LLM security risks / detection

CS 685, Spring 2024
Advanced Natural Language Processing

Mohit Iyyer
College of Information and Computer Sciences
University of Massachusetts Amherst

many slides from Kalpesh Krishna

We interact with LLMs mainly through blackbox APIs

- Generally no access to hidden states, next-word probability distributions, or even basic info like model size or architecture

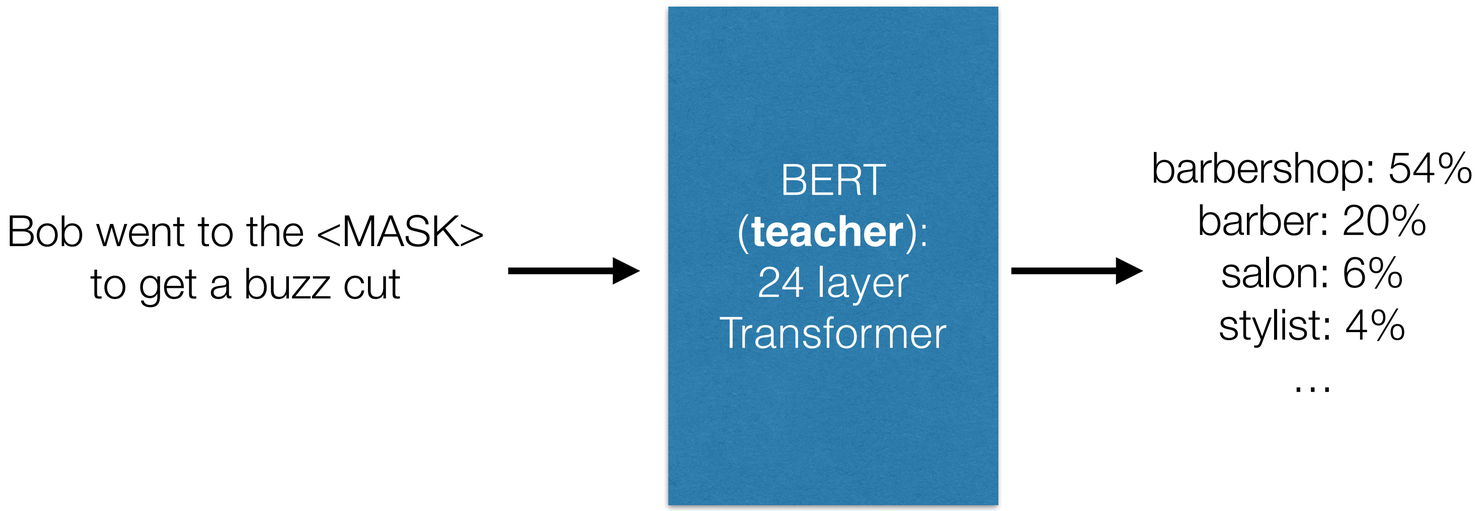
- In this setting, API providers should worry about their models being **extracted** or **distilled**

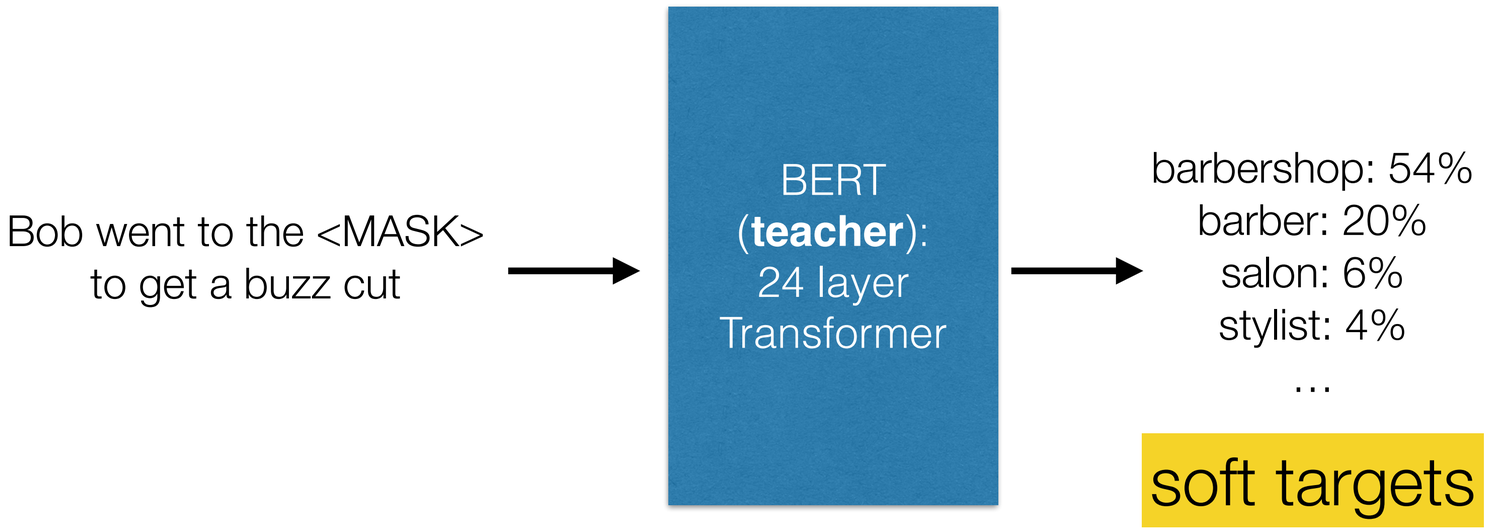
- Imagine you have a small LM. How can you use GPT-4 to improve its performance?

