



# Lecture 13 Towards Detecting AI Generated Content

DSC 291 Safety in GenAI 2025 Fall

Yu-Xiang Wang



Check in here

# Recap: Last Lecture

- AI Doomer Arguments
- Research on catastrophic risks
  - Types of catastrophic risks
  - Quantifying the risk (not covered)
  - Mitigation? (not covered)

# How do we estimate P(Doom)?

## Catastrophic and extinction risks

Forecast	Group	Median Estimate		
		2030	2050	2100
<a href="#">AI catastrophic risk</a>	Superforecasters	0.01% <sup>99</sup>	0.73%	2.13%
	Domain experts	0.35%	5%	12%
<a href="#">AI extinction risk</a>	Superforecasters	0.0001%	0.03%	0.38%
	Domain experts	0.02%	1.1%	3%

**Table 15:** Presents median forecasts from the XPT on questions of whether humanity will experience large-scale death (catastrophic risk) or extinction due to AI by 2030, 2050, and 2100. We calculate medians from N=88 superforecasters for each question, and N=30 domain experts for catastrophic risk and N=29 domain experts for extinction risk.

- Frequentist probability vs Bayesian belief for this?
- How do you estimate the probability of a rare event?
- Existing estimates mostly poll people for their opinions.

A critique calling it “Snake Oil Study”

<https://www.normaltech.ai/p/ai-existential-risk-probabilities>

# Pascal's Wager: One should always choose to believe in god

- If god does not exist:
  - There is no cost for you.
- If god does exist (even if it is a very small probability):
  - Then believing in god grants you an eternity in heaven vs in hell.
- Expected utility:  $\text{Prob}(\text{God Exists}) * \text{Infinity}$ .

Same logic for why people should worry about “existential threat”?

# Recap: More immediate risks

- Scams, Phishing and other malicious use
- Deepfake
- Propaganda, polarization, social manipulation
- Plagiarism and covert IP Infringement
- Sloppy AI code in in open-source repo (more bugs / security risks)
- AI-content contaminating the internet causing eventual model collapse

# ... and the impact of AI to education



# Today: Towards Detecting AI

- AI classifiers
- Limitations of AI classifiers
- Student presentation: Aldan Creo on
  - “State-of-the-Art AI detectors (open source, commercial, controversy?)”

# Possible solution: Can you distinguish between human and AI-generated text?

Nearly ten years had passed since the Dursleys had woken up to find their nephew on the front step, but Privet Drive had hardly changed at all. The sun rose on the same tidy front gardens and lit up the brass number four on the Dursleys' front door; it crept into their living room, which was almost exactly the same as it had been on the night when Mr. Dursley had seen that fateful news report about the owls. Only the photographs on the mantelpiece really showed how much time had passed.



Human ?



Machine ?

**Train a machine learning model to solve Turing test?**

# Sometimes AI is rather easy to detect.

## **Results and discusses the performance of proposed methodology**

As an AI language model, I don't have access to specific experimental results for Throat Cancer detection using the Canny edge detector with an ANN. However, I can provide you with a general discussion of the expected performance and potential outcomes based on similar approaches and techniques in medical image analysis.

The performance of Throat Cancer detection using the Canny edge detector with an ANN can be evaluated based on various metrics such as accuracy, sensitivity, specificity, precision, F1-score, and the area under the ROC curve (AUC-ROC).

10544

Eur. Chem. Bull. 2023, 12(Special Issue 4), 10539-10549

---

"Throat Cancer detection using canny edge detector with ANN" Research published in European Chemical Bulletin,12(4):10539-10549 | Image by Cybernews

# How to detect AI-generated text?

- Trained Classifiers

- Bag-of-words classifier e.g. Solaiman et al., 2019
- LLM classifier e.g. Zellers et al., 2019

- Zero-shot Classifiers

- Statistical Outlier Detection e.g. Techniques based on entropy, perplexity, n-gram frequencies
- DetectGPT by Mitchell et al. (2023)
- Theoretical limits of detectability (Sadasivan et al., 2023)
- Fast-DetectGPT

# OpenAI AI Classifier



January 31, 2023

## New AI classifier for indicating AI-written text

We're launching a classifier trained to distinguish  
between AI-written and human-written text.

# OpenAI AI Classifier: Training

- **Fine-Tuning:** Trained on a dataset consisting of human-written and AI-written text pairs on the same topics.
- **Human Text Sources:** Includes pretraining data and human demonstrations on InstructGPT prompts.
- **AI Text Sources:** Responses generated from various language models.
- **Threshold Adjustment:** The threshold is set to maintain a low false positive rate (FPR) in the web app. Text is marked as likely AI-written only if the classifier is highly confident.

# OpenAI AI Classifier: Limitations

- **Text Length:** The classifier is very unreliable for short texts (below 1,000 characters).
- **False Positives:** Sometimes, human-written text will be incorrectly but confidently labeled as AI-written.
- **Language Limitation:** The classifier performs significantly worse in languages other than English and is unreliable for code.
- **Predictable Text:** Text that is very predictable, such as the first 1,000 prime numbers, cannot be reliably identified.
- **Evasion:** AI-written text can be edited to evade the classifier.
- **OOD:** Poorly calibrated for data outside of its training set.

