Dataset Creation and Analysis for Evaluating AI Text Detection on University-Level Student Essays

Neil Sorkin, Department of Computer Science, University of Maryland, neilsorkin19@gmail.com

Tom Goldstein, Department of Computer Science, University of Maryland, tomg@umd.edu

Github repo: https://github.com/neilsorkin19/ai-essay-evaluation

"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." Source: "UMD records at least 50 AI-related academic integrity cases in 2022-23", The Diamondback, November 27, 2023

Introduction

Since the release of ChatGPT in November 2022, students around the world have embraced ChatGPT and other AI tools to complete course work faster than they could before. A 200 word reflection assignment that used to take 30 minutes can now be completed in 30 seconds with such tools. Some professors are not pleased with this, so they have found their own tools to detect AI writing in student assignments. GPTZero, Turnitin AI detector, Winston Al, Originality.Al, Undetectable, GLTR, Sapling, Content at Scale, Copyleaks, Crossplag, and Writer to name a few. All of them claim to be the best detectors, and they all would like professors to think that they have a false positive rate less than 1%. This study independently evaluates GPTZero.

Selection of GPTZero as the AI Text Detector

This study focuses on GPTZero, prompted by an incident encountered by Neil Sorkin during a business course at the University of Maryland. Midway through the semester, the course instructors, who initially did not restrict the use of AI for coursework in their syllabus, unexpectedly began enforcing a prohibition against it. At the end of a class, they made an announcement that they had used an AI detector, and ran class submissions through it. Two students were requested to stay after class for an alleged use of AI in their assignments.

One of these students, who Sorkin later spoke with after class, was graduating that semester and lived in San Diego. The student faced the possibility of extending their studies due to the accusation of AI use, despite having evidence from other AI detectors that suggested their work was not AI-written. This raised questions about the reliability and consistency of AI detection methods.

The tool employed by the course instructors was GPTZero. Upon registering for an account on GPTZero, a promotional offer of "300,000 words processed for free" was made if the role of the user was provided (Student, teacher, researcher, etc.), so it was used to process the collected essays. With more time, the same essays that were collected in this study would be put through other AI detectors.

There were plans to initially use Turnitin as it has an integration with Canvas Learning Management System (LMS), but the University of Maryland has disabled the AI writing detection feature according to correspondence with the IT department on September 29th 2023. They said "The administrators decided to NOT turn on the AI detection feature in UMD's Turnitin service. They made their decision based on concerns regarding reliability, bias, and overall accuracy of AI detection as evidenced by OpenAI shutting down their own AI detection software and universities like Vanderbilt suppressing the feature in Turnitin for similar reasons."

Dataset Creation

Institutional Review Board Process

The first step to evaluating AI text detectors on student essays is creating a dataset they have never seen. These essays are written by people, so the University of Maryland Institutional Review Board (IRB) had to be involved. The process to register was started on September 28th 2023, and finished on

November 1st 2023, so about 1 month. It took multiple rounds of back and forth, but eventually, the study was given "Exempt" status.

Survey Questions

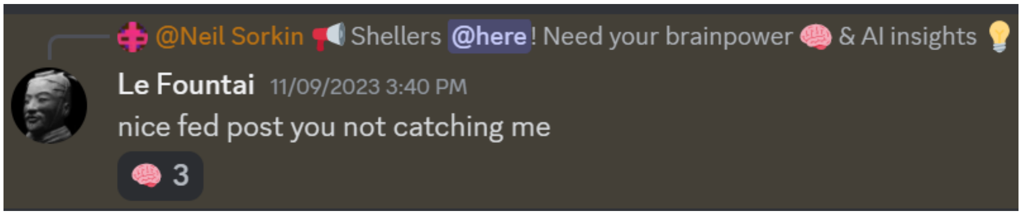
The data collected for each person was:

- Essay collection section
    - Is English your first language?
    - Essays written without the use of AI
    - Essays written with the use of AI
- Opinion questions section
    - If a student is caught copying their answers on a writing assignment from another student, what should the punishment be?
    - Suppose a student is using a Large Language Model (LLM), such as ChatGPT, when they are not allowed to use an LLM on an assignment. What should the consequence be when a student gets caught using an LLM?
- Demographic information
    - Gender identity
    - Race
    - Status as a student, professor, parent, alumni
    - If a student, a GPA range

The full survey Google Form with questions used for participants can be found on our Github: https://github.com/neilsorkin19/ai-essay-evaluation/blob/main/google_form_for_participants.pdf.

Recruitment

Subjects were recruited via personal messaging from Neil Sorkin. They were also recruited from Discord messaging groups known to contain university students. Sorkin used his personal connections with professors to try and get them to recruit students to fill out the survey. Because no incentive was given, it was difficult to get people to complete the survey. The survey required about 10 minutes of effort, fishing around for old essays, for no immediate gain. This likely led to sampling only from people who cared enough about the cause, or were good friends with Neil Sorkin. Additionally, students were suspicious of people who were asking for their AI written essays, as this could potentially get them caught for using AI when it was disallowed.



