

# **Evaluating AI Text Detectors on Student Essays**

by Neil Sorkin, Tom Goldstein

# Table of Contents

01

## **Problem**

How AI detectors are being used today, and consequences

02

## **IRB & Collection**

IRB process and gathering essays

03

## **Data Analysis**

Results of the data collected

04

## **Project Goals**

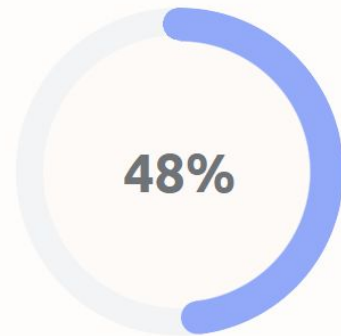
What I set out to do and what actually happened

“The University of Maryland’s Office of Student Conduct received **73 referrals for artificial intelligence-related academic integrity violations** during the 2022-23 academic year.” From these referrals, “**24 students ... received an “XF” mark on their transcript,** denoting failure with academic dishonesty.”

-“UMD records at least 50 AI-related academic integrity cases in 2022-23”, The Diamondback, November 27, 2023

# 01 Problem: Your career may depend on an AI text detector

- What is an AI text detector?
  - **Input:** Essay
  - **Output:** Probability the essay was AI written
- Why do professors use it?
  - To combat cheating
- Consequences
  - Expulsion, XF, points docked
  - Can result in denial from grad school/jobs
- **Detectors must be independently verified for accuracy, so we need essays**