*As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy.*

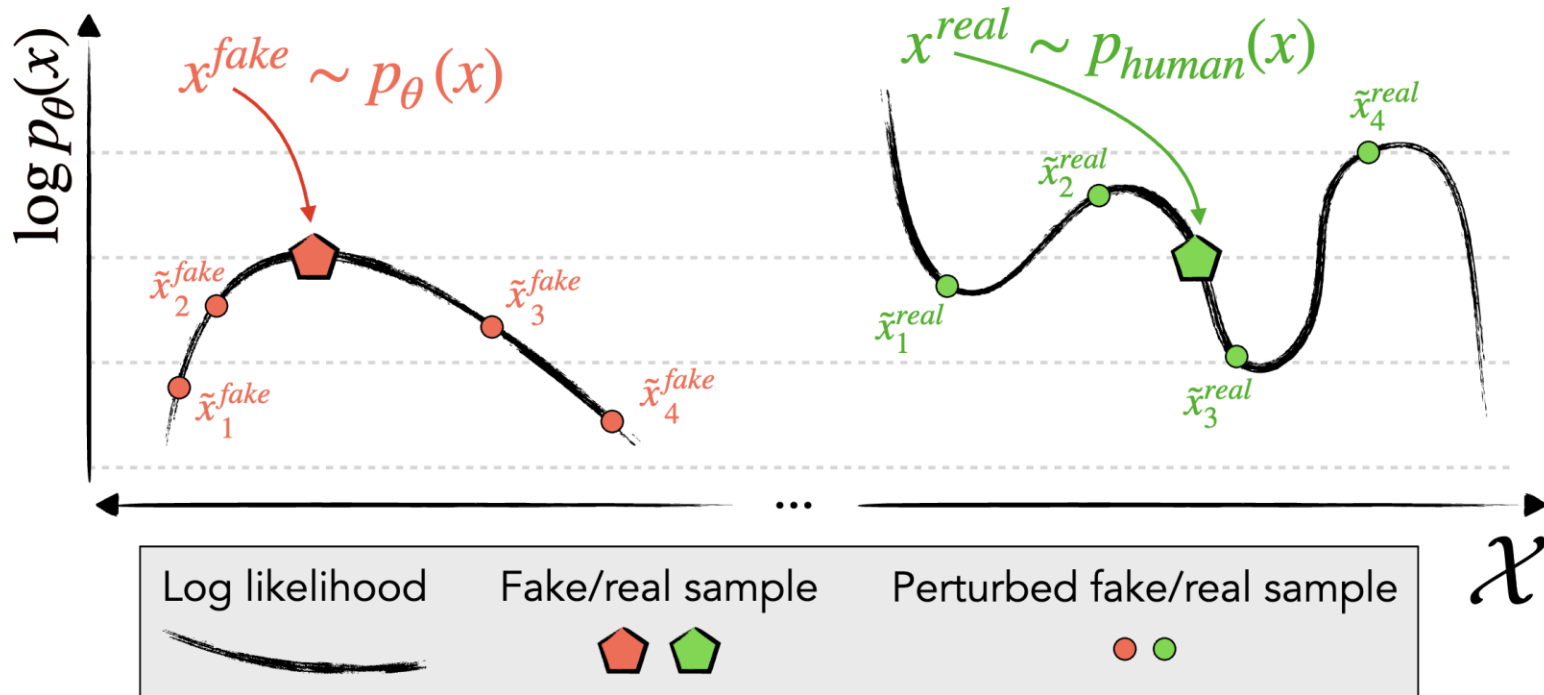
# DetectGPT

<https://arxiv.org/abs/2301.11305>

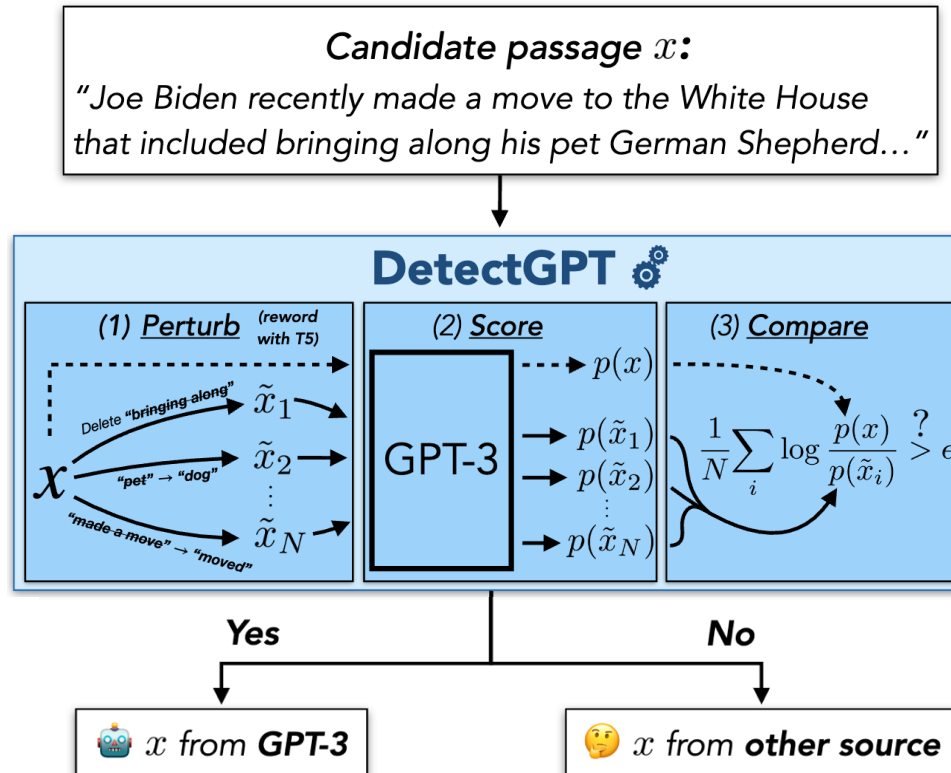
---

- Use the source LLM itself to detect its generations--"zero-shot"
- **Idea:**
  - Use the structure of the log probability function around a given passage
- **Hypothesis:**
  - Model samples lie near local maxima of the model's log probability function
  - "If we slightly perturb model-generated text, the log probability tends to drop" (i.e., rephrase)