@Neil Sorkin 📢 Shellers @here! Need your brainpower 🧠 & AI insights 💡

Le Fountai 11/09/2023 3:40 PM

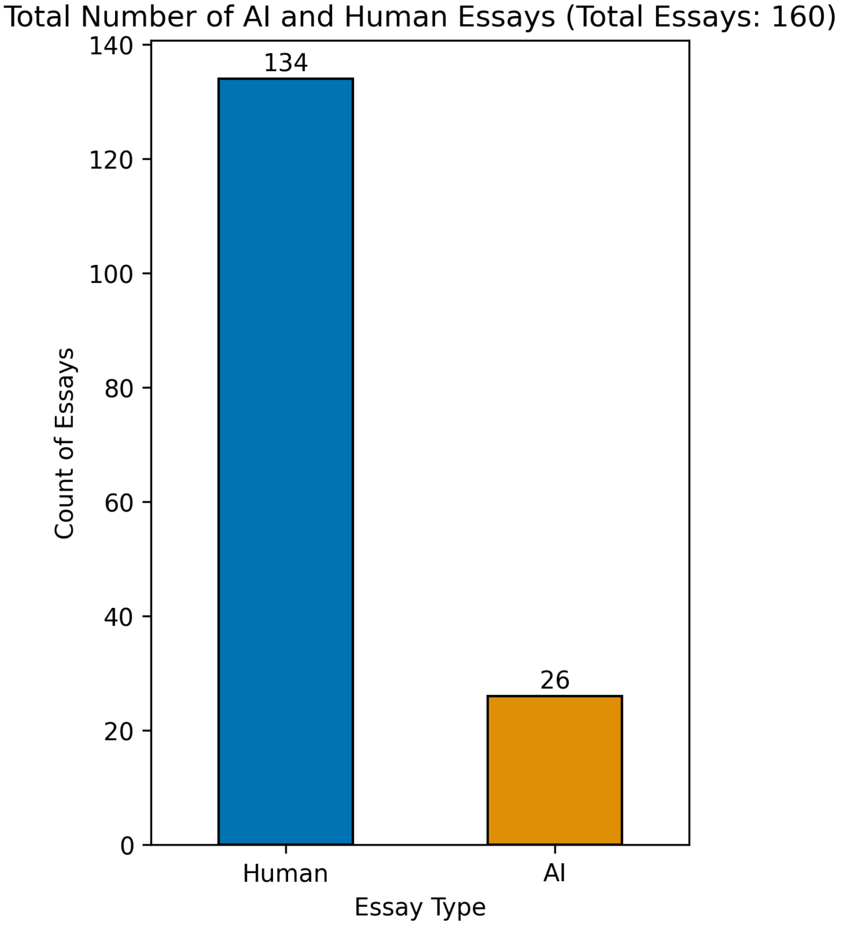
nice fed post you not catching me

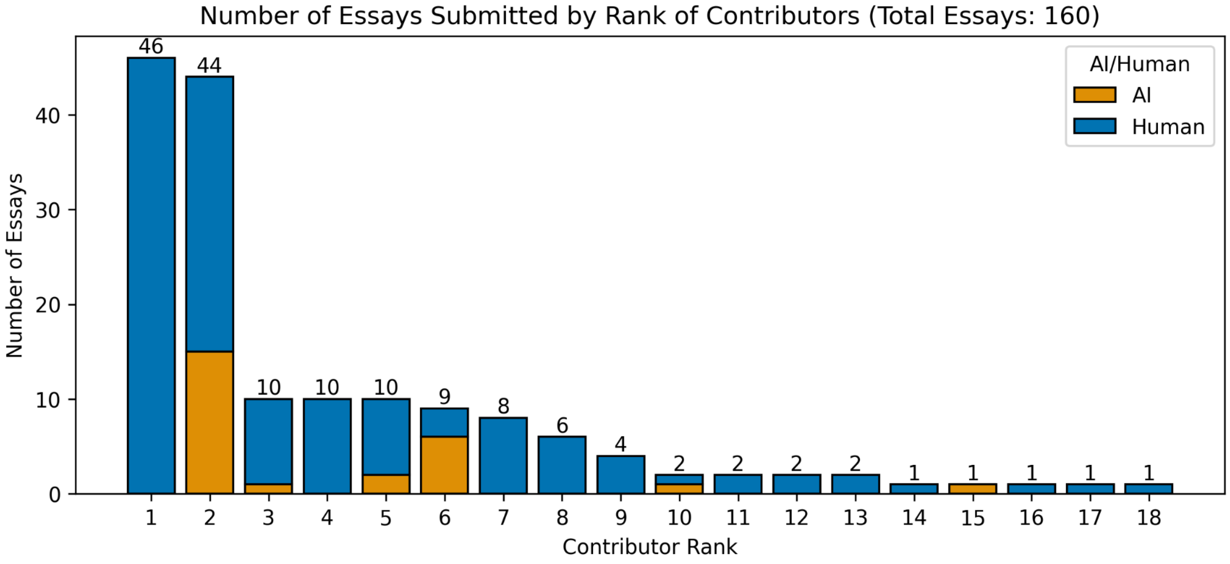
🧠 3

Attempt at recruiting participants in a Discord server

What was collected

During the essay collection period, starting November 5th 2023 and ending December 13th 2023, 160 total essays were collected. 134 essays were human written, and 26 were AI written.