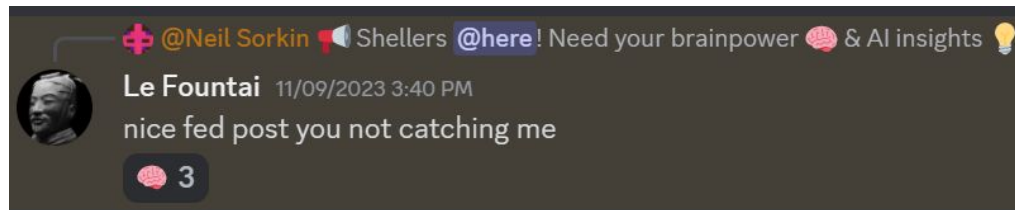
AI probability\*

GPTZero Example  
Output

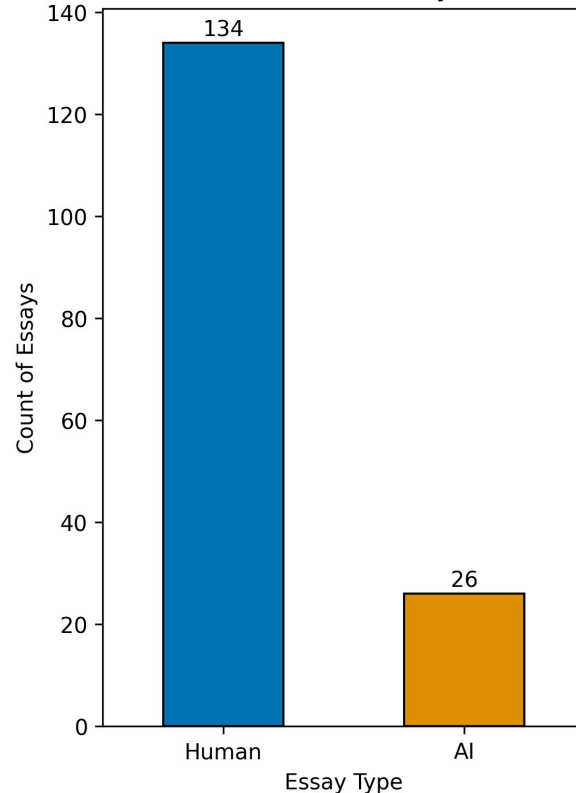


## 02 IRB and Essay Collection

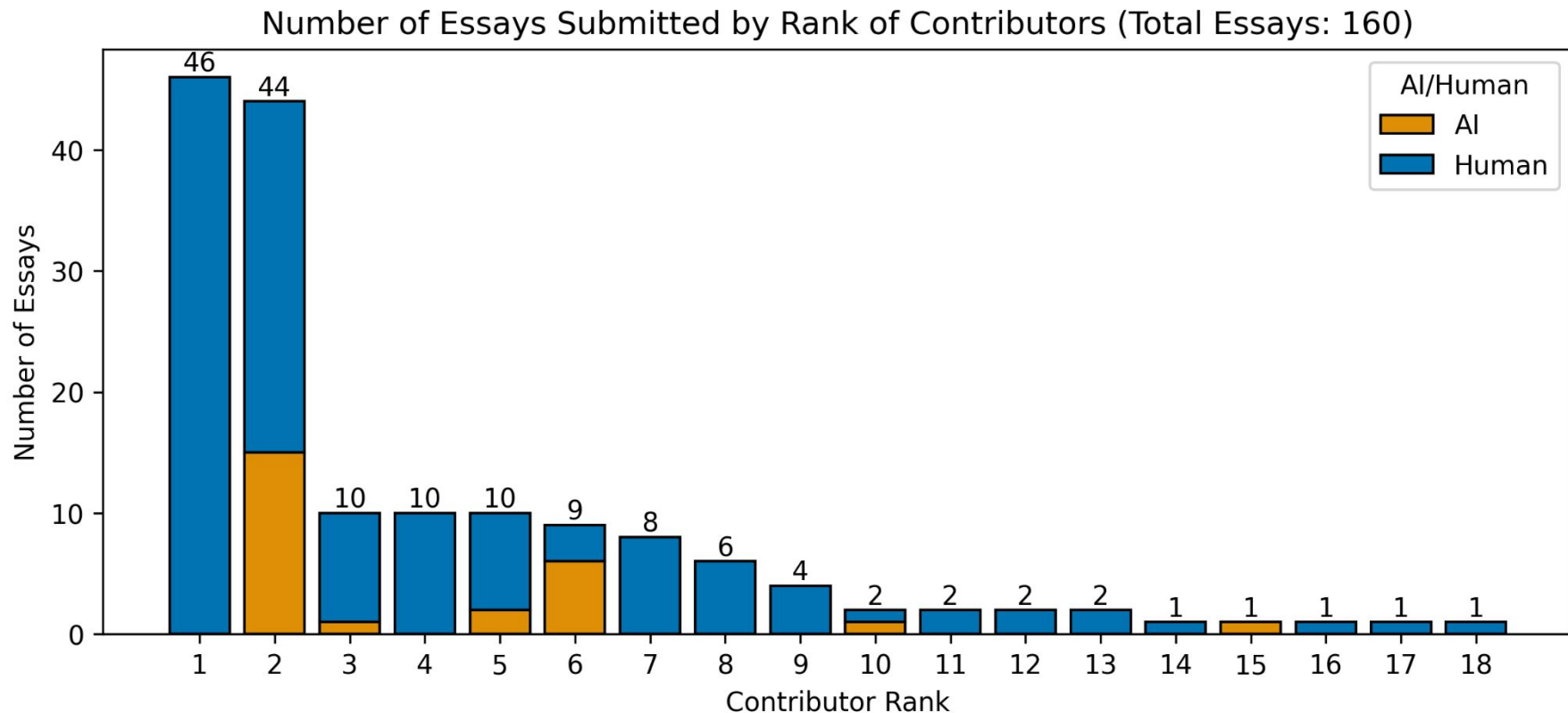
- IRB was not that bad [IRBNet](#)
- Collecting essays is difficult
  - ~10 minutes of effort for no reward
  - People are suspicious



Total Number of AI and Human Essays (Total Essays: 160)