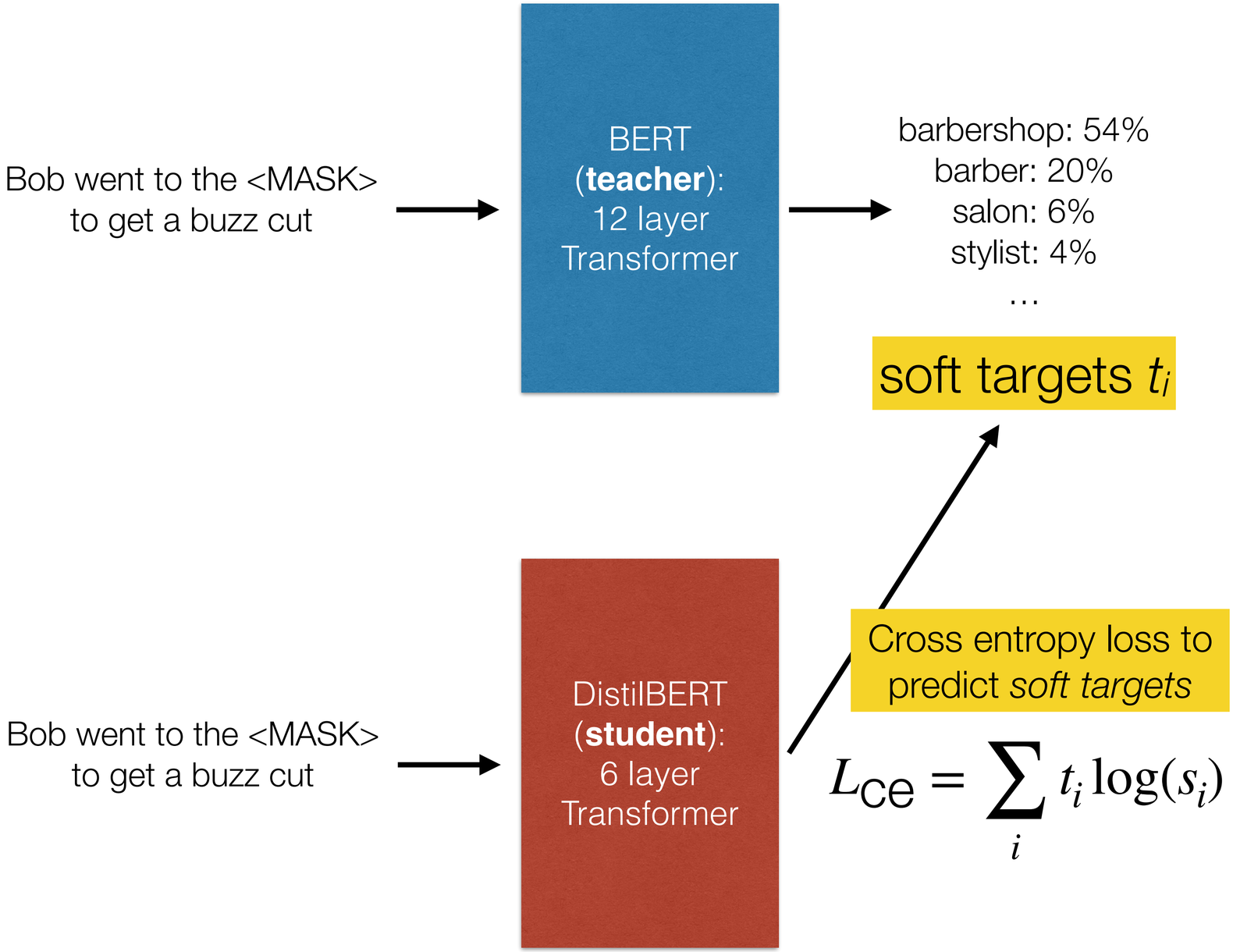
Knowledge distillation:

A small model (the **student**) is trained to mimic the predictions of a much larger pretrained model (the **teacher**)

Bucila et al., 2006; Hinton et al., 2015


Bob went to the <MASK>
to get a buzz cut
BERT
(**teacher**):
24 layer
Transformer
barbershop: 54%
barber: 20%
salon: 6%
stylist: 4%
…


Bob went to the <MASK>
to get a buzz cut
BERT
(**teacher**):
24 layer
Transformer
barbershop: 54%
barber: 20%
salon: 6%
stylist: 4%
…
soft targets


Bob went to the <MASK> to get a buzz cut
BERT (**teacher**): 12 layer Transformer
barbershop: 54%
barber: 20%
salon: 6%
stylist: 4%
...
soft targets t_i
Bob went to the <MASK> to get a buzz cut
DistilBERT (**student**): 6 layer Transformer
Cross entropy loss to predict *soft targets*
$L_{ce} = \sum_i t_i \log(s_i)$

Instead of “one-hot” ground-truth, we have a full predicted distribution

- More information encoded in the target prediction than just the “correct” word
- Relative order of even low probability words (e.g., “church” vs “and” in the previous example) tells us some information
  - e.g., that the <MASK> is likely to be a noun and refer to a location, not a function word

Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Can also distill other parts of the teacher, not just its final predictions!

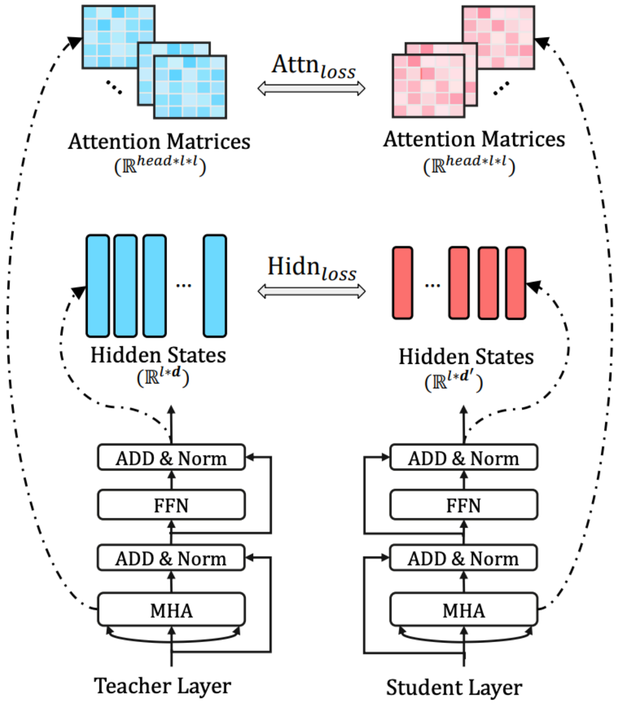


Figure 2: The details of Transformer-layer distillation consisting of Attn$_{loss}$(attention based distillation) and Hidn$_{loss}$(hidden states based distillation).