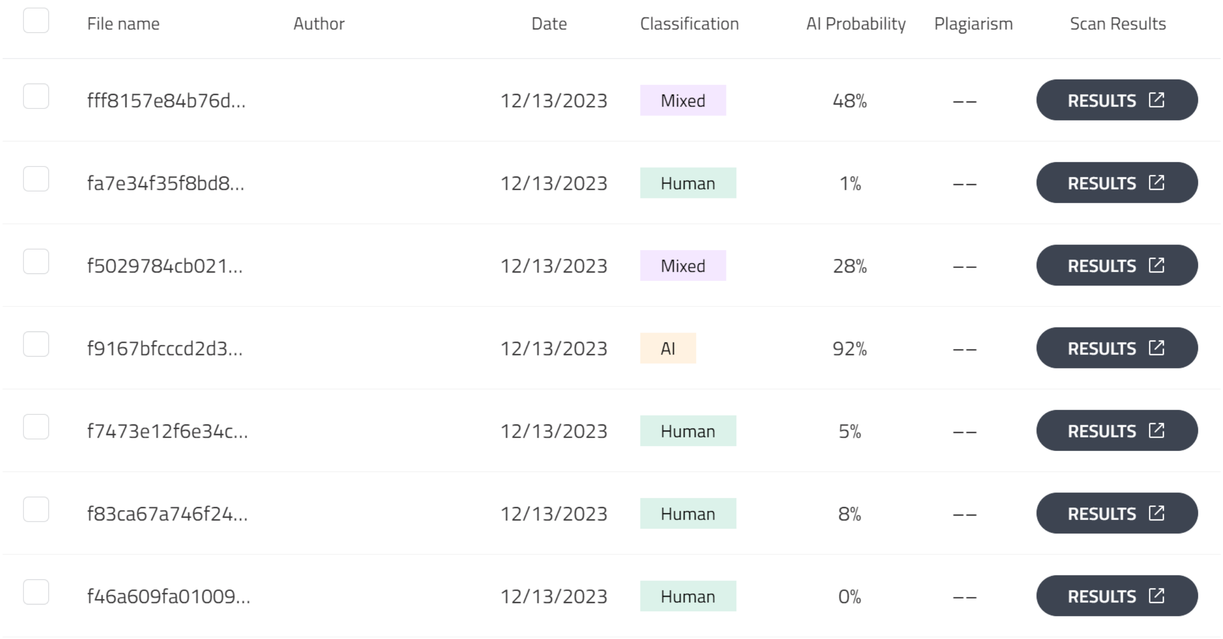


To maintain privacy of contributors, analysis will not be provided on who contributed what essays, but Neil Sorkin did contribute essays. He was one of 18 total contributors who provided a mix of human and AI essays. The Google Form only allowed 10 essay submissions per type (Human and AI written) per person. This resulted in a maximum of 20 essays per person. Some contributors were allowed to submit above this limit by directly speaking with the authors and providing the timestamp, which is the unique identifier for each contributor, that they submitted to be identified.

Data Analysis

File Processing and Upload

After a few steps of merging .csv files (utility scripts can be found here), removing any author metadata from the files, and changing the names of the files to hash strings to be uniquely identified later, the files were uploaded in batch to GPTZero. The results on the GPTZero web interface showed three different "Classification" labels: "Human", "Mixed", and "AI". In the .csv output, this was "HUMAN_ONLY", "MIXED", and "AI_ONLY" respectively.

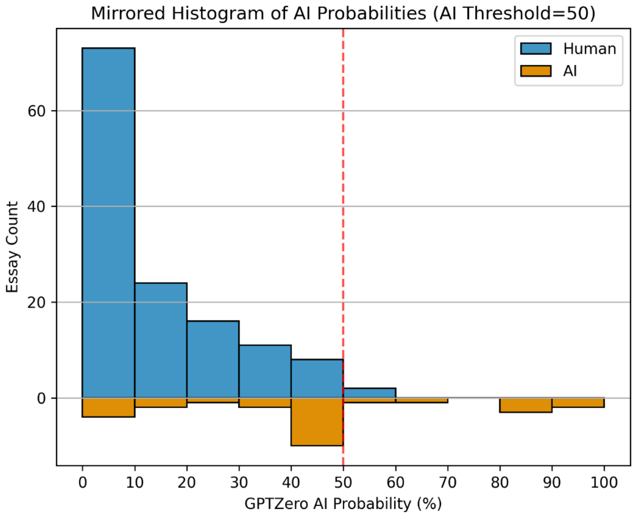


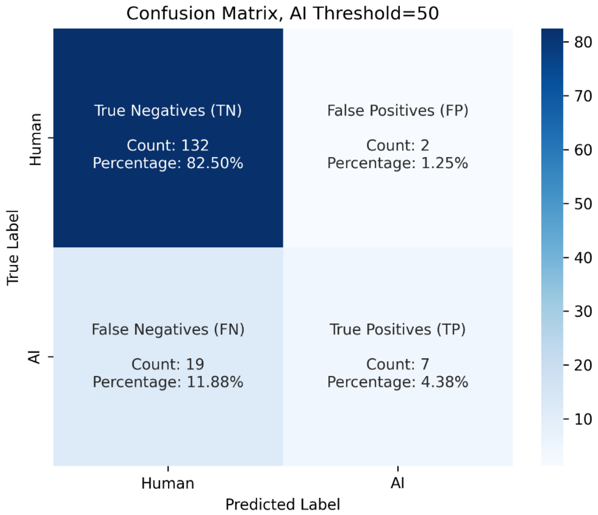
□	File name	Author	Date	Classification	AI Probability	Plagiarism	Scan Results
□	fff8157e84b76d...		12/13/2023	Mixed	48%	--	RESULTS ↗
□	fa7e34f35f8bd8...		12/13/2023	Human	1%	--	RESULTS ↗
□	f5029784cb021...		12/13/2023	Mixed	28%	--	RESULTS ↗
□	f9167bfcccd2d3...		12/13/2023	AI	92%	--	RESULTS ↗
□	f7473e12f6e34c...		12/13/2023	Human	5%	--	RESULTS ↗
□	f83ca67a746f24...		12/13/2023	Human	8%	--	RESULTS ↗
□	f46a609fa01009...		12/13/2023	Human	0%	--	RESULTS ↗