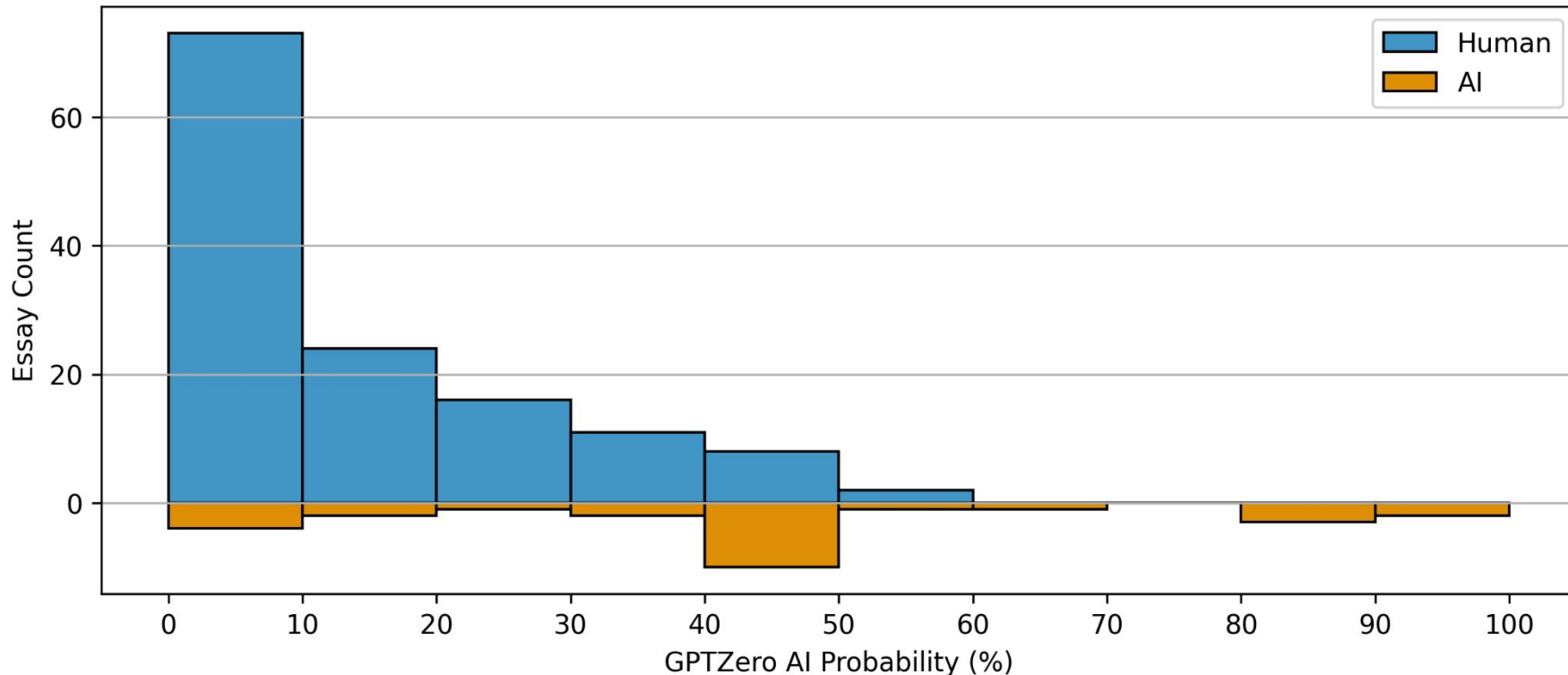
## 02 IRB and Essay Collection



# 03 Data Analysis (using GPTZero)

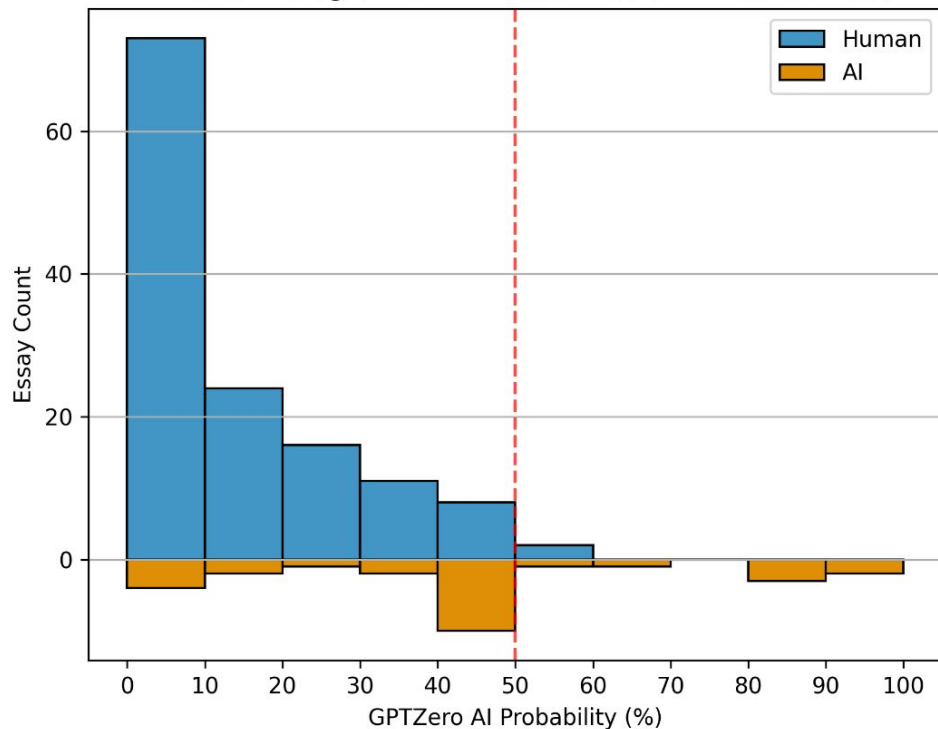