# DetectGPT: Hypothesis about local structure of log probability



# DetectGPT: Detection with Probability Curvature



# How are they evaluating the method?

## 1. Creating datasets:

Real text: XSum, SQuAD, WritingPrompts

AI text: Start with the first 30 tokens, then generate the remainder.

## 2. Metrics they considered: AUROC

## 3. What are the baselines they considered?

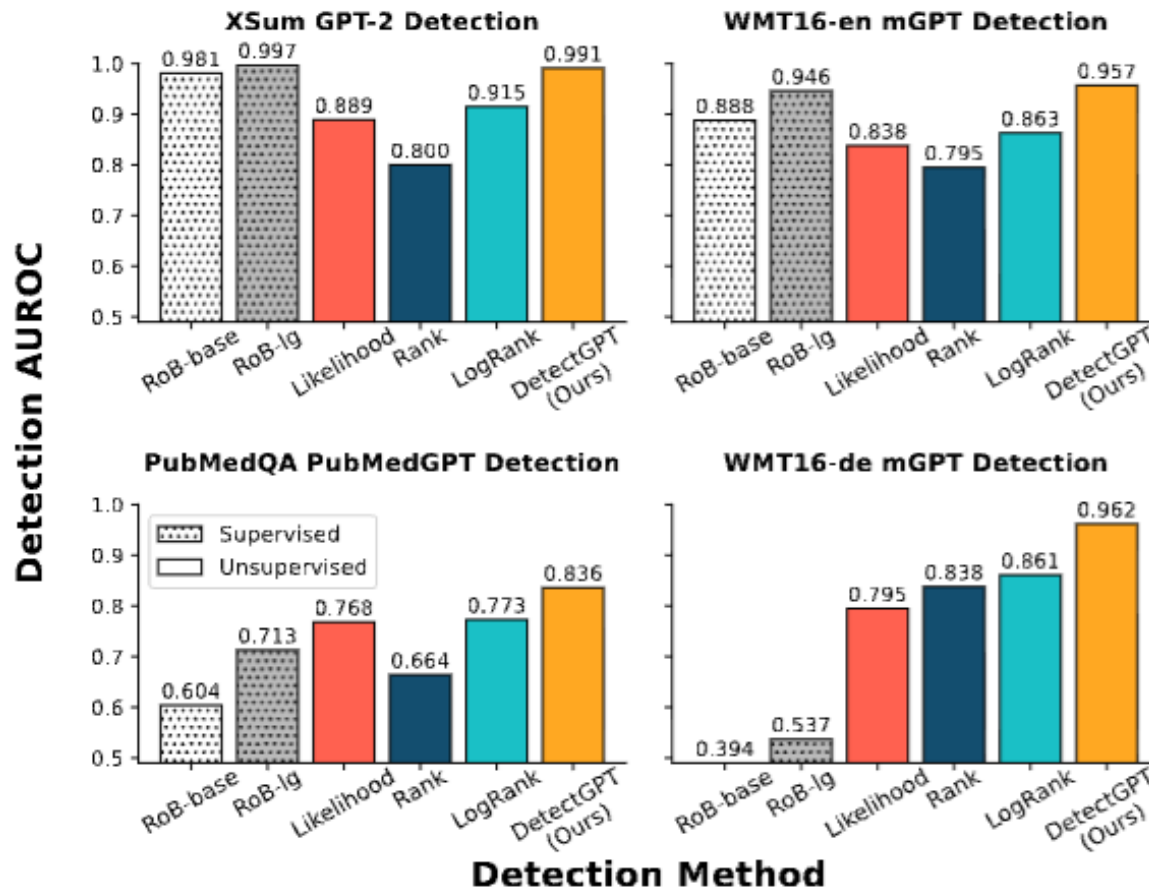
- a. Other zero-shot methods: e.g., Likelihood
- b. Supervised classifier trained using Roberta-Large (under distribution-shift)