GPTZero interface after uploading all 160 essays

Thresholds and Confusion Matrices

Unfortunately, it is not as simple as saying “GPTZero” has a 99% accuracy. There is another parameter, which is the percentage at which a professor will take disciplinary action (honor council referral, points docked, etc.). This threshold can be visualized in the following plots as a vertical red dotted line. We will analyze GPTZero's performance at various thresholds, but from the lens of minimizing harm to students by avoiding false positives, where a student's human-written essay is incorrectly identified as AI-generated. With this in mind, let's analyze the outcomes at the 50% and 90% thresholds:

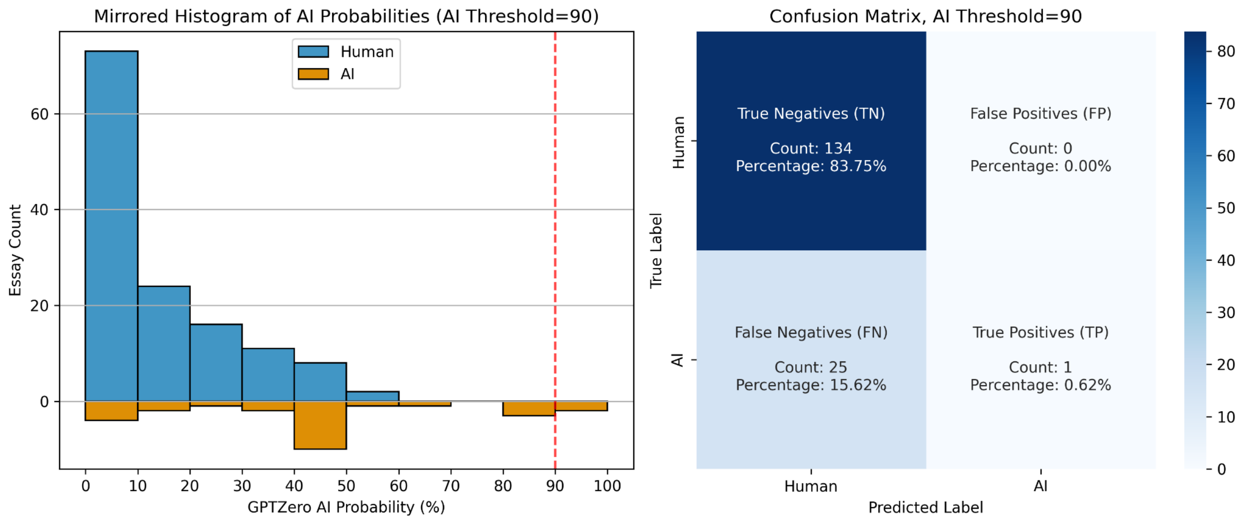




	Human	AI
Human	True Negatives (TN) Count: 132 Percentage: 82.50%	False Positives (FP) Count: 2 Percentage: 1.25%
AI	False Negatives (FN) Count: 19 Percentage: 11.88%	True Positives (TP) Count: 7 Percentage: 4.38%

At the 50% Threshold:

- True Negatives (TN): With 132 of 134 essays correctly identified as human-written in the top left quadrant.
- False Positives (FP): There are 2 instances of false positives in the top right quadrant. While low, any occurrence of false positives is concerning because of the potential for undue harm to students' academic records and stress.
- False Negatives (FN): The bottom left quadrant shows 19 of 26 false negatives. This suggests a weakness in detecting AI-generated essays, but from a standpoint of student well-being, it is less harmful than false positives.
- True Positives (TP): There are 7 true positives in the bottom right quadrant, indicating that some AI-generated essays are being correctly flagged.
- In this situation, the professor will accuse 9 students of using AI when only 7 of them have actually used AI. If all students deny using AI, then the professor will not know who to trust and may not have the time to decide for themselves, so all 9 may be referred to the honor council.



At the 90% Threshold:

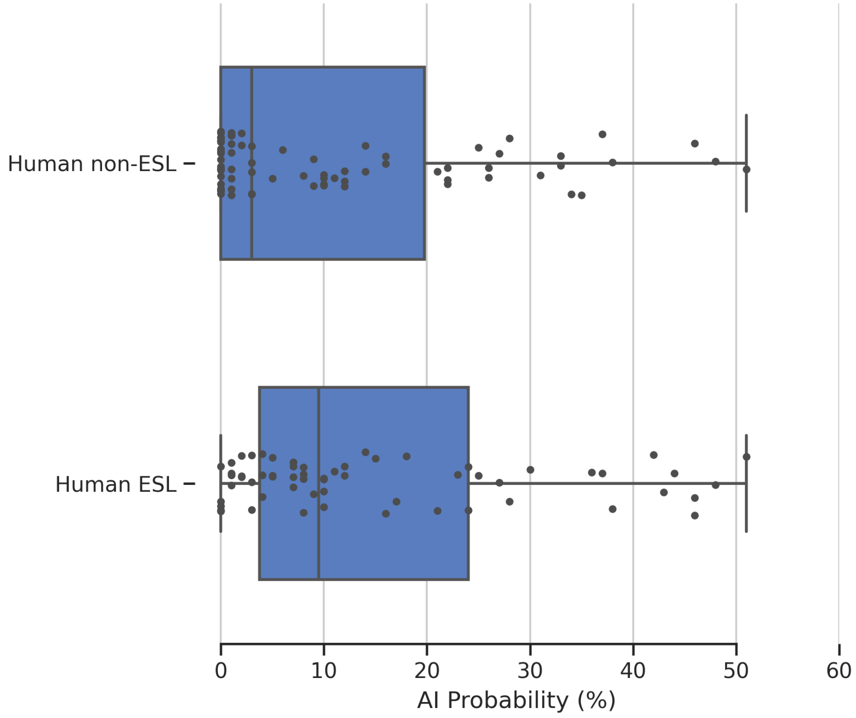
- True Negatives (TN): The count increases to 134 of 134 in the top left quadrant, which means the detector recognized all human writing correctly.
- False Positives (FP): Notably, the top right quadrant shows zero false positives. This is a crucial outcome, as it ensures that no student is incorrectly penalized for AI-generated content, aligning with the stance that even one false positive is too many.
- False Negatives (FN): There is an increase to 25 of 26 false negatives in the bottom left quadrant. Although this means some AI-generated content is missed, it is a trade-off for ensuring no student is falsely accused.
- True Positives (TP): The bottom right quadrant reveals a significant reduction to just 1 true positive, which may seem less effective but is the cost of ensuring fairness and protecting students from false accusations.
- In this situation, the professor will correctly accuse 1 student of using AI.

