Detecting Generative AI Usage in Application Essays

NEIL NATARAJAN, University of Oxford, UK
ELÍAS SÁNCHEZ HANNO, Schmidt Futures, USA
LOGAN GITTELSON, Schmidt Futures, USA

Student use of generative AI in essay-writing creates new challenges for education in the marking of essays and essay-based selection for scholarships, fellowships, and universities. In theory, new software purporting to detect AI-generated content offers a plausible solution, but little research has attempted to validate such solutions in real-world situations. We present a case study exploring the efficacy and implications of using one such detection product, GPTZero, in the selection process for an anonymous talent identification program that finds promising young people and provides them with opportunities that allow them to work together to serve others. We determine that GPTZero achieves good predictive performance at a variety of thresholds but suffers at extremely low false-positive tolerances. Also, its scores are heterogeneously biased across geographical and gender groups. However, we find an overall high AUC score for GPTZero's statistic, and we demonstrate useful aggregate analyses of the program's application process, wherein we find evidence for only limited use of AI-generated text in the most recent program's application cycle. Thus, we conclude that GPTZero does not perform sufficiently well to merit disqualifying applicants on its basis, but that it yields valuable insights into generative AI's impact on application processes.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Applied computing → Education**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: Generative Artificial Intelligence, Generative Artificial Intelligence Detection, Education

1 INTRODUCTION

Given the mass-market adoption of OpenAI's ChatGPT and student use thereof, scholarship programs, universities, and talent identification organisations that rely on essays must reconsider how essays factor into the selection process [2, 3]. Such reconsideration may lead some programs to encourage AI-generated text so long as it is truthful and well-written, while others may ban the use of generative AI altogether; in both cases, it is useful to be able to determine which applicants submitted essays written or augmented by AI. Researchers have developed methods such as DetectGPT and stylometric detection to make this determination, while commercial approaches include GPTZero, Turnitin, and Originality.ai, among others [4, 5, 8, 11, 12]. All have significant practical limitations – multiple experiments demonstrate even state-of-the-art detection methods to be both beatable and bias [4, 6, 13] – but there is limited evidence evaluating their *utility* in real settings [4, 6]. We present a case study illustrating the utility and limitations of AI detection using GPTZero. We make the following contributions: (a) we show that we cannot identify AI-generated content with high enough sensitivity to use said identifications for high-stakes decision making, (b) we confirm that our detector disproportionately identifies certain subgroups' genuinely human-written content as AI generated, and (c) we demonstrate that organisations can use AI detection technology to extract useful aggregate insights.

2 DATA

2.1 Applicant Data

We partner with an anonymous talent identification program that runs an application process to find promising young people from around the world. We use data from two of the program's application cycles, which we term Cycle 2022

and Cycle 2023. Each applicant submitted essays in response to five prompts. These prompts are not described in more detail to preserve our partner organisation's anonymity. Applicants additionally provided their gender identity and their countries of citizenship, the latter of which we have grouped into larger regional categories with shared cultural heritages. Applications for Cycle 2022 were due in early 2022, well before ChatGPT's public release, so we assume that these submissions were written without the use of generative AI [10]. However, applications for Cycle 2023 were due in early 2023, after generative AI tools had become widely available, so we cannot make this same assumption [5, 7, 12]. We use a total of 15, 149 human-written Cycle 2022 essays and 24, 815 human-written Cycle 2023 essays in our analysis. We additionally generate 5, 002 synthetic essays using OpenAI's ChatGPT API and Cycle 2022 prompts [1].

2.2 Statistic Calibration

We calculated likelihoods of each of our essays being AI-generated using an industry-leading generative AI detection tool, GPTZero [12]. GPTZero yielded various statistics for each essay, but we are primarily interested in overall likelihood *completely_generated_prob*.

However, though *completely_generated_prob* contains the suffix *_prob*, we are not guaranteed by GPTZero that this value is, in fact, a probability. Thus, we can use our large body of essays of known provenance (15, 149 applicant-submitted Cycle 2022 essays and 5, 002 researcher-prompted ChatGPT-generated essays) we can compare GPTZero estimates to actual probabilities to confirm whether the statistic is a well-calibrated probability estimate (I.e., if *completely_generated_prob* is a probability, we should find an average *completely_generated_prob* value of our entire provenance near to $\frac{5{,}002}{5{,}002+15{,}149}$) [9].

Figure 1 presents a calibration curve demonstrating large differences between GPTZero's predicted probability and fraction of true positives in a variety of samples of our data, and thus we conclude that *completely_generated_prob* is not well-calibrated on our data. We calibrate our statistics with Cycle 2022 data (real and synthetic), yielding $P(AI)$, the probability that an essay is completely AI-generated. We use this calibrated statistic in all subsequent analyses.