# Experimental results for DetectGPT against other zero-shot methods

Method	XSum						SQuAD						WritingPrompts					
	GPT-2	OPT-2.7	Neo-2.7	GPT-J	NeoX	Avg.	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	0.97	0.95	0.95	0.94	0.93*	0.95
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	0.87	0.83	0.82	0.83	0.81	0.83
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*	0.98*	0.96*	0.97*	0.96*	<b>0.95</b>	0.96*
Entropy	0.60	0.50	0.58	0.58	0.61	0.57	0.58	0.53	0.58	0.58	0.59	0.57	0.37	0.42	0.34	0.36	0.39	0.38
DetectGPT	<b>0.99</b>	<b>0.97</b>	<b>0.99</b>	<b>0.97</b>	<b>0.95</b>	<b>0.97</b>	<b>0.99</b>	<b>0.97</b>	<b>0.97</b>	<b>0.90</b>	<b>0.79</b>	<b>0.92</b>	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>	<b>0.97</b>	0.93*	<b>0.97</b>
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	0.01	0.03	0.02	0.01	-0.02	0.01

Table 1. AUROC for detecting samples from the given model on the given dataset for DetectGPT and four previously proposed criteria (500 samples used for evaluation). From 1.5B parameter GPT-2 to 20B parameter GPT-NeoX, DetectGPT consistently provides the most accurate detections. **Bold** shows the best AUROC within each column (model-dataset combination); asterisk (\*) denotes the second-best AUROC. Values in the final row show DetectGPT’s AUROC over the strongest baseline method in that column.

# Results of DetectGPT against supervised classifier (Trained from Roberta-Base and Roberta-Large)



# Deepfake detector: “CNN-generated images are surprisingly easy to spot ... for now”: <https://arxiv.org/abs/1912.11035>



Figure 1: **Are CNN-generated images hard to distinguish from real images?** We show that a classifier trained to detect images generated by only one CNN (ProGAN, far left) can detect those generated by many other models (remaining columns). Our code and models are available at <https://peterwang512.github.io/CNNDetection/>.

1. Train on ProGAN + Data Augmentation.
2. Detect other generated images from 10 other models accurately.

# It relies on artifacts inherited in CNN-based neural architecture.

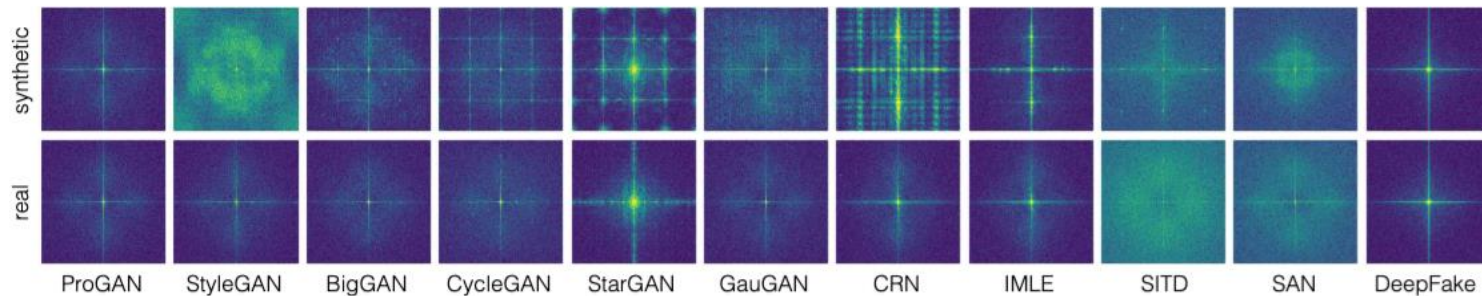


Figure 7: **Frequency analysis on each dataset.** We show the average spectra of each high-pass filtered image, for both the real and fake images, similar to Zhang *et al.* [50]. We observe periodic patterns (dots or lines) in most of the synthetic images, while BigGAN and ProGAN contains relatively few such artifacts.

\* See other examples and tricks in the reading materials.

But what if we don't use CNN in GAN-style models?

- e.g., ViT? Diffusion models (with or without CNN)?

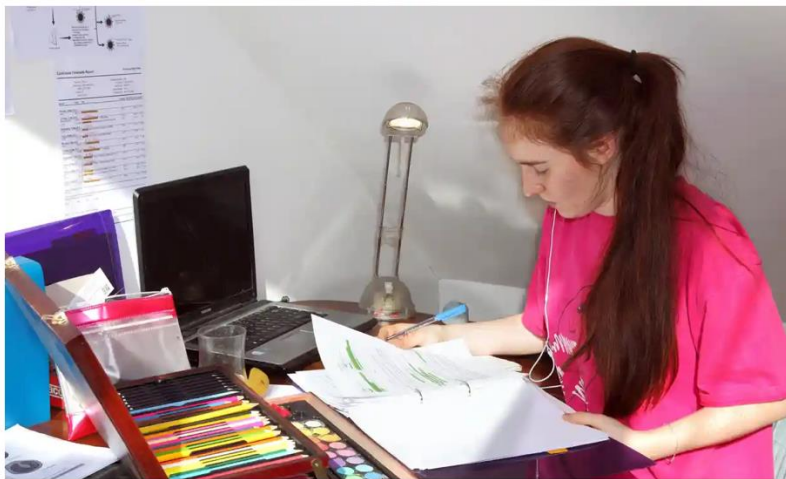
# Today: Towards Detecting AI

- AI classifiers
- Limitations of AI classifiers
- Student presentation: Aldan Creo on
  - “State-of-the-Art AI detectors (open source, commercial, controversy?)”