The comparison clearly shows that a higher threshold, such as 90%, is preferable when the priority is to protect students from the consequences of false positives. Any false positive can be too detrimental, potentially impacting a student's academic career and mental health. Therefore, it's crucial to err on the side of caution and adjust the threshold in such a way that the likelihood of false positives is minimized, even if this means accepting a higher number of false negatives. This stance prioritizes the ethical implications of using automated detection tools in educational settings and upholds the principle of "innocent until proven guilty."

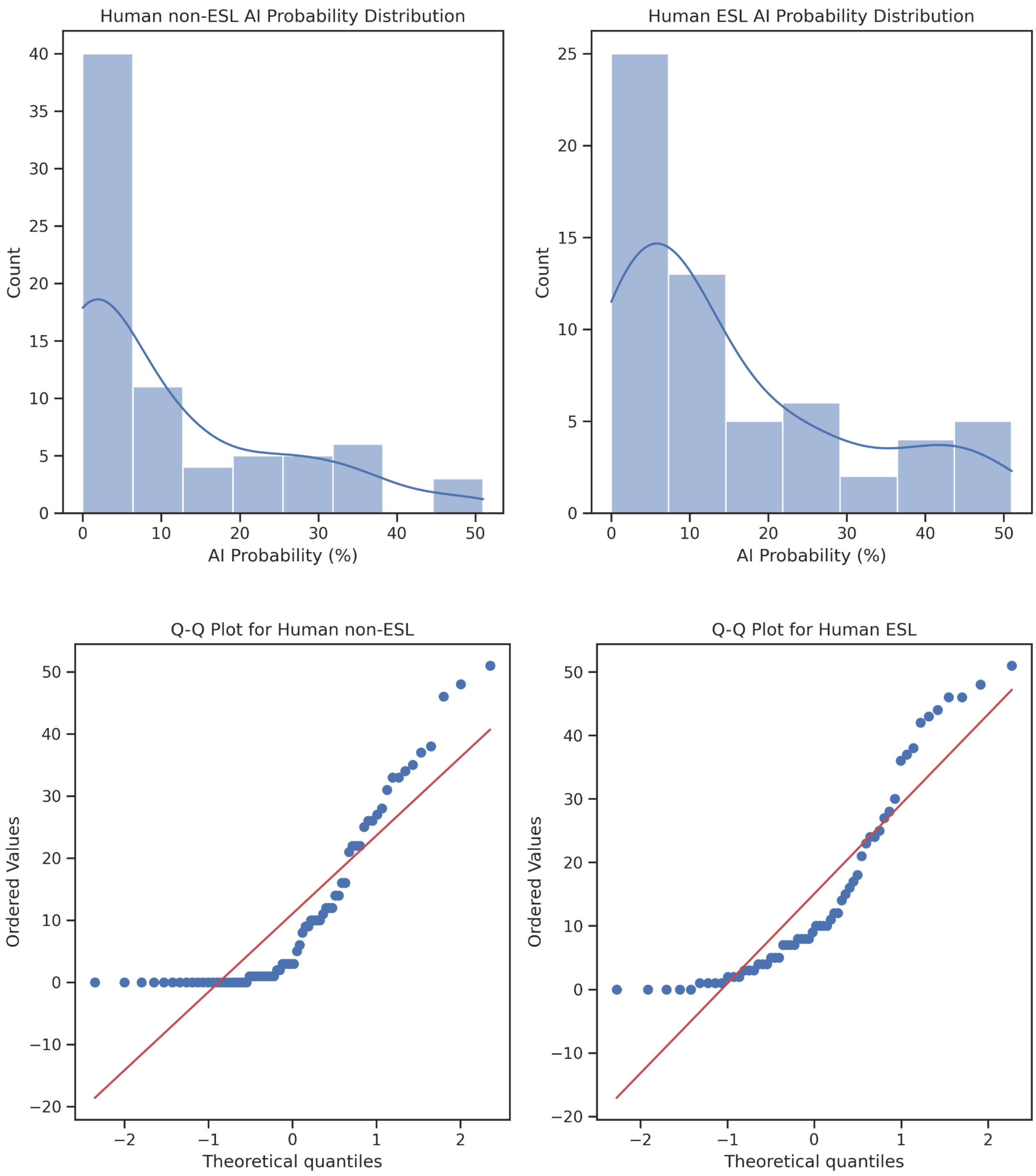
To be clear, these plots should not be taken to mean that as a professor using GPTZero, accuse people that have above 90% AI Probability score on their essays. The sample size is rather small to make decisions for an entire class of students. There may be thousands of submitted essays in a single semester of a course, so figuring out the edge cases will be very important before picking any acceptable threshold. Ultimately, more data is needed to conclusively pick a threshold.

