The Science & Implications of Detecting LLM-generated Text

Kayla Schilz & Dillon Blair

Table of contents

01

Large Language Models

02

Societal Impact

03

AI Detection Software

04

Future Research

05

Conclusion

Objectives

- Understand what an LLM is as well as its challenges
- Note examples of past and current LLMs
- Understand the Mechanics of AI detection software
- Learn examples of detection software
- Explore future questions within AI/Detection

What is a Large Language Model?

- Large Language Models (a.k.a LLM) are trained on huge sets of data to be able to recognize human text
  - Deep Learning
    - Subset of machine learning
    - Uses deep neural network
    - Aims to replicate the complex decision-making power of humans
  - Tokenization
    - Breaks human words into small pieces in order for ai to process
  - Fine tuning
    - Training a pre-trained model on a smaller data set
- “Large Language Models (LLMs) have emerged as cutting edge artificial intelligence systems that can process and generate text with coherent communication, and generalize to multiple tasks.”

Tokenization

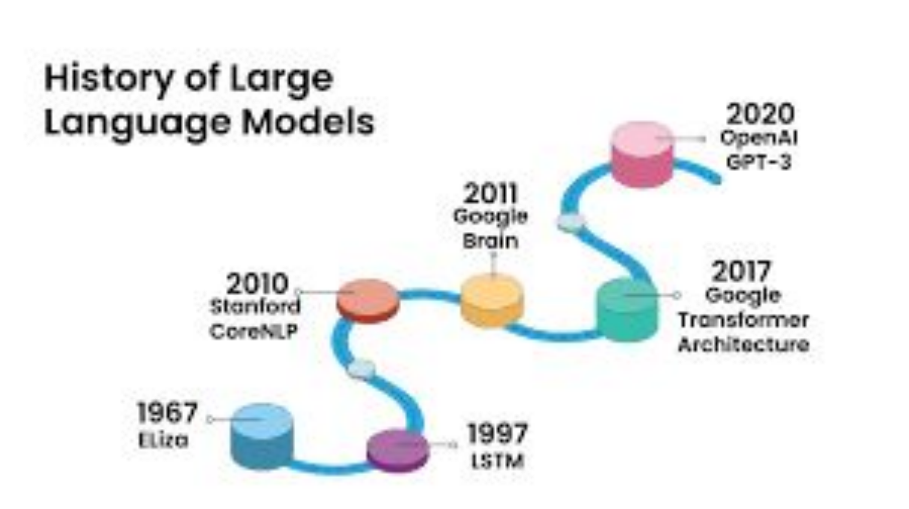
- Converting words into smaller units, called tokens
- This allows for AI to understand human text
- Benefits:
  - Data Security - sensitive data is now replaced thus harder to steal
  - Scalable Data processing - accommodating large amounts of data allowing for a seamless process
- Disadvantages:
  - Privacy Concerns
  - Over-reliance