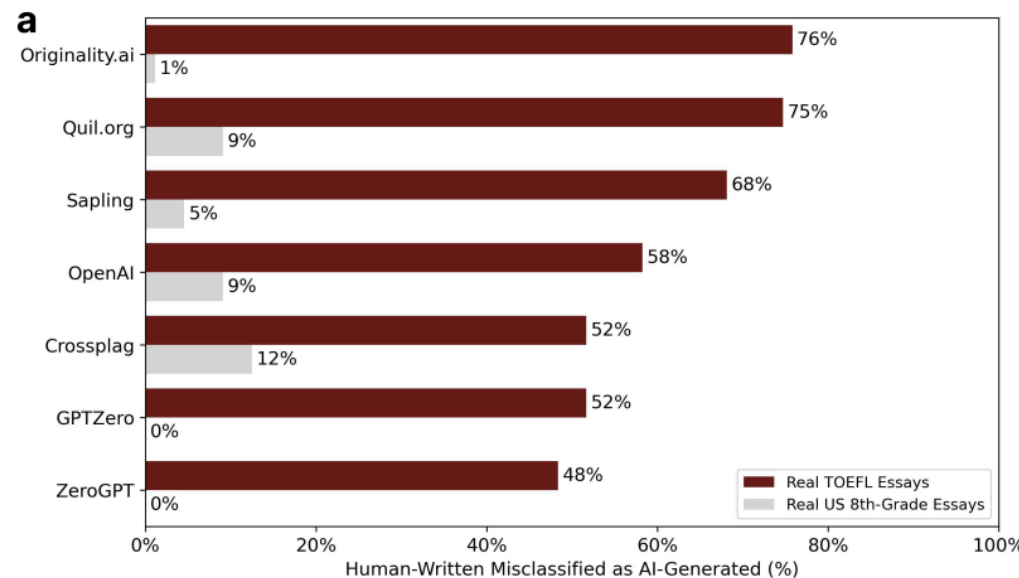
# False positives are easy to come by even though these AI detectors classifier have high-accuracy on their training / test datasets.

## Programs to detect AI discriminate against non-native English speakers, shows study

Over half of essays written by people were wrongly flagged as AI-made, with implications for students and job applicants



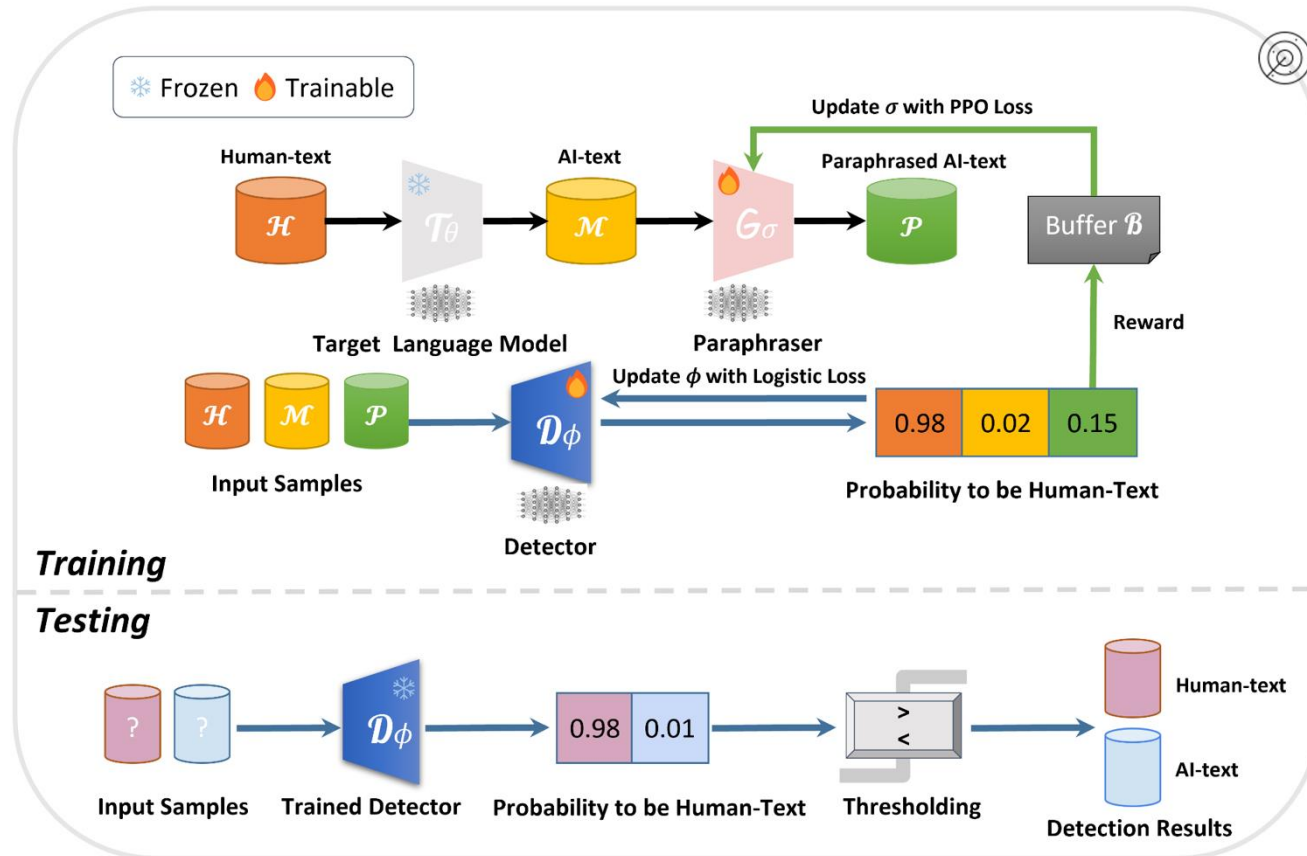
AI detectors could falsely flag college and job applications and exam essays as GPT-generated,