Non-Native English Writers Bias

The paper “GPT detectors are biased against non-native English writers” inspired our research to investigate if there is a difference between the writing of non-native English writers/English as a second language (ESL) and native English writers.

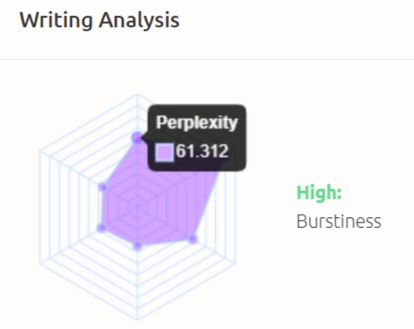


Comparison of AI Probability distributions for human written non-ESL and ESL students



The Mann-Whitney U statistical test was used to determine if the distributions were significantly different, which had a p-value of 0.0182. Both distributions are right skewed, so the Mann-Whitney U statistical test is appropriate in this case. This p-value implies that the AI Probability distributions for Human non-ESL and ESL groups are significantly different, agreeing with the Stanford study, but not to the same extreme as was seen in that case. This could be caused by the fact that our essays were much longer than their ESL TOEFL essays, which are generally between 150 and 225 words in length (LeapScholar, 2023).

The explanation for why these distributions differ was explained by an article titled “AI-Detectors Biased Against Non-Native English Writers” about the Stanford paper. They say AI text detectors use perplexity as a metric to determine if text is AI generated. Low perplexity is associated with AI writing and high perplexity is associated with human writing. However, “non-native speakers typically score lower on common perplexity measures such as lexical richness, lexical diversity, syntactic complexity, and grammatical complexity”.