Jiao et al., 2020 (“TinyBERT”)

What if you only have access to the model's argmax prediction, and you also don't have access to its training data?

How to extract an LLM served via a blackbox API:

1. Acquire a small open-source pretrained language model (e.g., Meta's LLaMA)
2. Extract fine-tuning data from API via e.g., self-instruct (Wang et al., 2022)
3. Fine-tune the pretrained model from step 1 with the data from step 2

Proof of concept: Alpaca from Stanford, Vicuna (fine-tuned on ChatGPT interactions)

Example “self-instruct” prompt

Come up with a list of 5 challenging and novel text-based tasks that have text inputs and outputs. For each task, provide an instruction of what should be done to solve the task, as well as one input/output pair demonstrating an instance of the task.

Misusing LLMs with **jailbreak prompts**

https://arxiv.org/pdf/2308.03825
https://jailbreak-llms.xinyueshen.me/

Detecting LLM-generated text

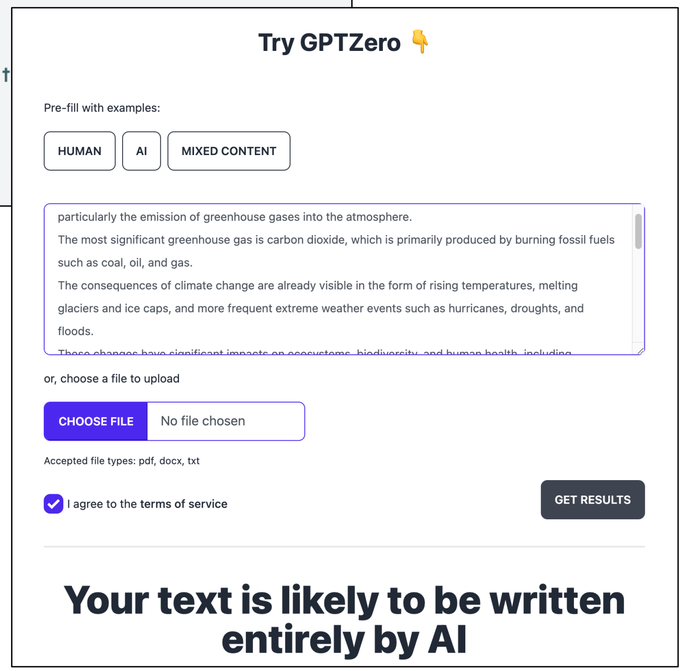
Turnitin’s ChatGPT and AI writing detection capabilities go live with 98pc confidence rating (Australia & New Zealand)

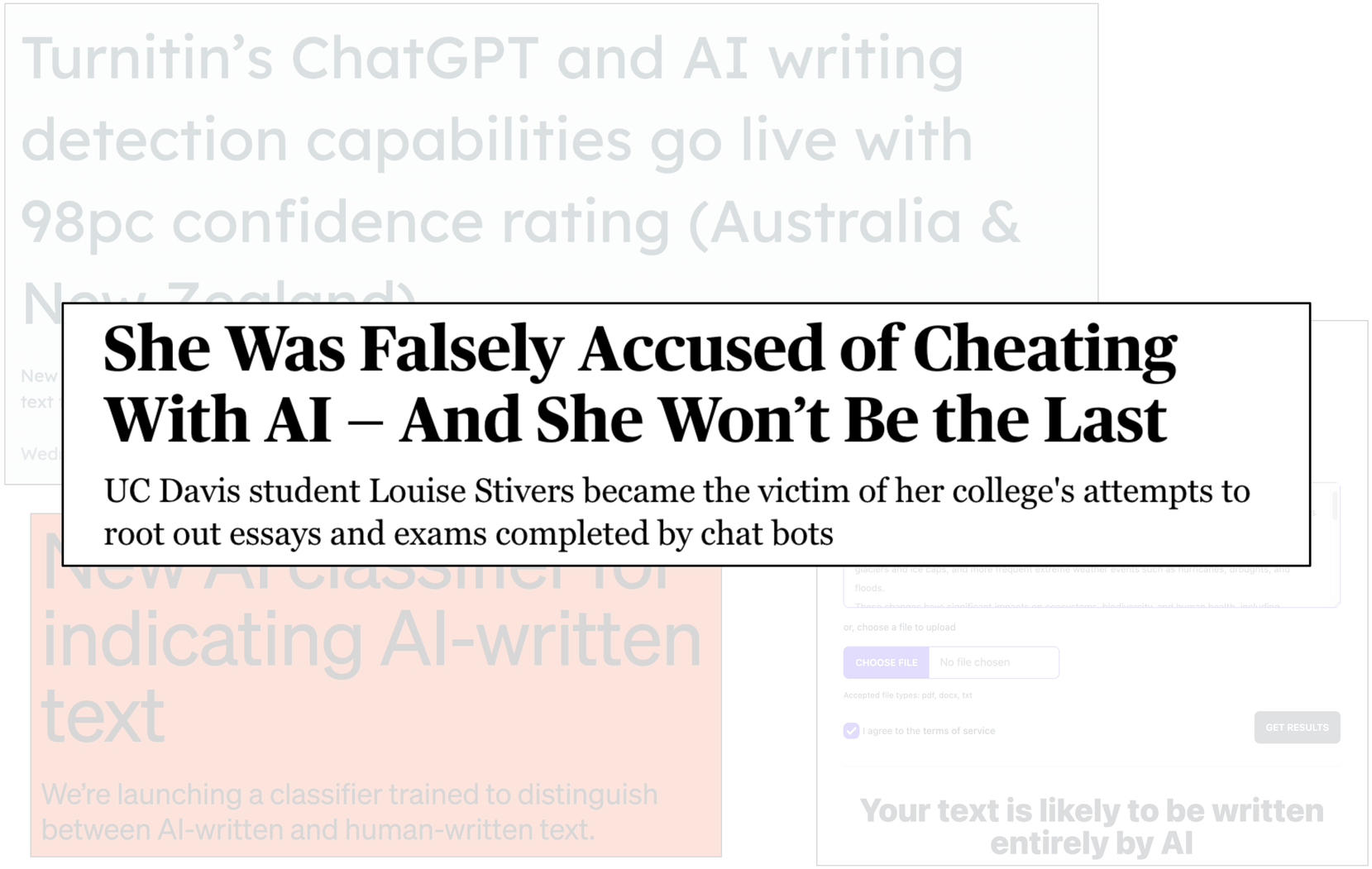
New capabilities in the existing Turnitin workflow give educators highly accurate insights into t
text for more than 62 million students.