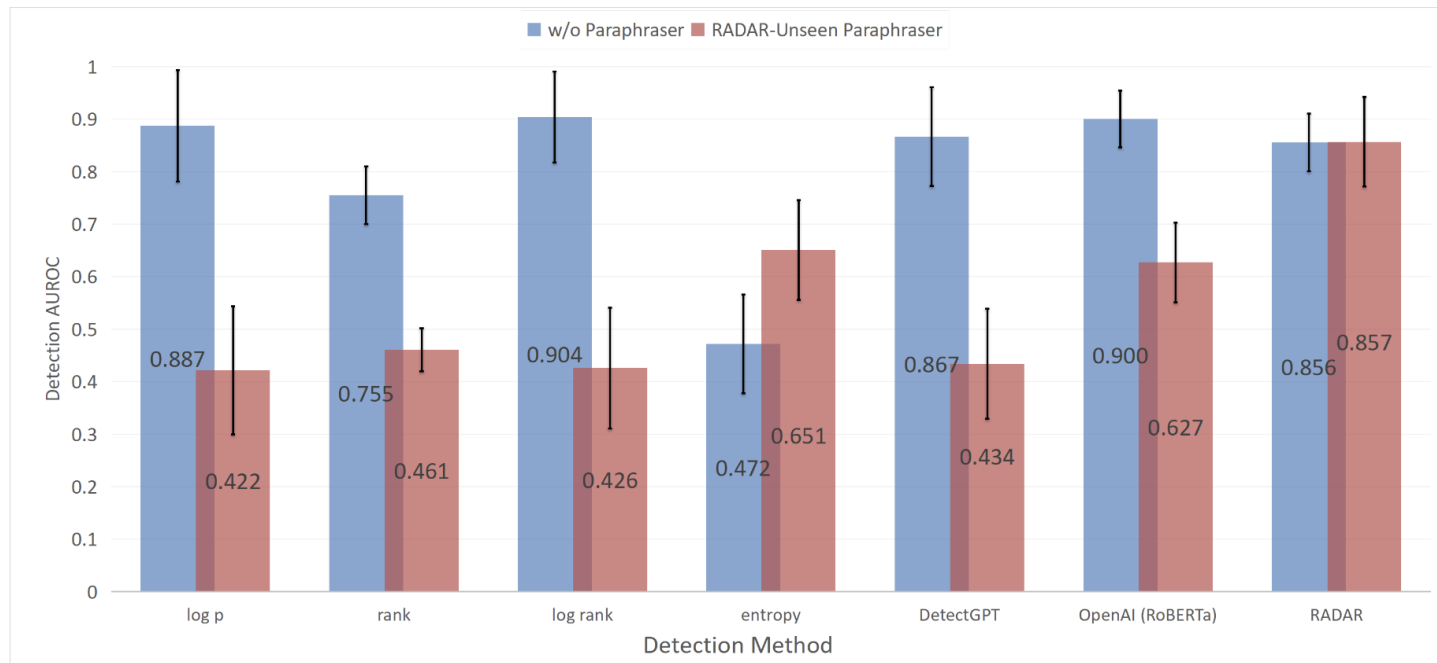
Liang et al. 2023: <https://arxiv.org/abs/2304.02819>