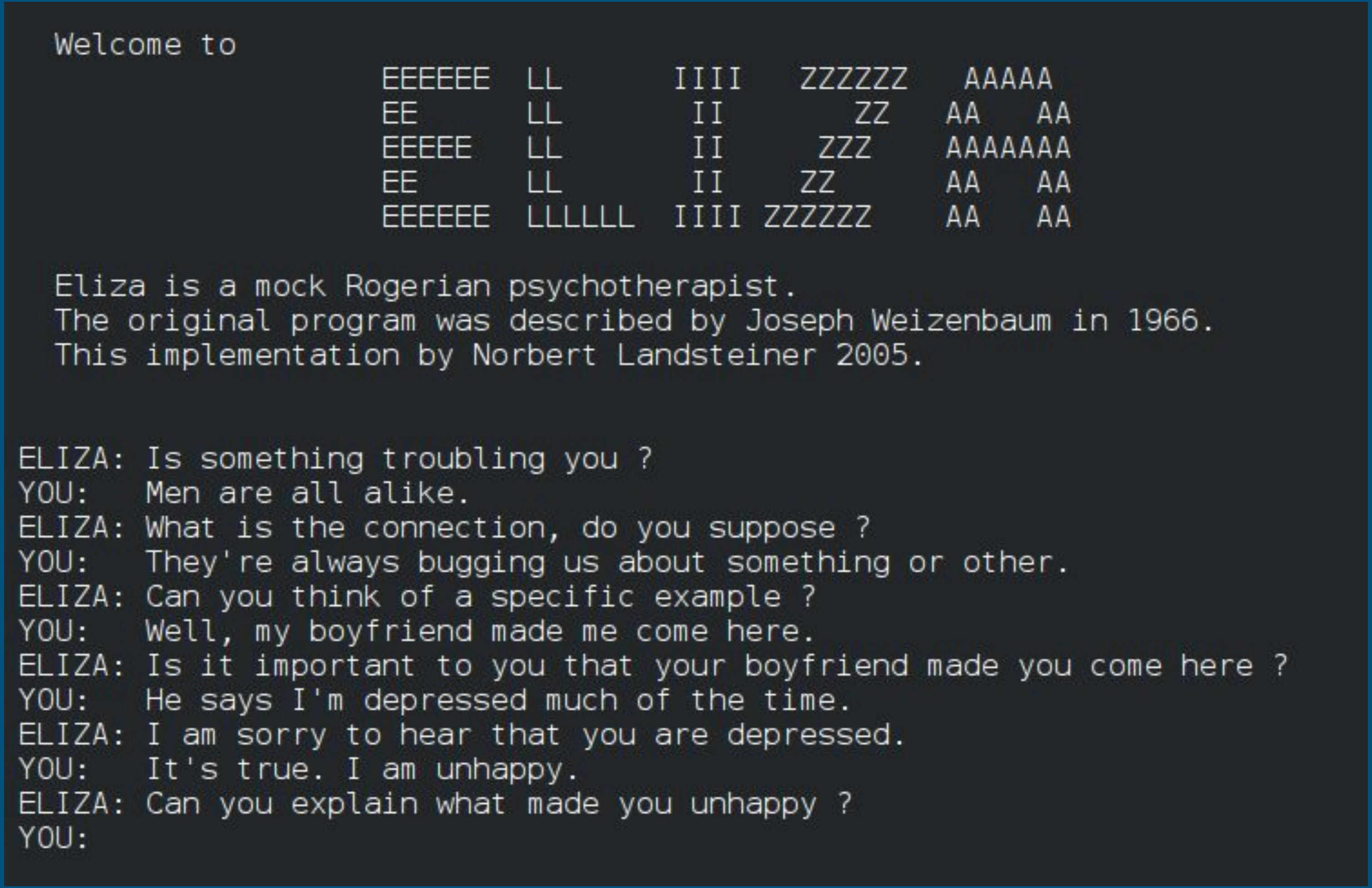
Fine Tuning

- Transfer Learning: The pre-trained LLMs perform well for various tasks
- Instruction-tuning: To enable a model to respond to user queries effectively, the pre-trained model is fine-tuned on instruction formatted data i.e., instruction and an input-output pair.
- Alignment-tuning: LLMs are prone to generate false, biased, and harmful text. To make them helpful, honest, and harmless models are aligned using human feedback. Alignment involves asking LLMs to generate unexpected responses and then updating their parameters to avoid such responses