Mirrored Histogram of AI Probabilities

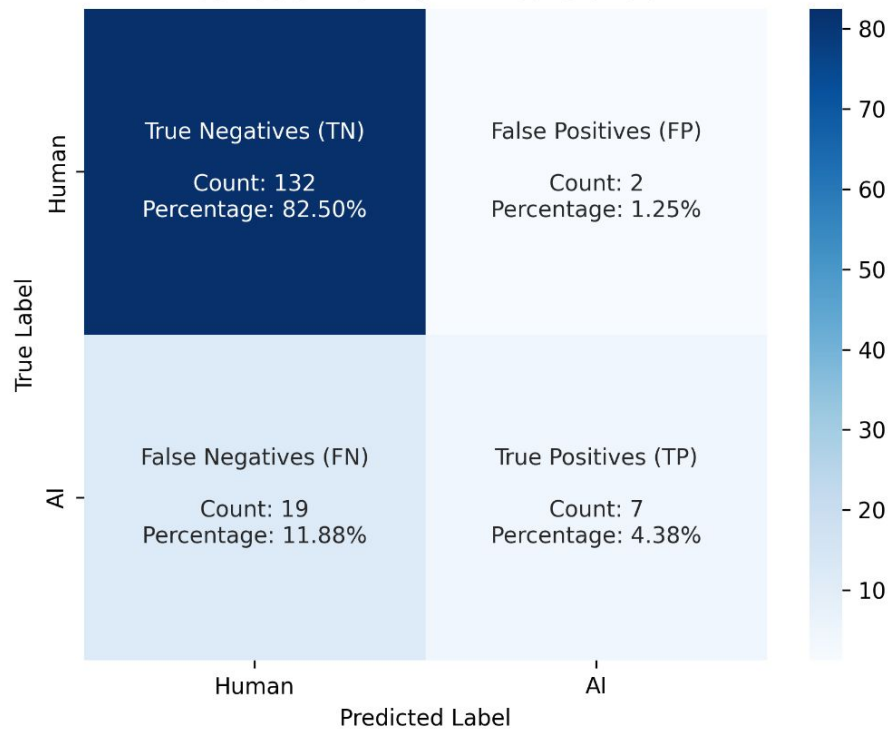


# 03 Data Analysis, Threshold at 50%

Mirrored Histogram of AI Probabilities (AI Threshold=50)