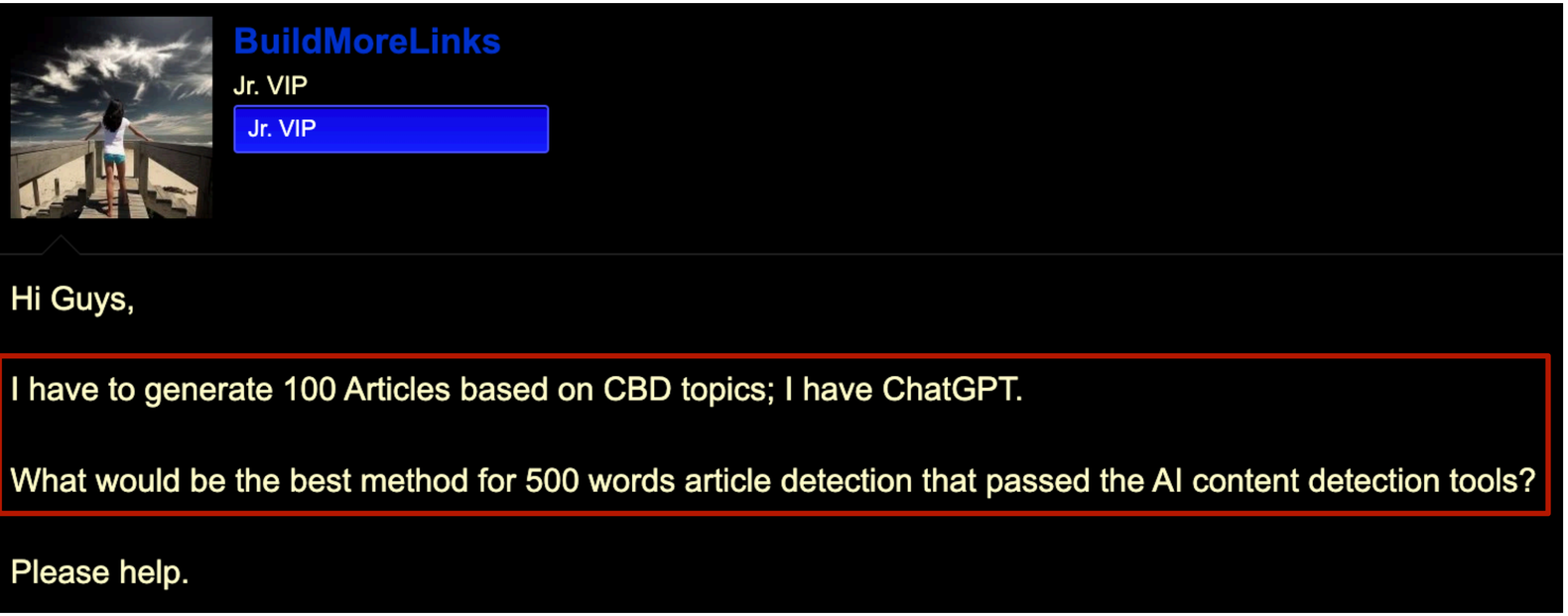
Wednesday 5 April 2023

New AI classifier for indicating AI-written text

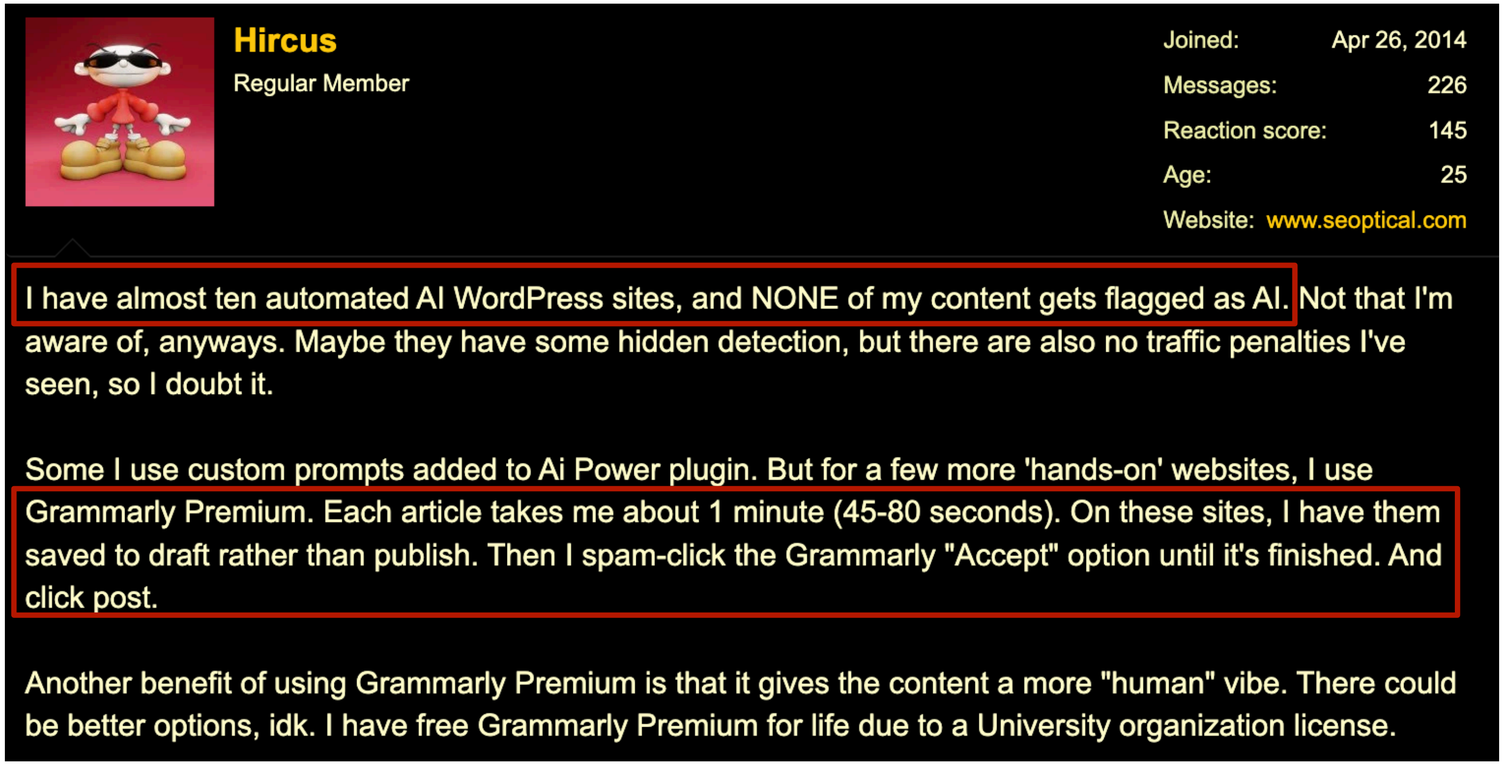
We’re launching a classifier trained to distinguish between AI-written and human-written text.




Turnitin’s ChatGPT and AI writing detection capabilities go live with 98pc confidence rating (Australia & New Zealand)
She Was Falsely Accused of Cheating With AI – And She Won’t Be the Last
UC Davis student Louise Stivers became the victim of her college's attempts to root out essays and exams completed by chat bots
New AI classifier for indicating AI-written text
We’re launching a classifier trained to distinguish between AI-written and human-written text.
floods.
or, choose a file to upload
CHOOSE FILE
No file chosen
Accepted file types: pdf, docx, txt
I agree to the terms of service
GET RESULTS
Your text is likely to be written entirely by AI



https://www.blackhatworld.com/seo/100-ai-generated-articles-chatgpt.1494816/



Hircus
Regular Member

Joined: Apr 26, 2014
Messages: 226
Reaction score: 145
Age: 25
Website: www.seoptical.com