Brief History of LLMs

- Eliza
- LSTM
- Stanford CoreNLP
- Google Brain
- Google Transformer Architecture
- OpenAI GPT-3





```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Examples of LLMs

- PaLM 2
    - Developed by Google
- Gemini (Bard)
    - Developed by Google
- LLaMA
    - Created by Meta and Microsoft
- LaMDA
    - Developed by DeepMind (Google)

What challenges do LLMs present?

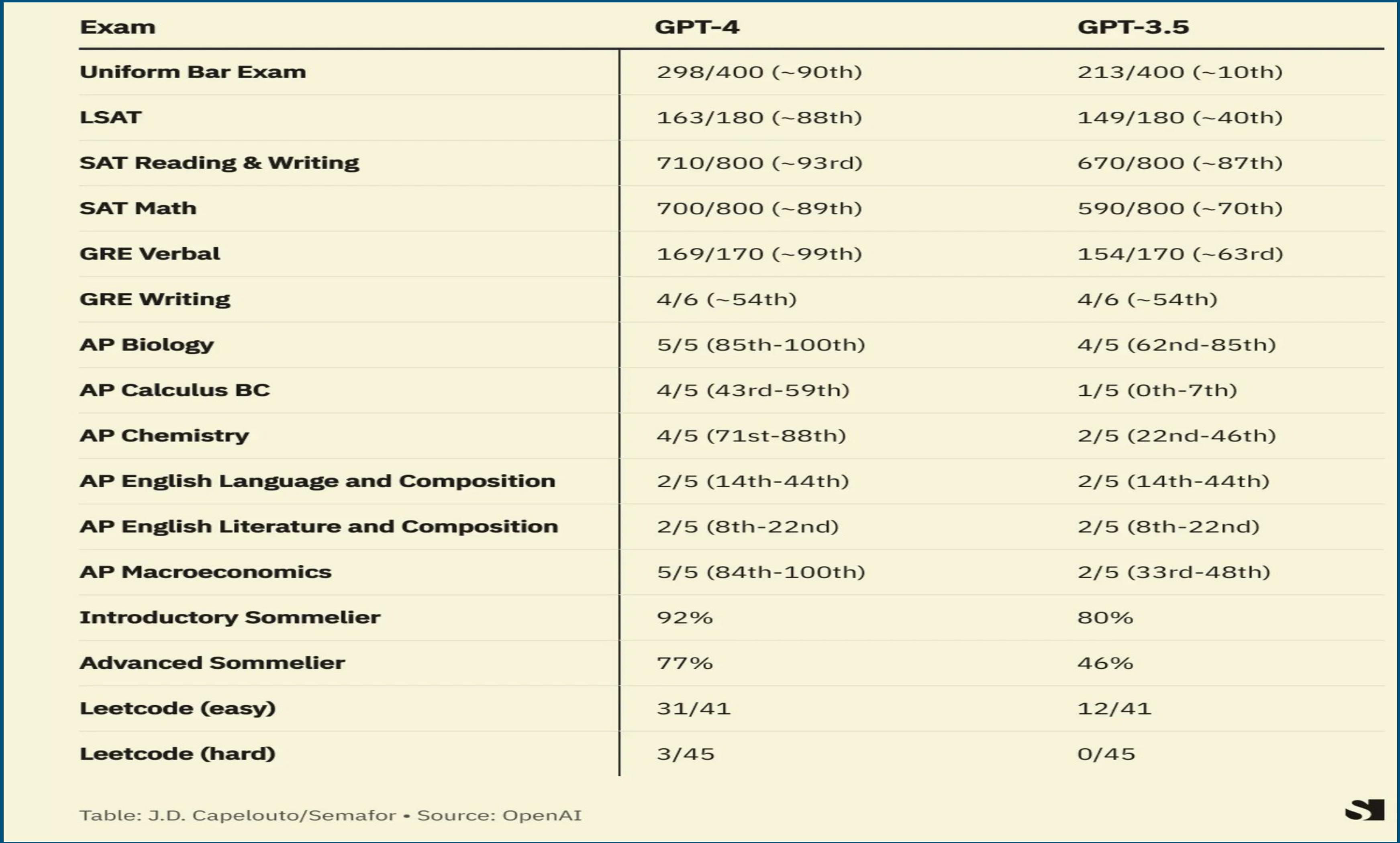
- “The improved efficiency and effectiveness of these models are changing how we do things across a multitude of domains. As a result, when we talk about LLMs, we’re not just measuring their technical competency, but also looking at their broader societal and professional implications.
- The advent of powerful natural language processing technologies presents many issues going forward. It has sparked concerns over:
    - Spreading misinformation online
    - Academic misconduct & Regression
    - Cybersecurity
    - Job loss
    - Creating trust

Misinformation

- “Historically, propaganda operations have relied on armies of low-paid workers or highly coordinated intelligence organizations to build sites that appear to be legitimate.” Not anymore with LLMs
- For example, the newswork Global Village Space, https://www.globalvillagespace.com/
- In journalism, the emergence of AI-generated "deepfake" news articles can threaten the credibility of news outlets and misinform the public on a variety of topics, such as elections, foreign affairs, etc.

Disruption in Education

- LLMs have made torrential waves across all levels of education
- Their ability to provide quick answers to complex problems threatens to undermine students critical thinking and problem-solving skills and they offer students an easy path towards engaging in academic dishonesty in the form of cheating.