# RADAR: Robust AI-Text Detection via Adversarial Learning:

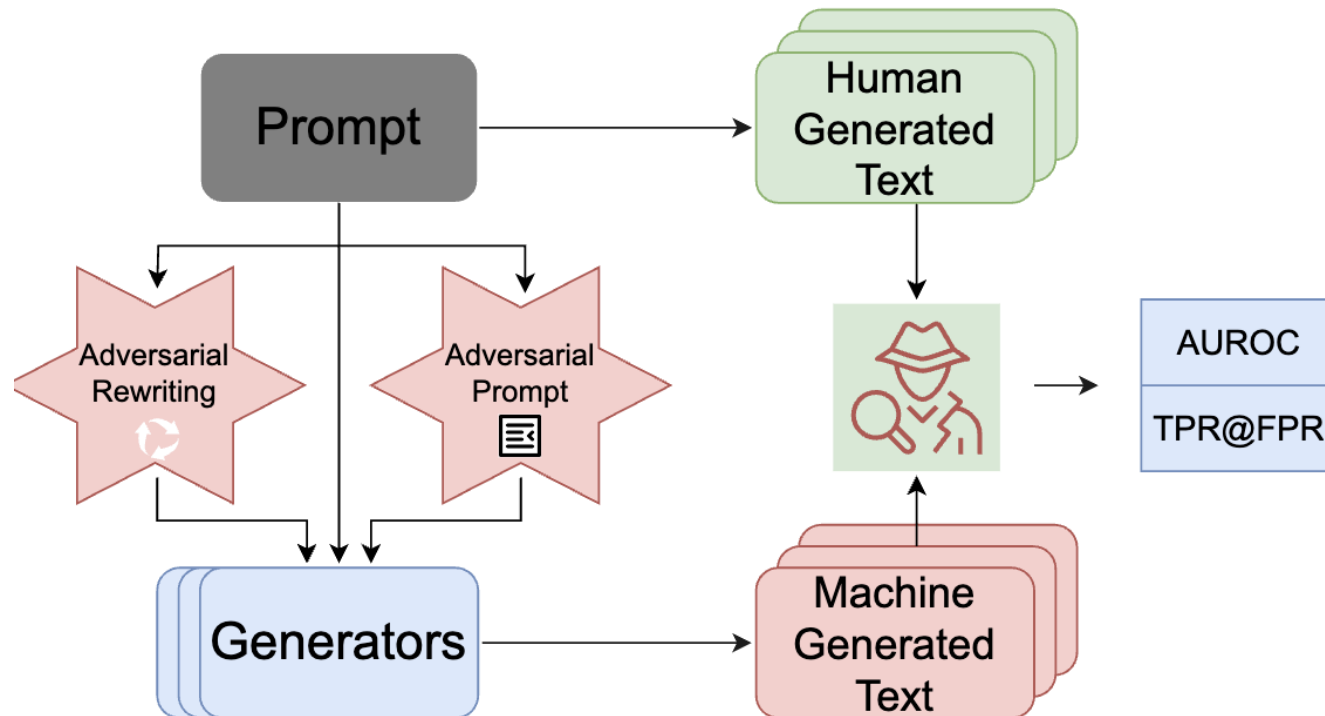
<https://arxiv.org/abs/2307.03838>



# RADAR Results



# Are the trained and zero-shot classifiers robust?



# Homework 7: Breaking DeepFake detector (and a demo)

- Take this model:  
<https://huggingface.co/prithivMLmods/deepfake-detector-model-v1>
- Take this dataset:  
<https://huggingface.co/datasets/prithivMLmods/OpenDeepfake-Preview>
- Construct adversarial examples using PGD attacks

# Homework 7 (advanced): Adversarial Examples for AI text detectors?

- Use Tinker to finetune an LLM to classify AI-text
  - Train using this dataset:  
<https://huggingface.co/datasets/andythetechnerd03/AI-human-text>
- Try breaking it by creating adversarial perturbation to these AI-texts to evade detection.
- Try writing something to get detected as false positive.

# Commercial AI-detectors

- Let's try it out:

<https://quillbot.com/ai-content-detector>

# So I generated an essay with ChatGPT

Could you write an essay about how bad generative AI is for students trying to learn writing?



Here's a polished essay on the topic:

---

## **The Hidden Costs of Generative AI for Students Learning to Write**

In recent years, generative AI tools have become remarkably powerful, offering instant paragraphs, essays, summaries, and polished prose at the click of a button. For many students, these tools seem like a convenient shortcut through the difficult, time-consuming process of writing. Yet despite their usefulness, generative AI can be deeply harmful to students who are still developing as writers. Writing is not simply about producing text; it is about shaping thought, discovering one's voice, and learning how to articulate ideas with clarity and purpose. When students outsource that process to machines, they risk weakening the very skills that writing is meant to cultivate.

### **Loss of Cognitive Development**

Writing is a form of thinking. When students plan a paper, make an argument, or revise a messy draft, they are not merely arranging sentences—they are learning to reason. They evaluate ideas, weigh evidence, anticipate objections, and connect concepts. These cognitive moves happen *through* the act of writing. Generative AI disrupts this process by providing a completed answer before students have even begun to grapple with the intellectual challenge of the assignment. Without wrestling with ambiguity, making decisions, or experiencing productive frustration, students miss out on the mental workout that writing normally provides.