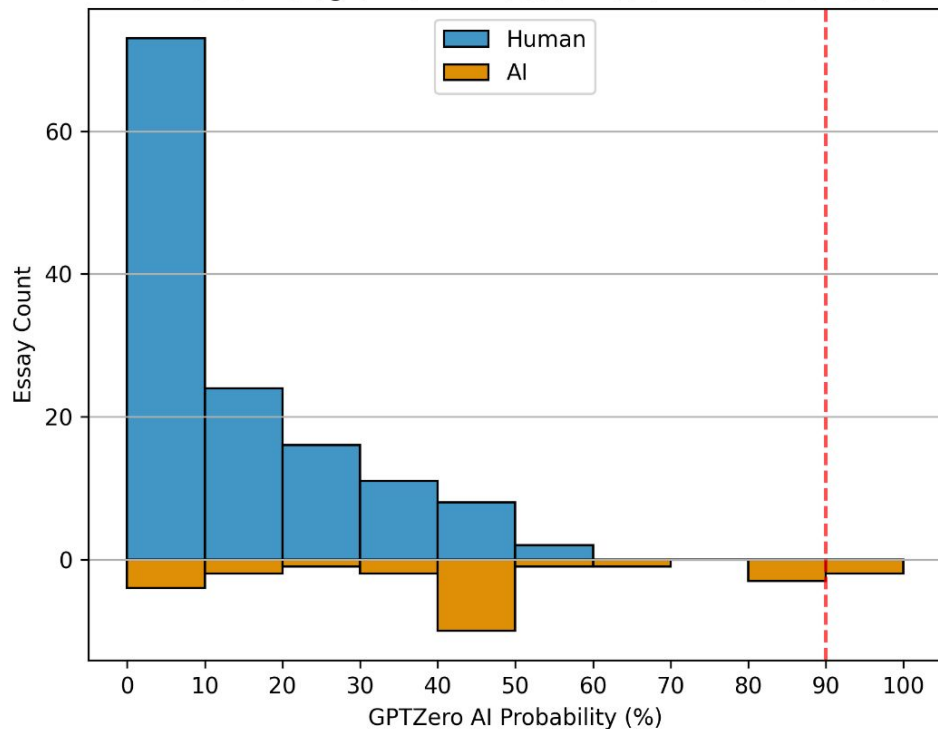
Confusion Matrix, AI Threshold=50



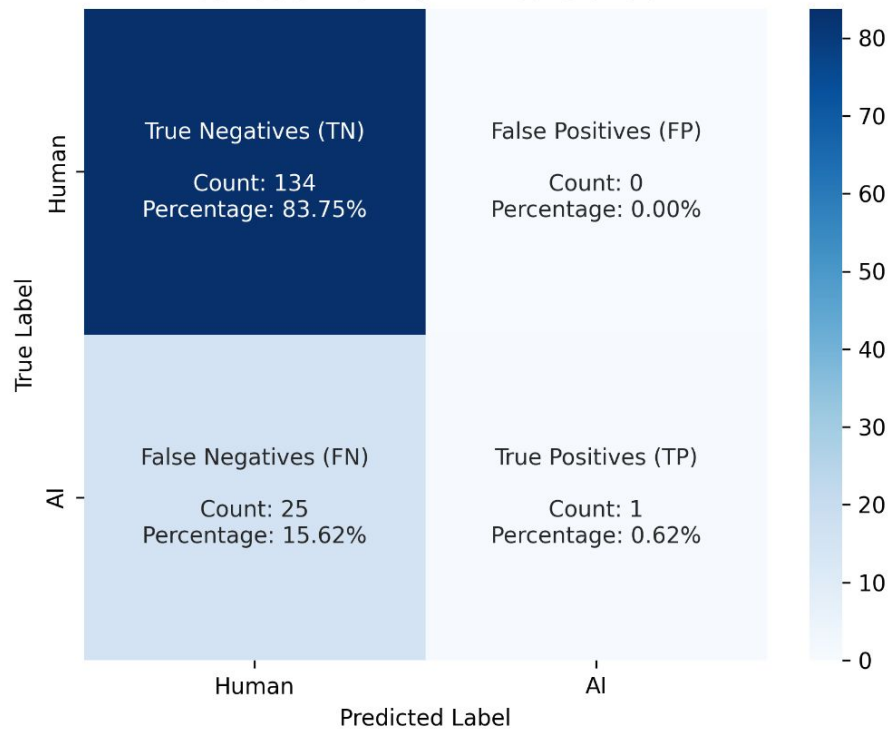


# 03 Data Analysis, Threshold at 90%

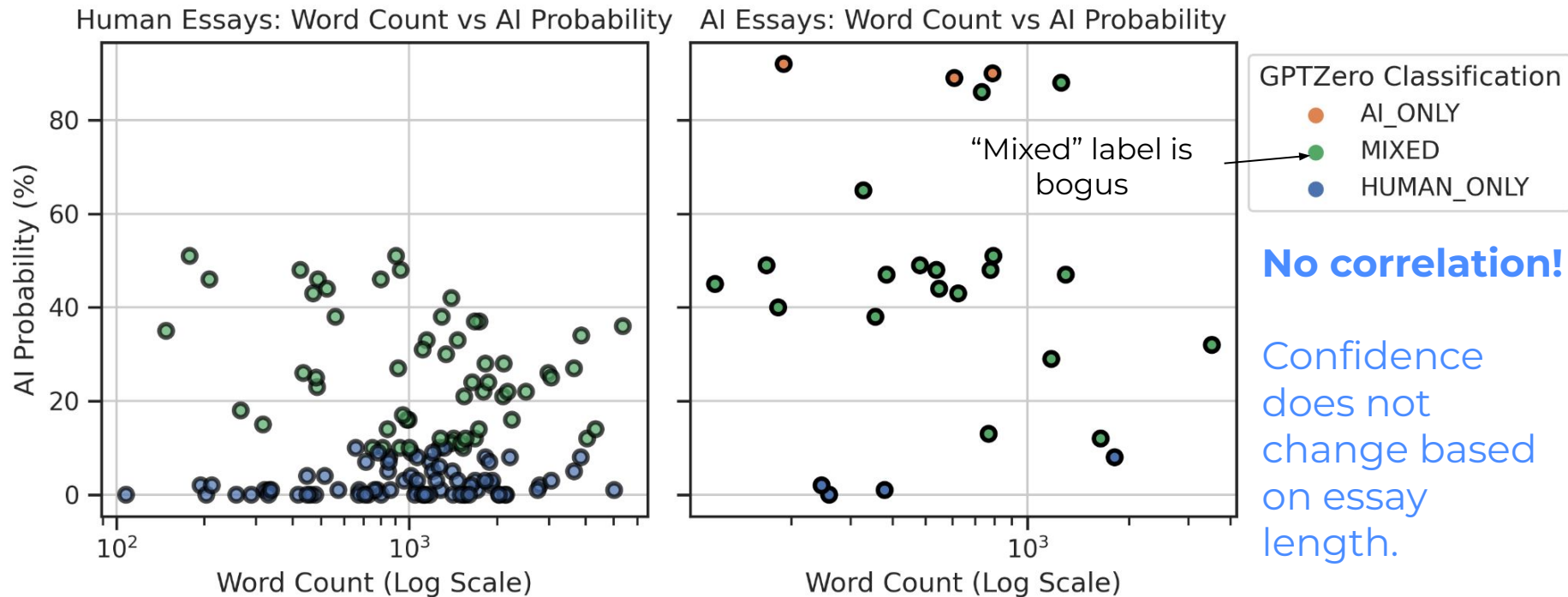
Mirrored Histogram of AI Probabilities (AI Threshold=90)