- An NYC school district banned chatGPT on all of its proprietary devices.
  - https://www.chalkbeat.org/newyork/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence/
- ChatGPT3 passes MBA warden exam
  - https://www.nbcnews.com/tech/tech-news/chatgpt-passes-mba-exam-wharton-professor-rcna67036
- Even at UNCO, we added an AI policy that can be found in the syllabi



Exam	GPT-4	GPT-3.5
Uniform Bar Exam	298/400 (~90th)	213/400 (~10th)
LSAT	163/180 (~88th)	149/180 (~40th)
SAT Reading & Writing	710/800 (~93rd)	670/800 (~87th)
SAT Math	700/800 (~89th)	590/800 (~70th)
GRE Verbal	169/170 (~99th)	154/170 (~63rd)
GRE Writing	4/6 (~54th)	4/6 (~54th)
AP Biology	5/5 (85th-100th)	4/5 (62nd-85th)
AP Calculus BC	4/5 (43rd-59th)	1/5 (0th-7th)
AP Chemistry	4/5 (71st-88th)	2/5 (22nd-46th)
AP English Language and Composition	2/5 (14th-44th)	2/5 (14th-44th)
AP English Literature and Composition	2/5 (8th-22nd)	2/5 (8th-22nd)
AP Macroeconomics	5/5 (84th-100th)	2/5 (33rd-48th)
Introductory Sommelier	92%	80%
Advanced Sommelier	77%	46%
Leetcode (easy)	31/41	12/41
Leetcode (hard)	3/45	0/45

Table: J.D. Capelouto/Semafor • Source: OpenAI

Cybersecurity, Job Loss, & Building Trust

- Dangers of enhanced phishing and social engineering attacks
- With the rise of LLMs, there is a concern over job loss due to automation
- If there is an infallible way to detect AI generated content, people will be more willing to trust these tools.

How do you detect LLM-generated text?

- People are developing software that attempts to identify AI-generated content in a way that distinguishes it from human-generated content.
- There are two main types:
  - Black-box detection
  - White-box detection

Black-box Detection

External entities (detector software) restricted to API-level access to LLM