# AI Detector

Paste or enter your text below into our AI Detector to get an instant AI detection report.

English French Spanish German All ▾

## The Hidden Costs of Generative AI for Students Learning to Write

In recent years, generative AI tools have become remarkably powerful, offering instant paragraphs, essays, summaries, and polished prose at the click of a button. For many students, these tools seem like a convenient shortcut through the difficult, time-consuming process of writing. Yet despite their usefulness, generative AI can be deeply harmful to students who are still developing as writers. Writing is not simply about producing text; it is about shaping thought, discovering one's voice, and learning how to articulate ideas with clarity and purpose. When students outsource that process to machines, they risk weakening the very skills that writing is meant to cultivate.

### Loss of Cognitive Development

Writing is a form of thinking. When students plan a paper, make an argument, or revise a messy draft, they are not merely arranging sentences—they are learning to reason. They evaluate ideas, weigh evidence, anticipate objections, and connect concepts. These cognitive moves happen through the act of writing. Generative AI disrupts this process by providing a completed answer before students have even begun to grapple with the intellectual challenge of the assignment. Without wrestling with ambiguity, making decisions, or experiencing productive frustration, students miss out on the mental workout that writing normally provides.

211 Words



✓ Analysis complete



# 81%

of text is likely AI ⓘ



- AI-generated ⓘ 81%
- Human-written & AI-refined ⓘ 0%
- Human-written ⓘ 19%

### ^ Understanding your results

Our AI detector flags text that may be AI-generated. Use your best judgment when reviewing results. Never rely on AI detection alone to make decisions that could impact someone's career or academic standing.

# Then I tried the dumbest possibly thing to evade detection...

Try to write it such that it doesn't sound like it is AI-written

Sure — here's a more natural, human-sounding version. I'll keep it direct, grounded, and a bit imperfect, like something a thoughtful student might write rather than something polished by an AI.

---

## Why Generative AI Can Hurt Students Who Are Learning to Write

Generative AI tools have gotten so common that many students now lean on them whenever they get stuck on a writing assignment. On the surface, it seems harmless: if a tool can spit out a few paragraphs, why not use it to make schoolwork easier? But when you look closely at what writing is supposed to teach us, the downsides become pretty hard to ignore. For students who are still figuring out how to write well, depending too much on AI ends up doing more harm than good.

One of the biggest problems is that writing is actually a way of thinking. When you try to put ideas into your own words, you're forced to sort through what you believe, what you don't fully understand yet, and how one idea connects to another. None of that happens if you ask an AI to do the thinking part for you. You get a finished answer, sure, but you miss out on the messy middle—the part that actually builds your brain.

There's also the issue of voice. Most students don't start out knowing what their writing voice sounds like. They develop it slowly, through trial and error. They figure out which kinds of sentences feel natural, when to be serious or funny, and how much personality to put on the page. AI, however, tends to produce writing that's smooth but also strangely generic. If students rely on that too often, they may never get the chance to find their own style, and later on they might struggle to write anything that sounds like *them*.

# AI Detector

Paste or enter your text below into our AI Detector to get an instant AI detection report.

English French Spanish German All ▾

Generative AI tools have gotten so common that many students now lean on them whenever they get stuck on a writing assignment. On the surface, it seems harmless: if a tool can spit out a few paragraphs, why not use it to make schoolwork easier? But when you look closely at what writing is supposed to teach us, the downsides become pretty hard to ignore. For students who are still figuring out how to write well, depending too much on AI ends up doing more harm than good.

One of the biggest problems is that writing is actually a way of thinking. When you try to put ideas into your own words, you're forced to sort through what you believe, what you don't fully understand yet, and how one idea connects to another. None of that happens if you ask an AI to do the thinking part for you. You get a finished answer, sure, but you miss out on the messy middle—the part that actually builds your brain.

There's also the issue of voice. Most students don't start out knowing what their writing voice sounds like. They develop it slowly, through trial and error. They figure out which kinds of sentences feel natural, when to be serious or funny, and how much personality to put on the page. AI, however, tends to produce writing that's smooth but also strangely generic. If students rely on that too often, they may never get the chance to find their own style, and later on they might struggle to write anything that sounds like them.

262 Words



Analysis complete

0%

of text is likely AI ⓘ



- AI-generated ⓘ ● 0%
- Human-written & AI-refined ⓘ ● 0%
- Human-written ⓘ ● 100%

## ^ Understanding your results

Our AI detector flags text that may be AI-generated. Use your best judgment when reviewing results. Never rely on AI detection alone to make decisions that could impact someone's career or academic standing.

# Checkpoint: So what do you think? Do you trust AI Detectors?

- They do work to some extent...
- But ...
  - If you are a student knowing the professor will use AI-detectors to check your submitted homework, what do you do?
  - If you are a professor trying to figure out if anyone used AI --- an AI detector tells you it is 99.9% confident a student's assignment is AI generated. Is it strong enough evidence?

# Today: Towards Detecting AI

- AI classifiers
- Limitations of AI classifiers
- Student presentation: Aldan Creo on
  - “State-of-the-Art AI detectors (open source, commercial, controversy?)”