Confusion Matrix, AI Threshold=90

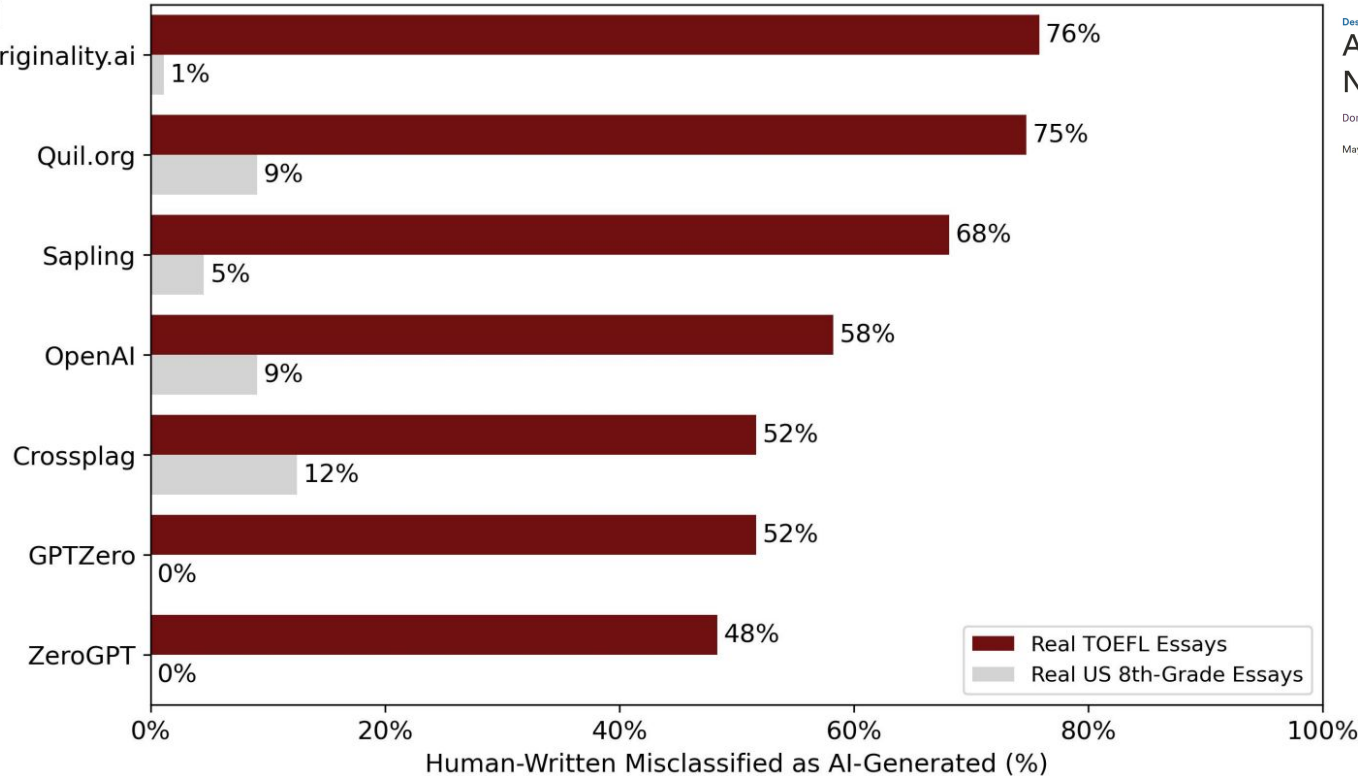


# 03 Data Analysis, Word Count vs AI%



# 03 Data Analysis, Non-Native English Writer Bias

a



Design and Human-Computer Interaction, Language Processing, Machine Learning

## AI-Detectors Biased Against Non-Native English Writers

Don't put faith in detectors that are "unreliable and easily gamed," says scholar.