1) Gather LLM generated data and human authored data
2) Human evaluation (exclamation marks)
3) Then Feature detection:
    a) Statistical disparities
    b) Linguistic patterns
    c) Fact verification
4) Classification Models use the previous features to binarily distinguish between human and ai generated text:
    a) Support Vector Machines, Naive Bayes, and Decision Trees.

- Deep learning approaches involve using AI models to detect vs. using explicitly extracted features

White-box Detection

“In white-box detection, the detector processes complete access to the target language model, facilitating the integration of concealed watermarks into its outputs to monitor suspicious or unauthorized activities.”

- Post-hoc watermarking
- Inference-time watermarking

Post-Hoc Watermarking

“Given an LLM-generated text, post-hoc watermarks will embed a hidden message or identifier into the text. There are two main categories of post-hoc watermarking methods:

- Rule-based approaches
  - Exploits syntactic and semantic structures to embed watermarks
- Neural-based approaches
  - Involves three components, a watermark encoder network, a watermark decoder network, and a discriminator network

Inference-time watermark

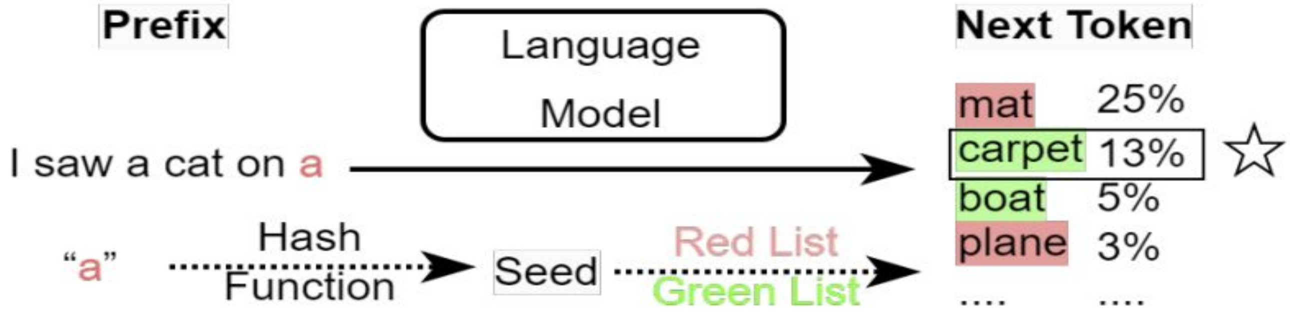


Figure 4. Illustration of inference time watermark. A random seed is generated by hashing the previously predicted token "a", splitting the whole vocabulary into "green list" and "red list". The next token "carpet" is chosen from the green list.

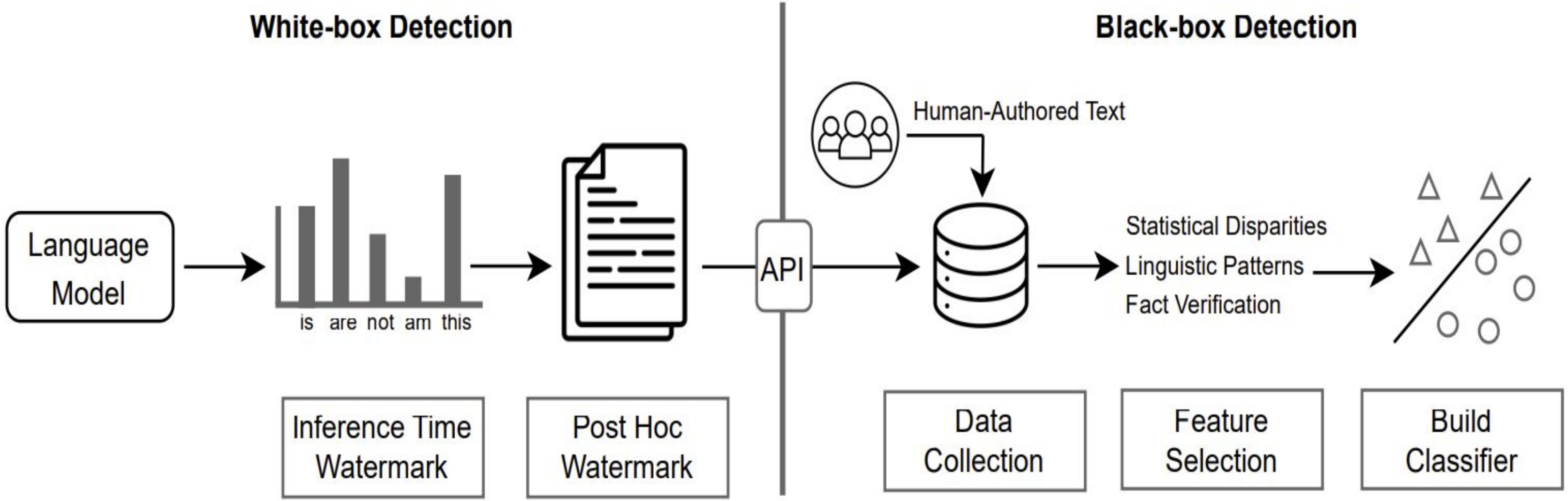
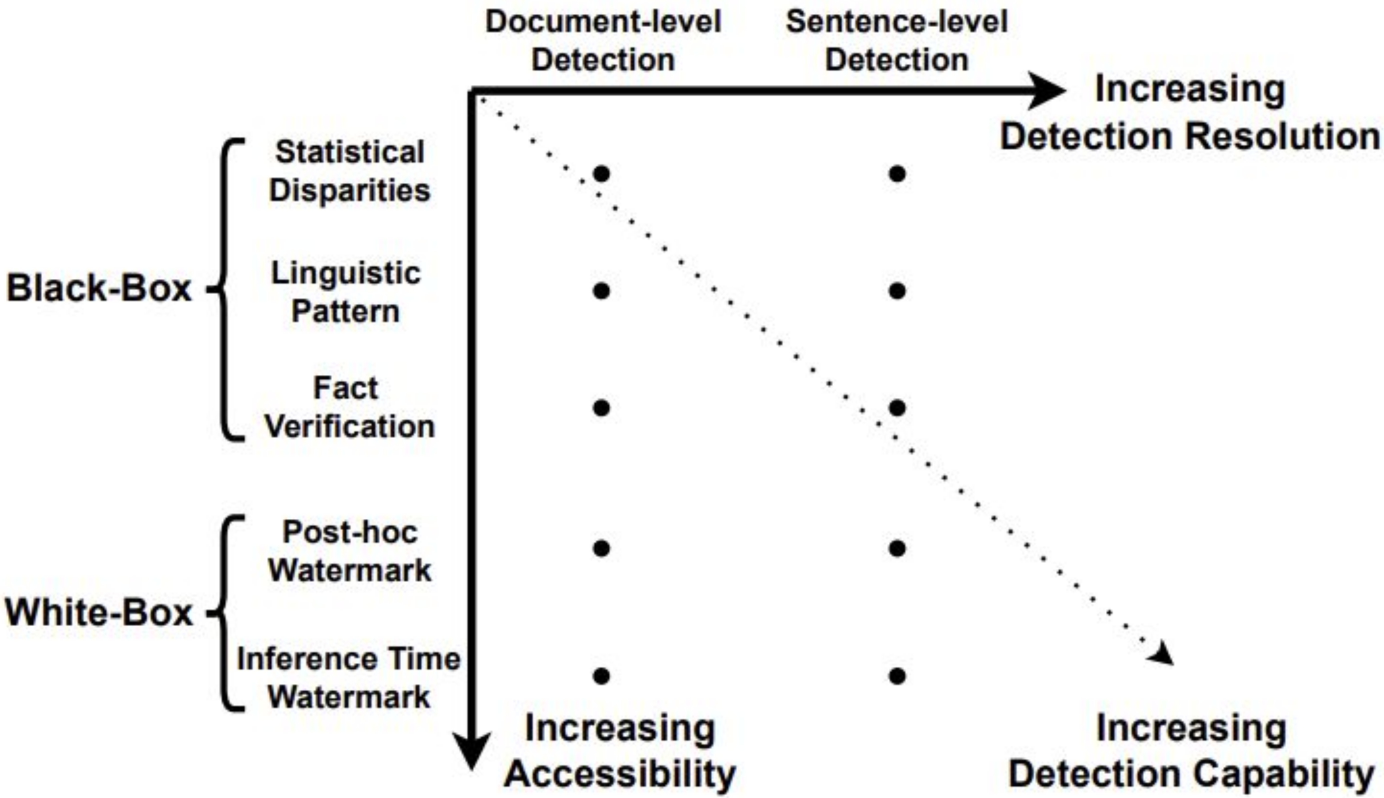


Figure 1. An overview of the LLM-generated text detection.



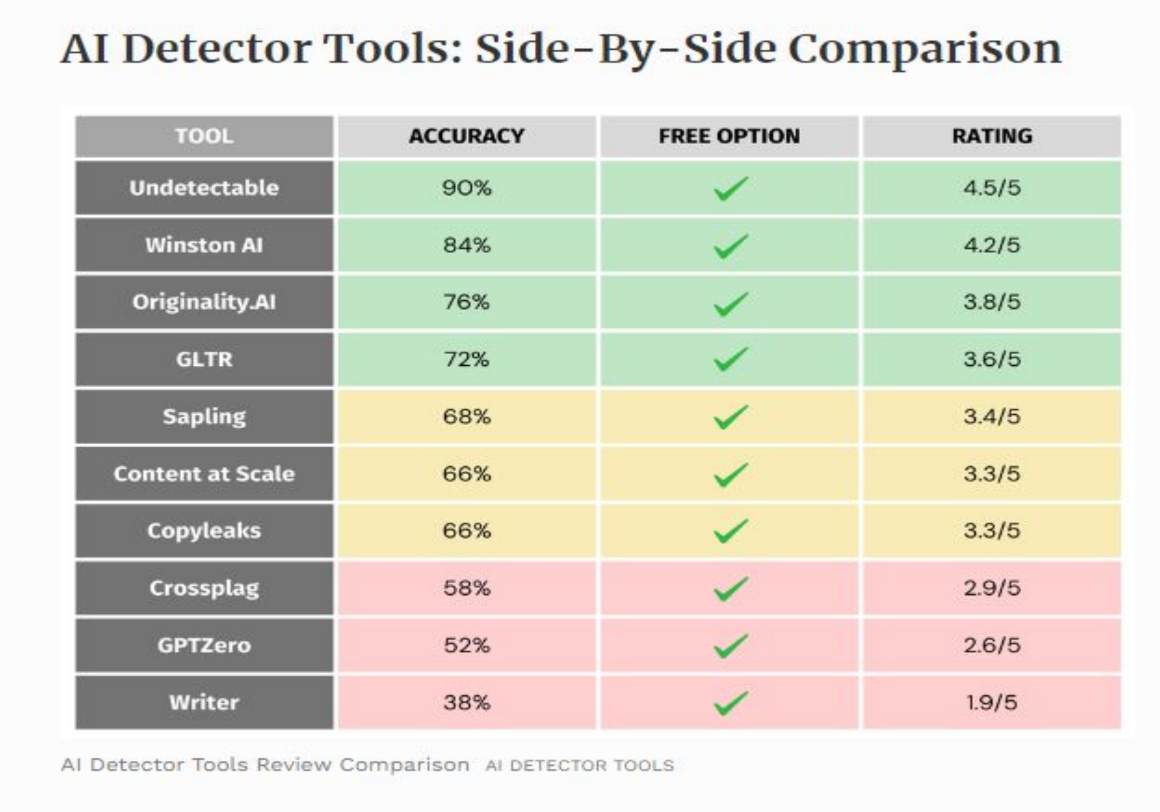
		Document-level Detection	Sentence-level Detection