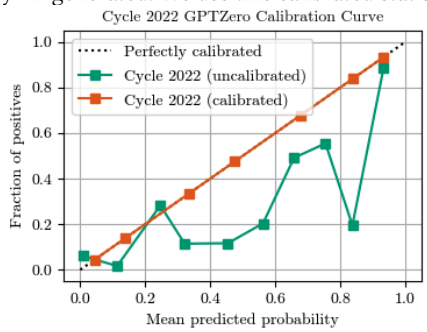

Fig. 1. Calibration Curve of GPTZero's Predicted Probability AI-Generated, Cycle 2022

3 RESULTS

3.1 Disqualifying Applicants via AI Detection

To determine how accurately we can identify AI-generated content, we compare the output statistic from our detector with actual knowledge of whether an essay was human- or computer-generated in Cycle 2022. We draw a receiver operating characteristic (ROC) curve and report the area under that curve in Figure 2. The area under this ROC curve is a relatively high 0.91, indicating good predictive performance at a variety of thresholds. However, as marked by '0.02,

0.66' in Figure 2, at low false positive rates (E.g., 0.02), GPTZero fails to identify more than a third of AI-generated essays.

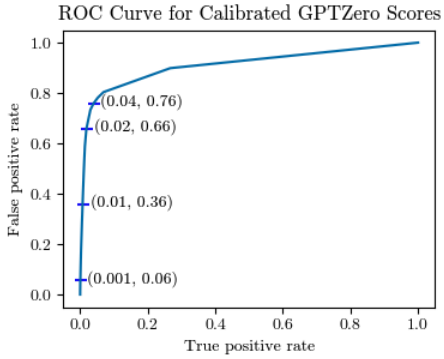

Fig. 2. Receiver Operating Characteristic Curve of Calibrated GPTZero Scores, Cycle 2022

We next evaluate whether GPTZero is biased, on average, against genuine submissions from applicants with specific backgrounds. We conduct analyses of variance (ANOVAs) of the calibrated probability across gender and region categories in applicant-submitted essays from Cycle 2022 in Table 1. As we know this data to be human-generated, this tests whether our detection method has heterogeneous false positive rates.

Table 1. Analysis of Variance for $P(AI)$, Cycle 2022

Dimension	F Statistic	p Value
Gender Identity	41.4	< 0.01
Region	62.9	< 0.01

Table 1 shows statistically significant variation across both dimensions in Cycle 2022. Column 3 of Table 2 reports that, on human-written Cycle 2022 data, male applicants' essays receive slightly lower estimates, on average, than applicants identifying as female or reporting some other gender category, which raises the risk that using such scores in decision-making would bias outcomes against content genuinely created by women. Column 3 of Table 2 also show genuinely human-generated applications from the Five Eyes – a group consisting of Australia, Canada, New Zealand, the United Kingdom, and the United States – get above average GTPZero scores. This is consistent with the finding that most large language models are trained primarily on texts from these countries, and thus produce output more closely resembling these texts [1]. However, it challenges previous work suggesting AI detectors are biased against non-native English speakers [6].

Despite good overall accuracy, GPTZero does not perform sufficiently well for the program to disqualify applicants on its basis alone. The program selected only 2% of completed applications, so any threshold with a false positive rate of 0.02 or more would risk rejecting as many qualified candidates on the basis of erroneous AI detection as were ultimately admitted. The '0.02, 0.66' threshold in Figure 2 demonstrates that a 2% false positive rate fails to identify more than a third of AI-generated essays. This would be untenable for the program, especially given the introduction of new algorithmic biases. We conclude that the program should not use GPTZero in high-stakes decisions such as disqualifying a candidate. However, $P(AI)$ may still be useful in determining which essays warrant closer inspection for veracity and plagiarism, and it's value in aggregate analyses is yet-underexamined.

Table 2. Calibrated Probability AI-Generated and Inter-Cycle Change by Applicant Demographics

Demographic Group	Cycle 2022 $\widehat{E(P(AI))}$	Cycle 2023 $\widehat{E(P(AI))}$	Inter-Cycle Δ t Score	p Value
Male	0.09	0.11	8.63	< 0.01
Female	0.11	0.11	−0.40	0.69
Other	0.14	0.12	−1.10	0.27
Caribbean	0.17	0.15	−0.47	0.64
East/Southeast Asia	0.13	0.13	0.74	0.46
Five Eyes	0.21	0.14	−5.86	< 0.01
Former Soviet Union	0.13	0.15	0.59	0.56
Indian Subcontinent	0.08	0.09	2.65	0.01
Latin America	0.13	0.09	−1.44	0.15
Mid East/North Africa	0.11	0.12	1.67	0.09
Sub-Saharan Africa	0.09	0.09	2.15	0.03
Other Europe	0.20	0.12	−2.75	0.01
Pacific Islands	0.04	N/A		
All Submissions	0.10	0.11		

'Five Eyes' consists of Australia, Canada, New Zealand, the UK, and the US.