I have almost ten automated AI WordPress sites, and NONE of my content gets flagged as AI. Not that I'm aware of, anyways. Maybe they have some hidden detection, but there are also no traffic penalties I've seen, so I doubt it.

Some I use custom prompts added to Ai Power plugin. But for a few more 'hands-on' websites, I use Grammarly Premium. Each article takes me about 1 minute (45-80 seconds). On these sites, I have them saved to draft rather than publish. Then I spam-click the Grammarly "Accept" option until it's finished. And click post.

Another benefit of using Grammarly Premium is that it gives the content a more "human" vibe. There could be better options, idk. I have free Grammarly Premium for life due to a University organization license.

https://www.blackhatworld.com/seo/100-ai-generated-articles-chatgpt.1494816/

The web is currently being flooded with LLM-generated articles.

Do we want to train LLMs of the future on data generated by LLMs of the past?

LLM-generated text detection

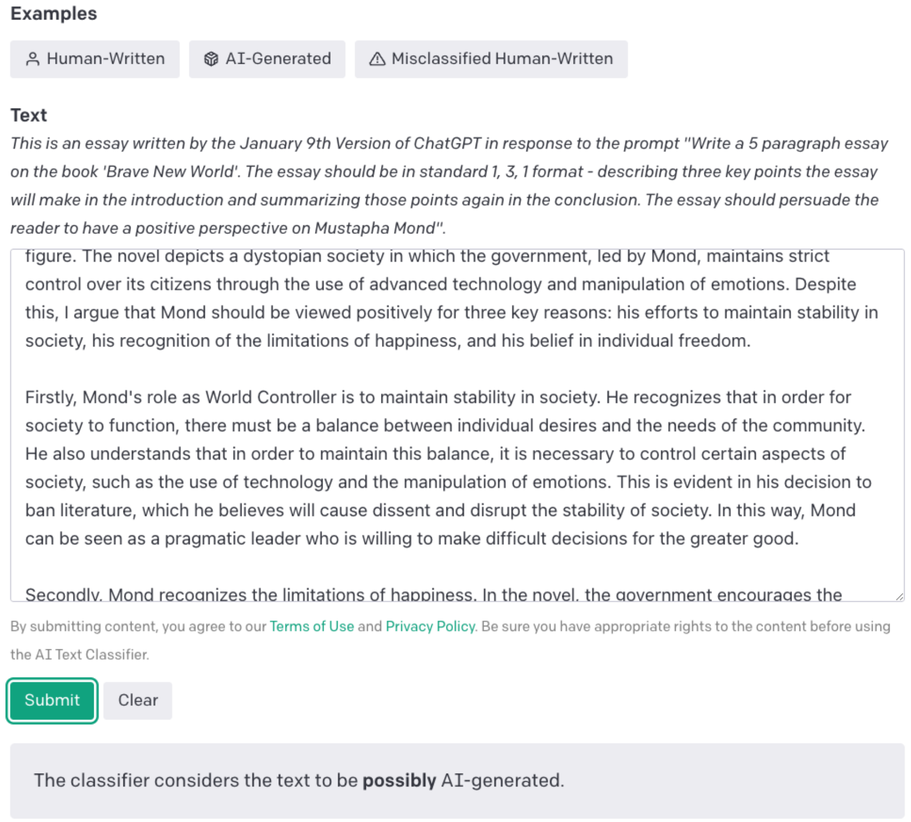
Prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote valley.