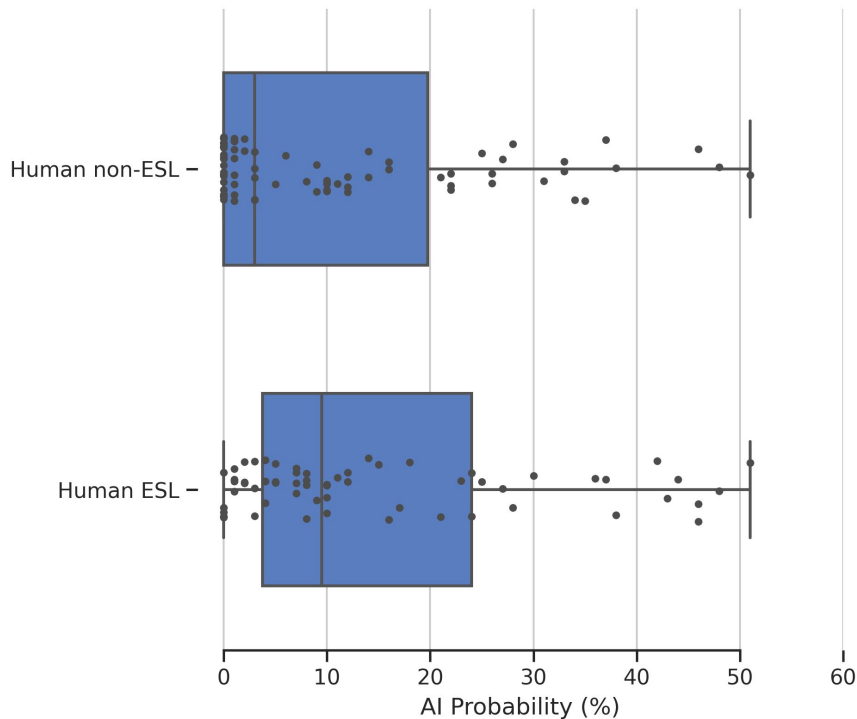
May 15, 2023 | Andrew Myers



**TOEFL** = Test of English as a Foreign Language

## 03

## Data Analysis, Non-Native English Writer Bias

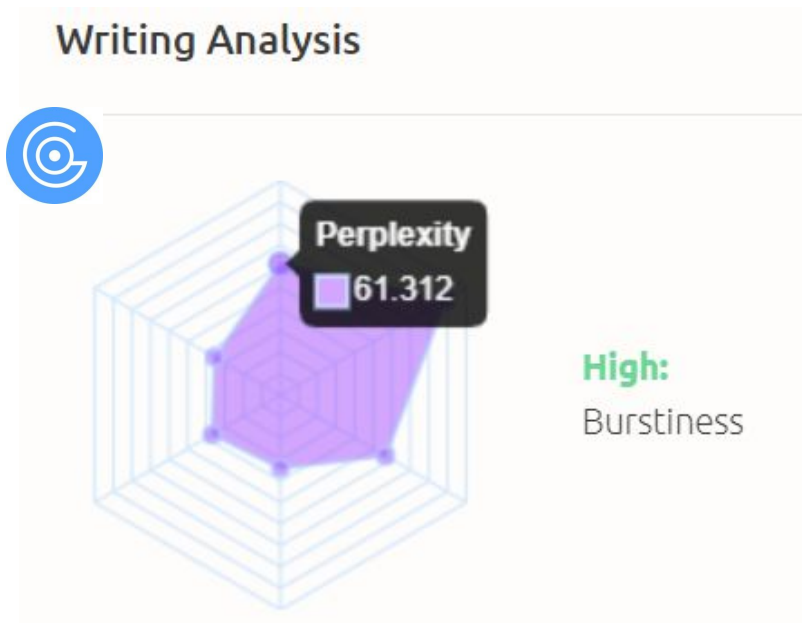


## Comparison of AI Probability Distributions

- **Categories:** Human non-ESL vs. Human ESL
- **Statistical Test:** Mann-Whitney U
- **P-value:** 0.0182
- **Interpretation:** Statistically significant difference

Implication: The AI Probability distributions for Human non-ESL and ESL groups are significantly different. This agrees with the Stanford study.