3.2 The Analytic Utility of GPTZero

We focus our subsequent analysis primarily on the potential use of generative AI by applicants in Cycle 2023. Seeking to avoid the disproportionate effects noted above, we focus primarily on within-group changes. We note here that the mean probability of an essay being AI-generated within a corpus is exactly the expected proportion of AI-generated content within that corpus. Thus, we test for changes in mean $P(AI)$. Table 2 presents mean $P(AI)$ for 2023 (column 3) and 2022 (column 5), as well as test statistics of whether there is a difference in means (columns 6 and 7). Because this is an exploratory analysis, statistics are not adjusted for multiple hypothesis testing.

We find statistically significant increases in the $P(AI)$ for only two overlapping subgroups: male applicants and applicants from the Indian subcontinent. In both cases the magnitude of the increase is small, suggesting that at least in Cycle 2023, the use of generative AI was limited. However, other findings preclude interpreting this change as a direct measure of increased generative AI use. In two regions, the Five Eyes and Europe (excluding the United Kingdom and former Soviet Union), we found a statistically significant decrease in the calibrated estimated probability that essays were completely AI generated. Since we can assume very few if any applicants in Cycle 2022 had access to generative AI, this cannot be interpreted as a decrease in AI use. It may be that both regions' high average scores in 2022 were a fluke of the cohort, and that these regions reverted to the mean in 2023. Alternatively, it is possible that, in these regions, Cycle 2023 applicants used AI detection tools to ensure that their content would not be flagged by our detector (although this would require such detector usage to offset any actual generative AI use) [12]. In either case, this analysis surfaces interesting discrepancies demanding further interrogation in future cycles.

4 DISCUSSION

Our natural experiment on Cycle 2022 data enjoyed optimal conditions for AI detection; the synthetic essays were not paraphrased or edited, and all of the human-written essays were real submissions to the program. Despite this, at sensitivity levels comparable to acceptance rates into selective programs, we do not achieve high specificity. Furthermore,

we find heterogeneous biases in our detection of AI essays. Thus, we conclude that the program should not use this technology in making high-stakes decisions such as disqualifying candidates. We do, however, see good performance at a variety of thresholds. Therefore, we find utility in using AI detection for aggregate analyses, so long as these analyses account for the risk of heterogeneous biases. We demonstrate one such analysis on data supplied by the program: examining the potential scale of AI usage by demographic group. We conclude that organisations looking to assess the use of generative AI in essays they receive have an opportunity to do so with GPTZero or other AI detection tools.

This analysis is limited in scope to only one method for AI generation and one method for identification of that generation. Interesting avenues for future analyses include using other methods of essay generation and AI detection and seeking out differential performance on essays in response to particular prompts. In particular, we intend further work to focus on understanding the effect of paraphrasing software on our results, as paraphrased models have been shown to vary [4, 8]. We also hope to expand our analysis to multiple detection tools and subsequent application cycles.

REFERENCES

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs].

[2] Catherine A. Gao, Frederick M. Howard, N. Markov, E. Dyer, S. Ramesh, Yuan Luo, and Alexander T. Pearson. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv* (2022). https://doi.org/10.1101/2022.12.23.521610 S2ID: b36acdfc67612d707c95d1ed282672d3ca262be7.

[3] Guangwei Hu. 2023. Challenges for enforcing editorial policies on AI-generated papers. *Accountability in Research* 0, 0 (Feb. 2023), 1–3. https://doi.org/10.1080/08989621.2023.2184262 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08989621.2023.2184262.

[4] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, J. Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. (2023). ARXIV_ID: 2303.13408 S2ID: 2969e8a14237f8244d3c825ff19bdfb3cc7fddf1.

[5] Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike. 2023. New AI classifier for indicating AI-written text. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

[6] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4, 7 (2023), 100779. https://doi.org/10.1016/j.patter.2023.100779

[7] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. https://doi.org/10.48550/arXiv.2303.11032 arXiv:2303.11032 [cs].

[8] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. https://doi.org/10.48550/arXiv.2301.11305 arXiv:2301.11305 [cs].

[9] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning (ICML '05)*. Association for Computing Machinery, New York, NY, USA, 625–632. https://doi.org/10.1145/1102351.1102430

[10] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

[11] Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, K. Trapeznikov, Scott W. Ruston, and Huan Liu. 2023. Stylometric Detection of AI-Generated Text in Twitter Timelines. *ArXiv* (2023). https://doi.org/10.48550/arxiv.2303.03697 ARXIV_ID: 2303.03697 S2ID: 18e0b11dd5b1b413d33308da0379836752aaaec1.

[12] Edward Tian and Alexander Cui. 2023. GPTZero | Technology. https://gptzero.me/

[13] Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and S. Feizi. 2023. Can AI-Generated Text be Reliably Detected? *ArXiv* (2023). https://doi.org/10.48550/arxiv.2303.11156 ARXIV_ID: 2303.11156 S2ID: fb47aa3c541fc2a9b340c9d2a3572860811767d6.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009