Black-Box	Statistical Disparities	●	●
	Linguistic Pattern	●	●
	Fact Verification	●	●
White-Box	Post-hoc Watermark	●	●
	Inference Time Watermark	●	●

Increasing Detection Resolution

Increasing Accessibility

Increasing Detection Capability

Examples of Software



AI Detector Tools: Side-By-Side Comparison

TOOL	ACCURACY	FREE OPTION	RATING
Undetectable	90%	✓	4.5/5
Winston AI	84%	✓	4.2/5
Originality.AI	76%	✓	3.8/5
GLTR	72%	✓	3.6/5
Sapling	68%	✓	3.4/5
Content at Scale	66%	✓	3.3/5
Copyleaks	66%	✓	3.3/5
Crossplag	58%	✓	2.9/5
GPTZero	52%	✓	2.6/5
Writer	38%	✓	1.9/5

AI Detector Tools Review Comparison AI DETECTOR TOOLS

Additional Research Questions/Gaps in Research

- How will the writing styles of children growing up with these tools be impacted by LLM?
- How many fundamental skills will people need to develop when there are powerful LLMs?
- Will NPCs in video games feel much more real due to LLMs?

Conclusions

LLMs

A type of AI model that uses tokenization and fine tuning

Social Impact

Fear of misinformation, privacy concerns, and disruption of education

Detection

By using black and white box, there is a way to detect AI but it isn’t perfect

Which is AI? Summary of “The Great Gatsby”

"The Great Gatsby" by F. Scott Fitzgerald is a quintessential American novel set in the prosperous yet morally ambiguous Jazz Age of the 1920s. The story is narrated by Nick Carraway, a young man from the Midwest who moves to New York and becomes embroiled in the lives of his wealthy and enigmatic neighbor, Jay Gatsby, and his cousin Daisy Buchanan. Gatsby is obsessed with reclaiming Daisy's love, despite her marriage to the wealthy but morally bankrupt Tom Buchanan. Through lavish parties and extravagant displays of wealth, Gatsby tries to win Daisy back, but his efforts ultimately end in