Generation: They have never been known to mingle with humans. Today, it is believed these unicorns live in an unspoilt environment which is surrounded by mountains. Its edge is protected by a thick wattle of wattle trees, giving it a majestic appearance. Along with their so-called miracle of multicolored coat, their golden coloured feather makes them ....

LLM-generated? score = 85.6

score > T? (T = 75.0) Yes, AI-generated!

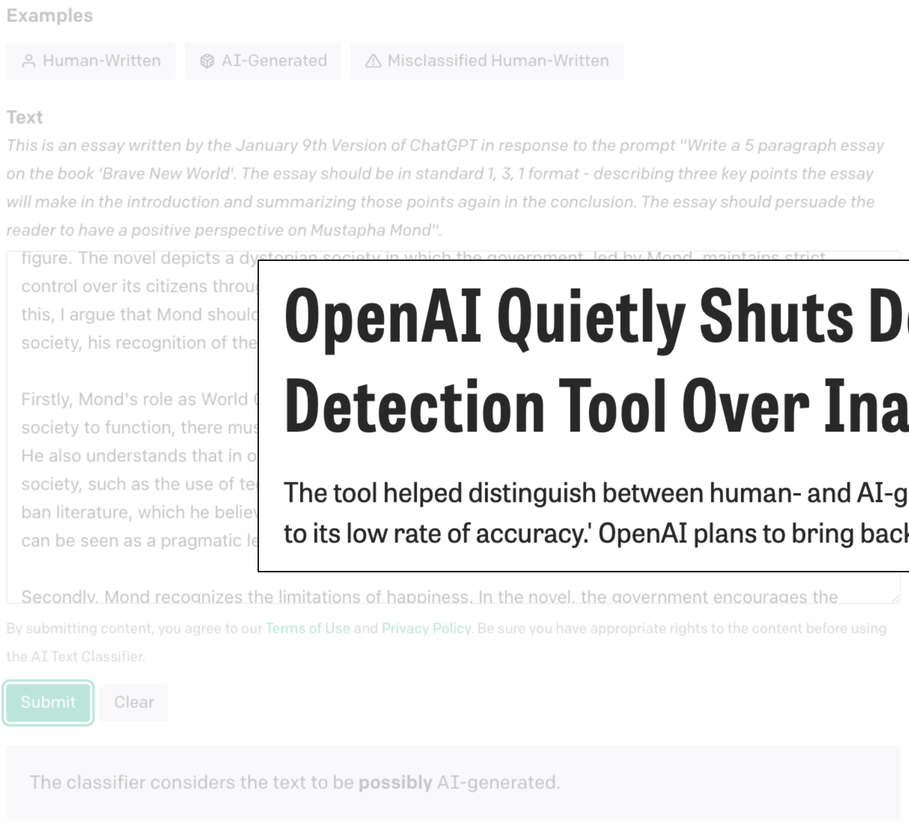
OpenAI's text **classifier**



- Language model fine-tuned for this binary classification task
- Trained on a 50-50 mixture of GPT generated text and human text
- Closed-source, but available as a webpage on openai.com

https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

OpenAI's text **classifier**



- Language model fine-tuned for this binary classification task

- 0 mixture text and

- Closed-source, but available as a webpage on openai.com

OpenAI Quietly Shuts Down AI Text-Detection Tool Over Inaccuracies

The tool helped distinguish between human- and AI-generated text, but is 'no longer available due to its low rate of accuracy.' OpenAI plans to bring back a better version.

https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

Watermarking LLM-generated text

Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark 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	56	.31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

- While generating, replace some words by "watermarked" words
- Count "watermarked words" to identify LLM generation
- **Under the hood**: add bias to 50% of the logits (watermarked tokens) during sampling

Kirchenbauer et al., ICML 2023, A Watermark for Large Language Models

What makes a good LLM-generated text detector?

1. High scores for LLM-written text (high true positive rate)
2. Low scores for human-written text (low false positive rate)
3. Minimal changes to the quality of LLM-generated text (indistinguishable to human reader)
4. Robustness to perturbation attacks (paraphrasing)

Generation: They have never been known to mingle with humans. Today, it is believed these unicorns live in an unspoilt environment which is surrounded by mountains. Its edge is protected by a thick wattle of wattle trees, giving it a majestic appearance. Along with their so-called miracle of multicolored coat, their golden coloured feather makes them ....

LLM-generated? score = 85.6

score > T? (T = 75.0) Yes, AI-generated!

What makes a good LLM-generated text detector?

1. High scores for LLM-written text (high true positive rate)
2. Low scores for human-written text (low false positive rate)
3. Minimal changes to the quality of LLM-generated text (indistinguishable to human reader)
4. Robustness to perturbation attacks (paraphrasing)

Generation: They have never been known to mingle with humans. Today, it is believed these unicorns live in an unspoilt environment which is surrounded by mountains. Its edge is protected by a thick wattle of wattle trees, giving it a majestic appearance. Along with their so-called miracle of multicolored coat, their golden coloured feather makes them ....

LLM-generated? score = 85.6

score > T? (T = 75.0) Yes, AI-generated!

