = S_{dis}(\hat{y}) + \alpha \times S_{gen}(\hat{y})
\]
Statistical Obfuscation: *Changing decoding strategy*

- **Misalignment of decoding strategies** between detector and generator leads to lower detection performance => simple and effective.

- Many detectors witnessed **13.3%--97.6% degradation** in recall of machine-generated texts.

<table>
<thead>
<tr>
<th>Detector and Baseline Decoding</th>
<th>Top-p</th>
<th>Recall Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT (Top-p 0.96)</td>
<td>0.8</td>
<td>-13.3</td>
</tr>
<tr>
<td>GLTR-GPT2 (Top-k 40 + Temperature 0.7)</td>
<td>0.98</td>
<td>-97.6</td>
</tr>
<tr>
<td>GROVER (Top-p 0.94)</td>
<td>0.98</td>
<td>-35.6</td>
</tr>
<tr>
<td>FAST (Top-p 0.96)</td>
<td>1.0</td>
<td>-9.7</td>
</tr>
<tr>
<td>RoBERTa (Top-p 0.96)</td>
<td>1.0</td>
<td>-22.0</td>
</tr>
</tbody>
</table>
Stylometric Obfuscation: From Adversarial Texts

- **Original text:**
  - “You don't have to know about music to appreciate the film's easygoing blend of comedy and romance”

<table>
<thead>
<tr>
<th>AO technique</th>
<th>Obfuscated text Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextFooler [1]</td>
<td>You don't have to know about music to acknowledging the film's easygoing mixtures of mockery and ballad</td>
</tr>
<tr>
<td>DeepWordBug [2]</td>
<td>You don't have to know about music to appreciate the film's easygoing blend of comedy and romance</td>
</tr>
<tr>
<td>Perturbation-in-the-Wild [3]</td>
<td>You don’t have to know about music to appreciate the film’s easygoing blend of comedy and romance</td>
</tr>
</tbody>
</table>

---

CS + Linguistics => Deepfake Obfuscation

Computer Science
- Speed
- Efficiency
- Transfer-ability

Linguistics
- Writer Profiling
- Writing Structure
- Stylometry

The *dem0cr@ts* are pitiful
The *demòcr@ts* are pitiful
The *democrats* are pitiful

Text Bugger (Machine)
VIPER (Machine)
DeepWordBug (Machine)
Outline

1. Introduction & Generation – 20 minutes
2. Hands-on Game: 10 minutes
3. Detection – 30 minutes
4. Obfuscation – 25 minutes
5. Conclusion – 5 minutes

https://adauchendu.github.io/Tutorials/
Open Problems & Challenges

DETECTION

1. Novel training for Human-based Evaluation
2. Explainable AA
3. Harder AA problems
4. Robust AA

<table>
<thead>
<tr>
<th>Multi-author dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-domain dataset</td>
</tr>
<tr>
<td>Multi/Cross-modal dataset</td>
</tr>
</tbody>
</table>

OBFUSCATION

1. Human-based Evaluation
2. Explainable AO
3. AO that preserves semantics
4. Robust AO
5. Hybrid-based AO
## Future of Deepfake Detection: "LLM utopia"

- A pattern in text that is hidden to human naked eyes but algorithmically identifiable as machine-generated
- Enable rigorous statistical significance test

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Num tokens</th>
<th>Z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No watermark</td>
<td>56</td>
<td>.31</td>
<td>.38</td>
</tr>
<tr>
<td>Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With watermark</td>
<td>36</td>
<td>7.4</td>
<td>6e-14</td>
</tr>
<tr>
<td>- minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

- LLMs are ubiquitous, continuously improving

- **Arms Race** among Generators, Detectors and Obfuscators

- We need to study them altogether