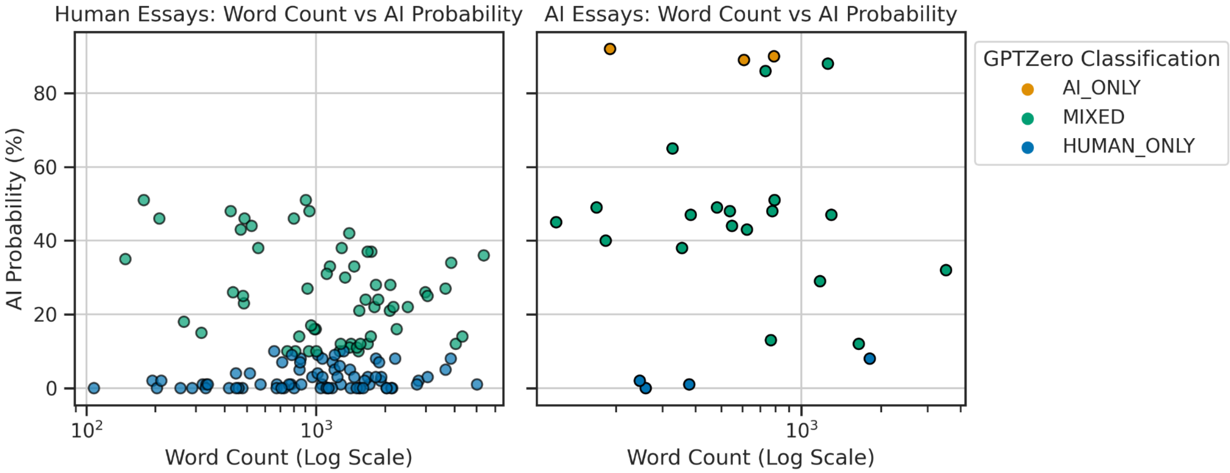
Writing Analysis

Perplexity
61.312

High:
Burstiness

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

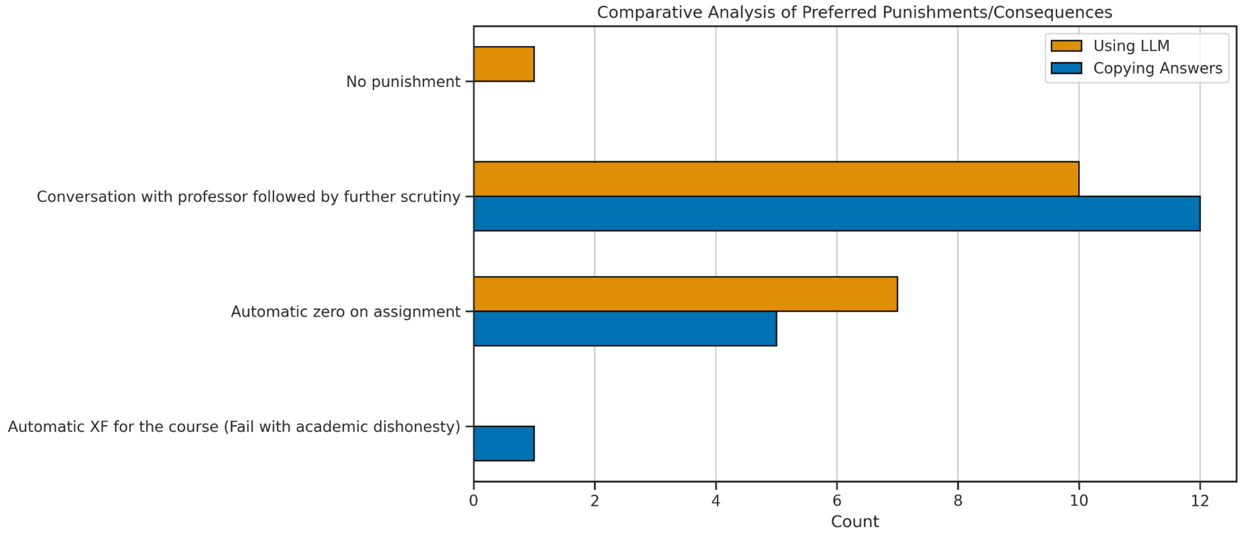
Word Count vs AI Probability



In examining the relationship between word count and AI probability, a Pearson correlation analysis yielded a coefficient r of -0.126, indicating a weak negative correlation. This suggests a slight decrease in AI probability with increasing word count. However, the coefficient of determination R^2 of 0.016 implies that only a small fraction (1.6%) of the variation in AI probability is explained by word count. Additionally, the correlation was not statistically significant (p-value = 0.114), suggesting that the observed relationship may be due to random variation. Therefore, the impact of word count on AI probability appears to be minimal and should be interpreted cautiously.

Initially, I would have expected that as GPTZero sees more of the essay, it will become more and more "sure" that it is AI or not AI, but that does not seem to be the case.

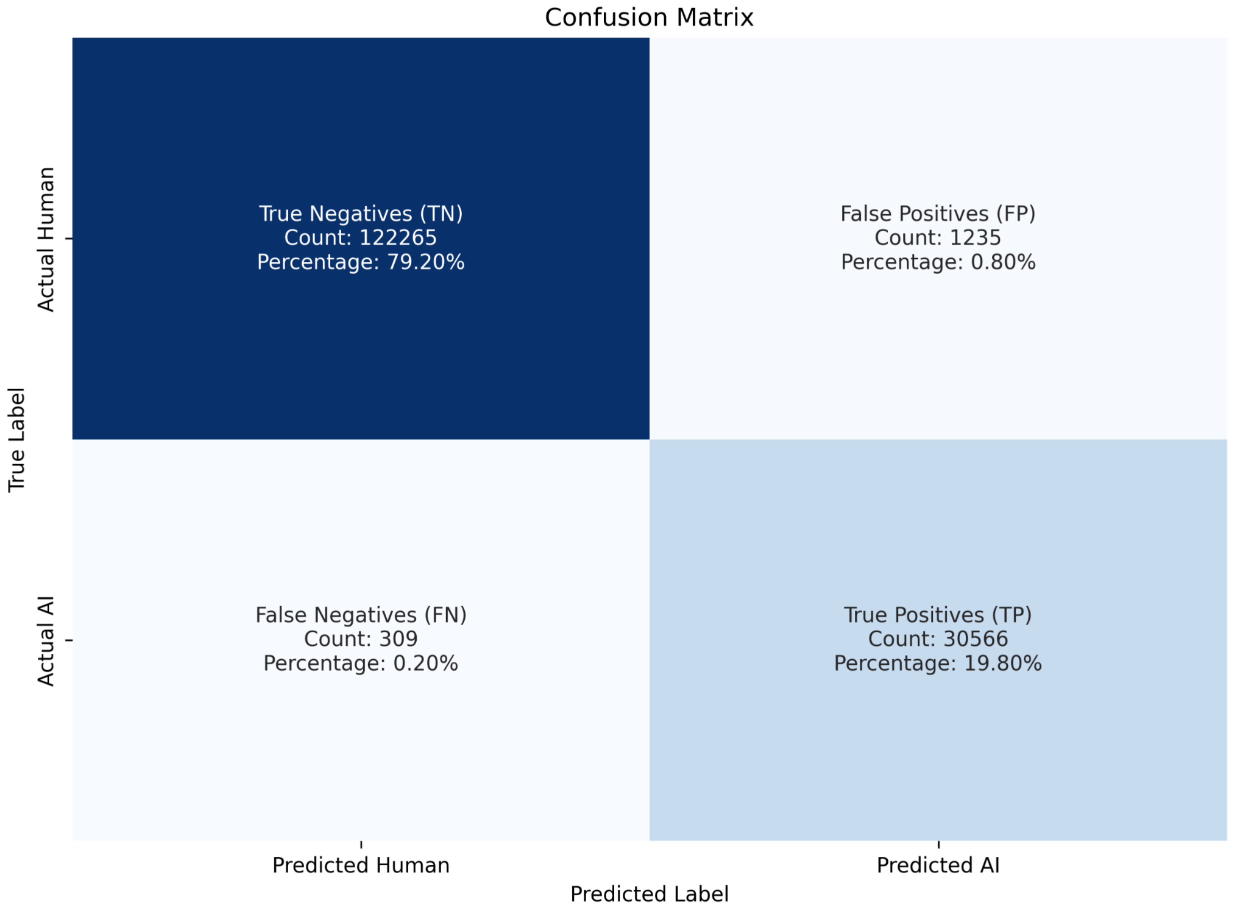
Punishment Opinions



As was mentioned earlier, the type of people who would complete this survey are more likely to be passionate about this topic, or be friends with the authors. However, what the opinion portion of the survey shows is that people prefer a less harsh punishment for people who use AI/LLMs on assignments compared to people who copy answers directly from another student. This portion of the survey could likely be conducted more quickly with much less level of effort if made separate from the essay attachment portion of the survey.