tragedy. The novel explores themes of the American Dream, the corrupting influence of wealth, and the emptiness of materialism. Fitzgerald's lyrical prose and intricate characterizations paint a vivid portrait of the era's excesses and disillusionments, making "The Great Gatsby" a timeless classic of American literature.

In F. Scott Fitzgerald's "The Great Gatsby," readers are transported to the lavish and tumultuous world of 1920s America. Narrated by Nick Carraway, the novel centers around the mysterious millionaire Jay Gatsby and his infatuation with Daisy Buchanan, Nick's cousin. Set against the backdrop of extravagant parties and opulent lifestyles, the story delves into the complexities of love, wealth, and the pursuit of the American Dream. Gatsby's relentless pursuit of Daisy, despite her marriage to the brutish Tom Buchanan, leads to tragic consequences. Through Nick's eyes, readers witness the moral decay and superficiality of the Jazz Age, as well as the fragility of human desires. "The Great Gatsby" remains a poignant critique of the excesses of the Roaring Twenties and a timeless exploration of the elusive nature of happiness and fulfillment.

Sources

- https://arxiv.org/pdf/2307.06435.pdf
- https://steemit.com/science/@etherealcreation/eliza-beginning-of-era-of-artificial-intelligence
- https://arxiv.org/pdf/2309.16021.pdf
- https://arxiv.org/pdf/2309.15840.pdf
- https://contentatscale.ai/blog/how-do-ai-detectors-work/#:~:text=AI%20detection%20tools%20employ%20a,unique%20to%20AI%2Dproduced%20material
- https://arxiv.org/abs/2307.02599
- https://arxiv.org/abs/2303.07205
- https://www.chalkbeat.org/newyork/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence/
- https://www.nbcnews.com/tech/tech-news/chatgpt-passes-mba-exam-wharton-professor-rcna67036