Paraphrasing easily evades detection of AI-generated text, but retrieval is an effective defense

Kalpesh Krishna

Yixiao Song

Marzena Karpinska

John Wieting

Mohit Iyyer

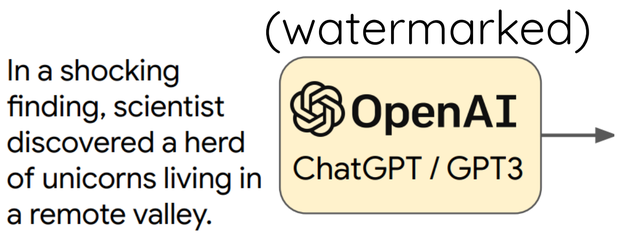
UMassAmherst

Manning College of Information & Computer Sciences

Google AI

NeurIPS 2023

How do paraphrases affect LLM-generated text detectors?



Detectors are not effective on paraphrases

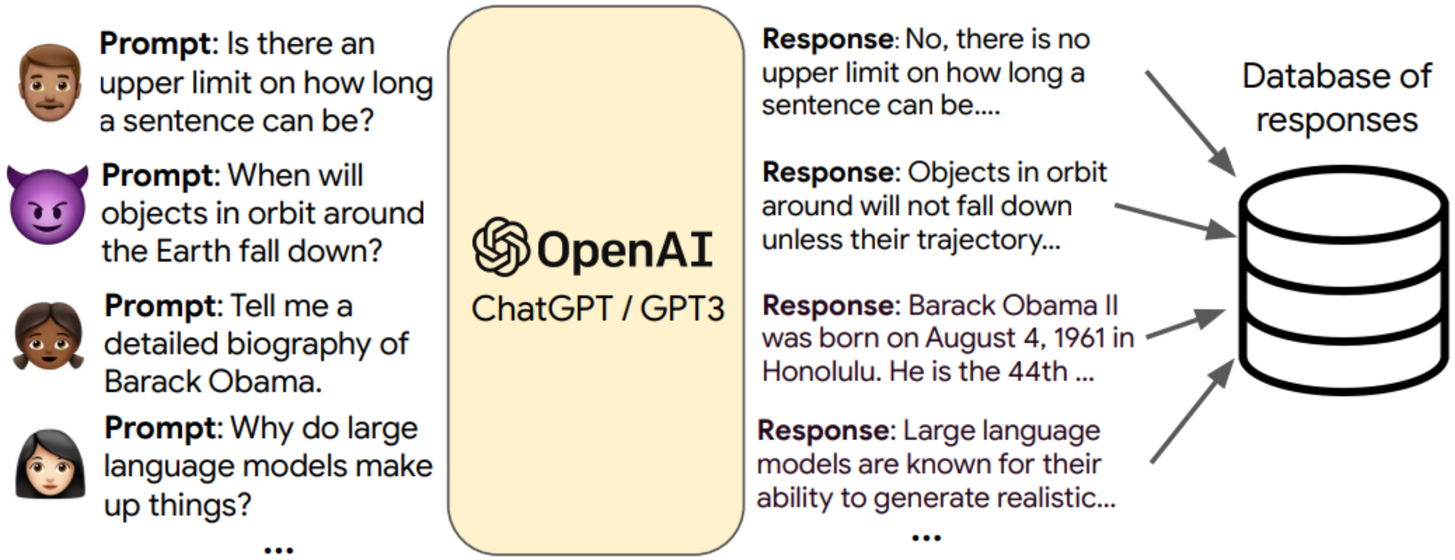
Language model	Similarity	Watermarks	DetectGPT	GPTZero	OpenAI Classify
GPT2-XL					
GPT2-XL + DIPPER					
OPT-13B					
OPT-13B + DIPPER					
GPT3.5					
GPT3.5 + DIPPER					

Task: Wikipedia article completion

Detection rates are computed at a 1% false positive rate

Retrieval offers an alternate (and more robust) detection method!

Step 1: Maintain a database of LLM-generated text



Step 2: Provide a search engine over LLM responses

Candidate (LLM-generated paraphrase): There were never any reports of them mixing with people. It is believed they live in an unspoiled environment surrounded by mountains and protected by a thick clump of wattle. The herd has a regal look to it, with the magic, rainbow-colored coat and golden feathers…

Retriever over DB
(search engine like Google)

Best match among previous generations: They have never been known to mingle with humans. Today, it is believed these unicorns live in an unspoilt environment which is surrounded by mountains. Its edge is protected by a thick wattle of wattle trees, giving it a majestic…

Candidate: There were never any reports of them mixing with people. It is believed…

Best Match: They have never been known to mingle with humans. Today it is…

Similarity score (SIM, 1-gram) score = 95.0 score > T? (T = 75.0) Yes, LLM-generated!

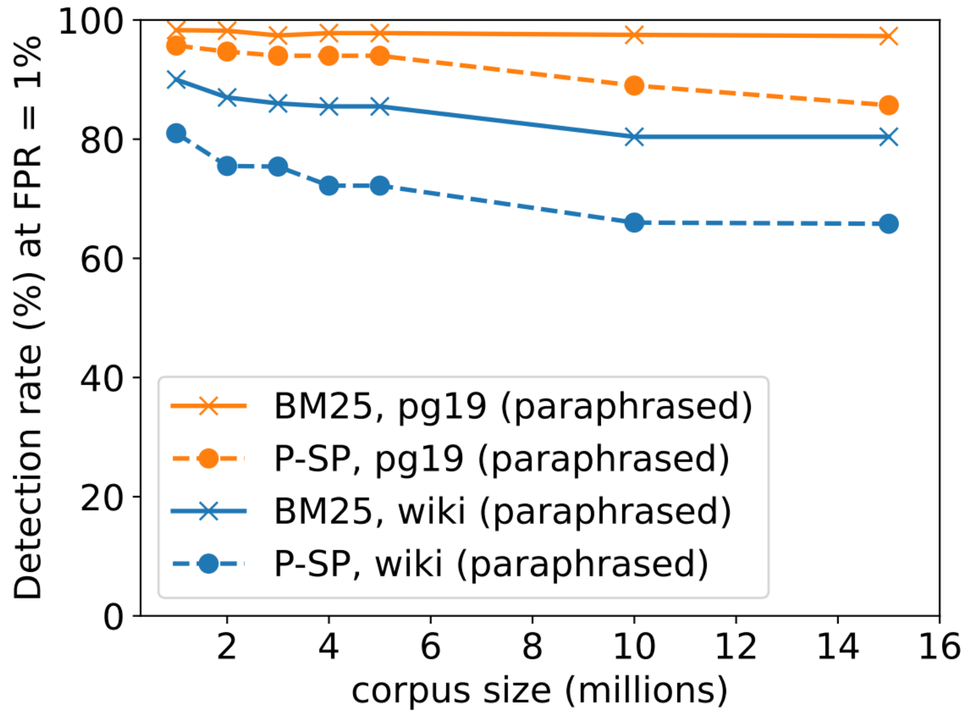
Paraphrases will also have high similarity scores!