## 03 Data Analysis, Non-Native English Writer Bias



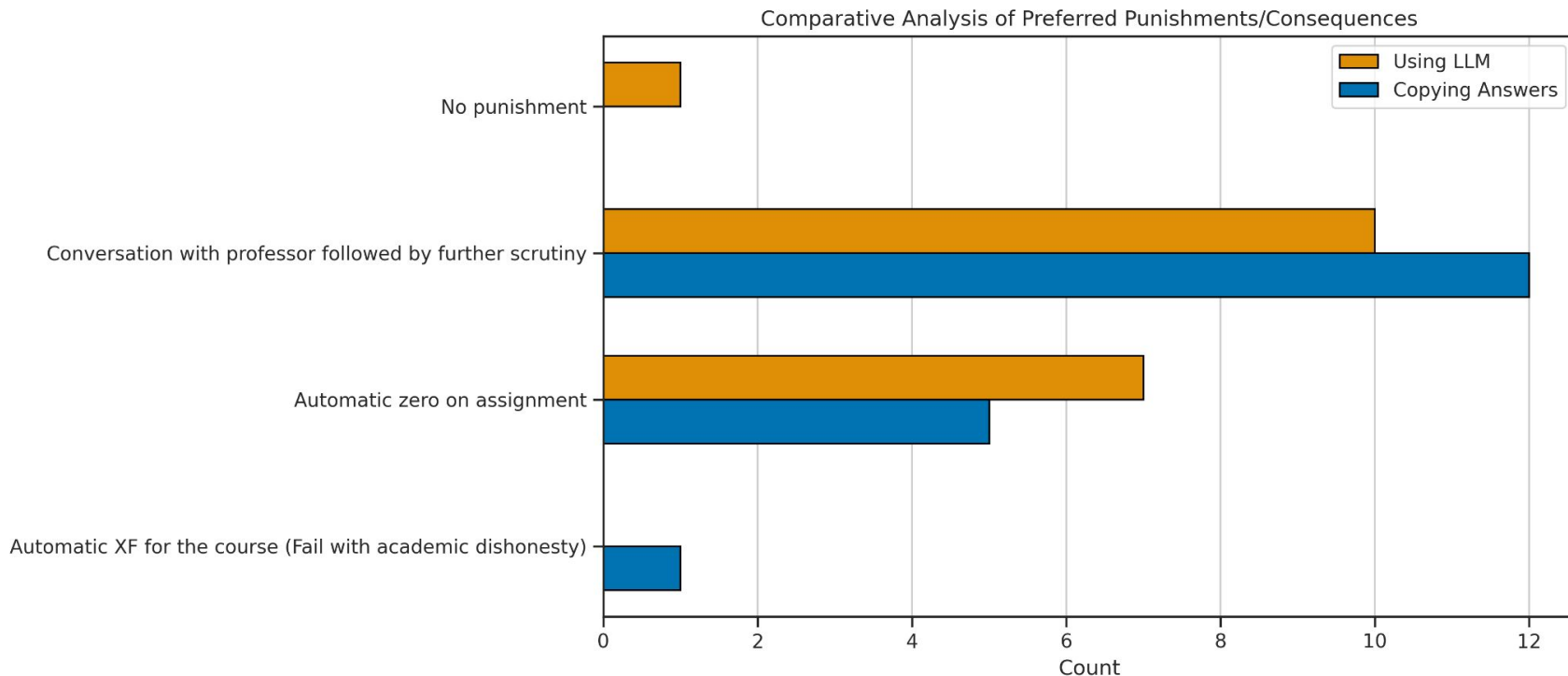
**GPTZero** Writing Analysis showing perplexity as a factor to determine “probability this text was entirely written by AI”

### Why is there a difference at all?





- “Non-native speakers typically score lower on common **perplexity** measures such as lexical richness, lexical diversity, syntactic complexity, and grammatical complexity”
- We should eat macaroni and \_\_\_\_\_
  - cheese = low perplexity / AI
  - CMSC473 = high perplexity / human

## 03 Data Analysis, Opinions

- What should the consequence be when a student gets caught **using an LLM**?
- A student is caught **copying their answers** on a writing assignment **from another student**, what should the punishment be?



## 04 Project Goals

-  Collect ~~>1000~~ **160** university essays
  - Very difficult
-  Evaluate ~~Turnitin~~ **GPTZero** on these essays
  - UMD shut down access to Turnitin, and no public access
-  Understand desired punishment policies for AI
-  ~~Explore strategies to bypass Turnitin detection~~

# Conclusion

- AI text detectors are **not perfect**
  - False negatives when cautious; false positives when aggressive
- One false positive is much worse than a false negative
- [GPTZero](#), and likely other detectors, are biased against English as a Second Language (ESL) writers
- We need **more data** to fill in the gaps, perhaps a paid study
- **Recommendation**
  - **Default:** Blacklist all AI text detectors.
  - **Criteria:** Accept only 0.01% false-positive rate.
  - **Ongoing:** Regular, independent re-evaluation required.



# Thanks

Do you have any questions?

Contact: [neilsorkin19@gmail.com](mailto:neilsorkin19@gmail.com)

Github:

[github.com/neilsorkin19/ai-essay-evaluation](https://github.com/neilsorkin19/ai-essay-evaluation)

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

# Sources

- <https://support.gptzero.me/hc/en-us/articles/15130070230551-How-do-I-interpret-burstiness-or-perplexity->
- <https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers>
- <https://arxiv.org/pdf/2304.02819.pdf>
- <https://dbknews.com/2023/11/27/umd-chatgpt-academic-integrity-cases/>

# Appendix



# UMD CS Grad School Application

Has disciplinary action been initiated  
against you at any of the institutions  
attended, including the University of  
Maryland?

Yes



Please explain