Discussion

Example of AI Text Detectors at Scale



The confusion matrix displayed provides a visual summary of the performance of an AI detection tool based on the University of Maryland's undergraduate student body, with the following assumptions:

- Each of the 30,875 students writes 5 essays per year, leading to a total of 154,375 essay submissions.
- The AI detection tool has a 1% false positive rate, which results in 1,235 human-written essays being incorrectly flagged as AI-generated. This is particularly significant as it represents students who would be wrongfully accused of using AI in their essays.

- An estimated 20% of the essays are assumed to be AI-generated, amounting to 30,875 essays. With a 99% true positive rate, the AI detection tool correctly identifies 30,566 of these as AI-generated.

The matrix shows a high number of true negatives (122,265 essays), indicating the tool's effectiveness in correctly identifying human-written essays. However, the presence of false negatives (309 essays) highlights instances where AI-generated essays are not detected. The key insight from this matrix is the stark reality of the impact of even a small false positive rate; when applied to the scale of a large university, it can have a considerable effect, leading to many students potentially facing unfounded academic integrity allegations. This underscores the importance of striving for extremely high accuracy in AI detection tools to protect student welfare and academic integrity.

While not shown, if the percentage of students using AI is 0%, then all accusations will be false accusations. As less students use AI on assignments, the number of false positives increases, but not the total number of accusations.

Conclusion

The evaluation of GPTZero has provided critical insights into the difficulties associated with using such detectors. While these tools offer potential in identifying AI-generated content, the balance between false positives and false negatives remains a delicate one. Adopting a cautious approach mitigates the risk of false positives but increases false negatives, and the inverse is true when the system is configured more aggressively.

The researchers believe that a single false positive, wrongfully accusing a student of academic dishonesty, far outweighs the cost of a false negative, where an AI-generated document is not detected. This is especially true when students who wish to attend graduate school may not be able to, based on them receiving an XF (failure with academic dishonesty) in a course.

Additionally, our study reaffirms previous findings that GPTZero, like other detectors, exhibits a bias against ESL writers. This bias raises significant concerns about fairness and equity, particularly for all 4,225 international students at the University of Maryland during the Fall 2023 semester (University of Maryland, n.d.).

More data is needed, and there is a need to fill the current gaps in understanding. Future studies, potentially with cash or extra credit incentivization for participation with stronger anonymity, could provide a more comprehensive dataset that can further refine the accuracy and fairness of these detectors.

Based on the findings, the recommendation is to exercise extreme caution in the deployment of AI text detectors within academic institutions. Until such tools can guarantee a false positive rate of no more than 0.01%, they should not be used as a definitive measure of academic integrity. Moreover, there should be a standing requirement for regular, independent re-evaluation of these tools to ensure they remain up-to-date with the evolving capabilities of AI text-generation technologies and to maintain a standard of fairness in their application.

References

Beam, C. (2023, September 14). The AI Detection Arms Race Is On. *Wired*. https://www.wired.com/story/ai-detection-chat-gpt-college-students/

Bhide, V. M. (2007). Game Theory and the Law - Plea Bargaining. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1013189

Cornell Law School. (n.d.). *plea bargain*. LII / Legal Information Institute. https://www.law.cornell.edu/wex/plea_bargain#

Devers, L. (2011). *Plea and Charge Bargaining*. https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/PleaBargainingResearchSummary.pdf

Gaur, A. (2023, November 27). *UMD records at least 50 AI-related academic integrity cases in 2022-23*. The Diamondback. https://dbknews.com/2023/11/27/umd-chatgpt-academic-integrity-cases/

Inazu, J. D. (2016). *Confident Pluralism: Surviving and Thriving through Deep Difference*. The University Of Chicago Press.

Jillson, E. (2021, April 19). *Aiming for truth, fairness, and equity in your company's use of AI*. Federal Trade Commission. https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

Koppelman, K. L. (2020). *Understanding Human Differences: Multicultural Education for a Diverse America* (6th ed.). Pearson Education.

LeapScholar. (2023, January 20). *TOEFL Writing Templates 2023: TOEFL Essays Simplified*. LeapScholar. https://leapscholar.com/blog/toefl-writing-templates-2022-2

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779–100779. https://doi.org/10.1016/j.patter.2023.100779

Myers, A. (2023, May 15). *AI-Detectors Biased Against Non-Native English Writers*. Stanford HAI. https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers

Sadasivan, V., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-Generated Text be Reliably Detected?* https://arxiv.org/pdf/2303.11156.pdf

Tian, E., & Cui, A. (2023). *GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods*. Gptzero.me; GPTZero. https://gptzero.me/

University of Maryland. (n.d.). *Reports & Statistics | Maryland Global*. Marylandglobal.umd.edu; Maryland Global. Retrieved December 16, 2023, from https://marylandglobal.umd.edu/global-learning-all/international-students-scholars/connect-isss/reports-statistics

University of Maryland, Division of Research. (n.d.). *Facts and Figures*. Division of Research. Retrieved December 16, 2023, from https://research.umd.edu/who-we-are/facts-and-figures

Weber-Wulff, D., Bjelobaba, S., Guerrero-Dib, J., Šigut, P., & Waddington, L. (2023). *Testing of Detection Tools for AI-Generated Text*. https://arxiv.org/ftp/arxiv/papers/2306/2306.15666.pdf

Williams, R. (2023, July 7). *AI-text detection tools are really easy to fool*. MIT Technology Review. https://www.technologyreview.com/2023/07/07/1075982/ai-text-detection-tools-are-really-easy-to-fool/