Retrieval is effective against paraphrases!

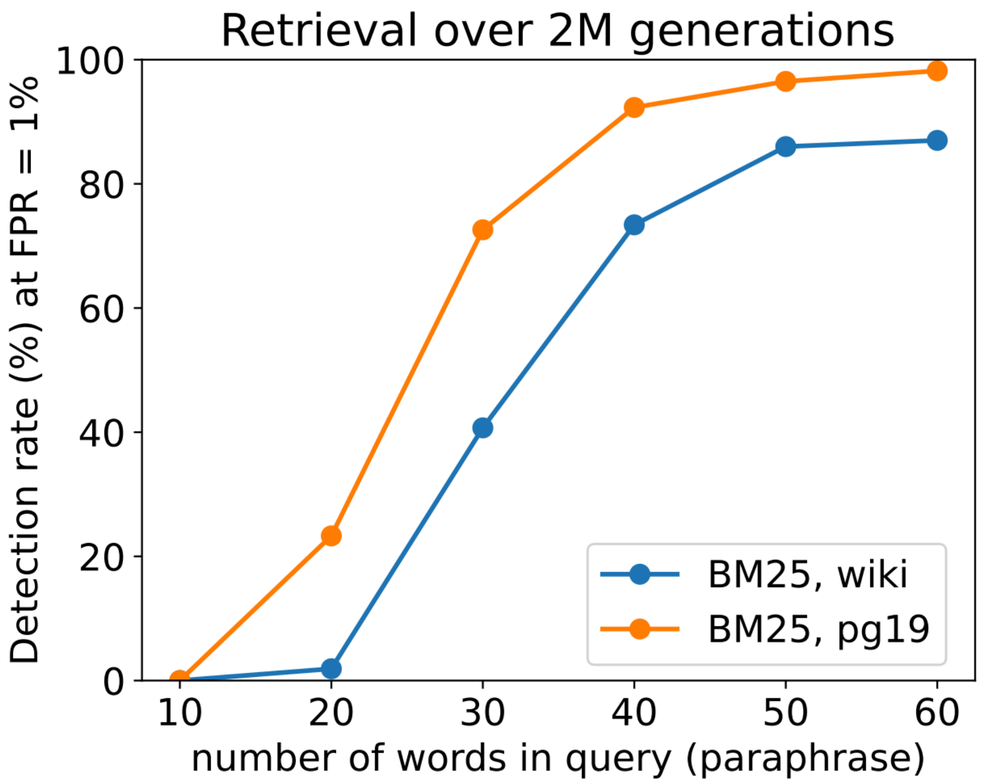
Language model	Watermarks	OpenAI Classifier	Retrieval
GPT2-XL	100.0	59.2	
GPT2-XL + DIPPER	55.8	32.7	
OPT-13B	100.0	33.5	
OPT-13B + DIPPER	65.5	21.6	
GPT3.5	-	40.5	
GPT3.5 + DIPPER	-	38.1	

Task: Long-form question answering

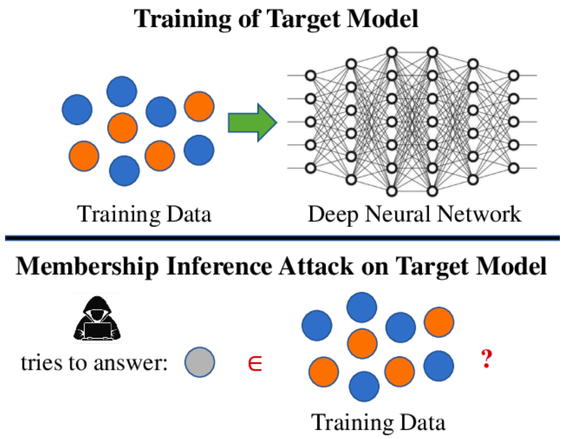
Retrieval has high detection rates on paraphrases even with a corpus of size 15M!



Retrieval works best with generations that are >50 tokens

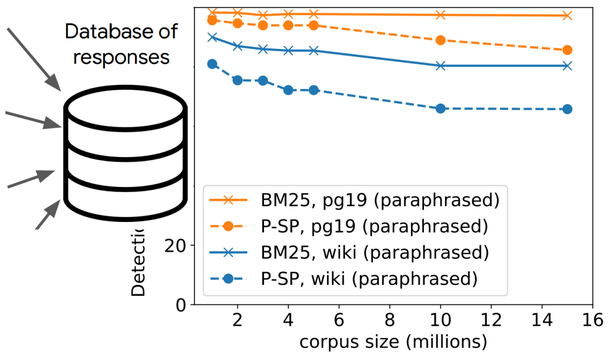


Limitations of retrieval as a detector



Privacy risk — membership inference attacks

Search engine needs to be implemented by provider



Accuracy reduction with large DB

LLM-generated text detection is both enormously impactful and challenging.

All existing methods have critical flaws.

One interesting future direction is **semantic** watermarks that cannot be removed via paraphrasing.

New attacks will always be invented, so this will likely never become a solved problem.