CHATGPT DETECTORS ARE BIASED AGAINST NON-NATIVE ENGLISH WRITERS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid adoption of generative language models has brought about substantial advancements in digital communication, while simultaneously raising concerns regarding the potential misuse of AI-generated content. Although numerous detection methods have been proposed to differentiate between AI and human-generated content, the fairness and robustness of these detectors remain underexplored. In this study, we evaluate the performance of several widely-used GPT detectors using writing samples from native and non-native English writers. Our findings reveal that these detectors consistently misclassify non-native English writing samples as AI-generated, whereas native writing samples are accurately identified. Furthermore, we demonstrate that simple prompting strategies can not only mitigate this bias but also effectively bypass GPT detectors, suggesting that GPT detectors may unintentionally penalize writers with constrained linguistic expressions. Our results call for a broader conversation about the ethical implications of deploying ChatGPT content detectors and caution against their use in evaluative or educational settings, particularly when they may inadvertently penalize or exclude non-native English speakers from the global discourse.

INTRODUCTION

Generative language models based on GPT, such as ChatGPT (OpenAI, 2022), have taken the world by storm. Within a mere two months of its launch, ChatGPT attracted over 100 million monthly active users, making it one of the fastest-growing consumer internet applications in history (Hu, 2023; Paris, 2023). While these powerful models offer immense potential for enhancing productivity and creativity (Lee et al., 2022; Kung et al., 2023; Terwiesch, 2023), they also introduce the risk of AI-generated content being passed off as human-written, which may lead to potential harms, such as the spread of fake content and exam cheating (Else, 2023; Gao et al., 2022; Kreps et al., 2022; Editorial, 2023; ICML, 2023).

Recent studies reveal the challenges humans face in detecting AI-generated content, emphasizing the urgent need for effective detection methods (Else, 2023; Gao et al., 2022; Kreps et al., 2022; Clark et al., 2021). Although several publicly available GPT detectors have been developed to mitigate the risks associated with AI-generated content, their effectiveness and reliability remain uncertain due to limited evaluation (OpenAI, 2019; Jawahar et al., 2020; Fagni et al., 2021; Ippolito et al., 2019; Mitchell et al., 2023; Solaiman et al., 2019; Gehrmann et al., 2019; Heikkil"a, 2022; Crothers et al., 2022). This lack of understanding is particularly concerning given the potentially damaging consequences of misidentifying human-written content as AI-generated, especially in educational settings (Rosenblatt, 2023; Kasneci et al., 2023).

Given the transformative impact of generative language models and the potential risks associated with their misuse, developing trustworthy and accurate detection methods is crucial. In this study, we evaluate several publicly available GPT detectors on writing samples from native and non-native English writers. We uncover a concerning pattern: GPT detectors consistently misclassify non-native English writing samples as AI-generated while not making the same mistakes for native writing samples. Further investigation reveals that simply prompting GPT to generate more linguistically diverse versions of the non-native samples effectively removes this bias, suggesting that GPT detectors may inadvertently penalize writers with limited linguistic expressions.

Our findings emphasize the need for increased focus on the fairness and robustness of GPT detectors, as overlooking their biases may lead to unintended consequences, such as the marginalization of non-native speakers in evaluative or educational settings. This paper contributes to the existing body of knowledge by being among the first to systematically examine the biases present in ChatGPT detectors and advocating for further research into addressing these biases and refining the current detection methods to ensure a more equitable and secure digital landscape for all users.

RESULTS

GPT DETECTORS EXHIBIT BIAS AGAINST NON-NATIVE ENGLISH AUTHORS

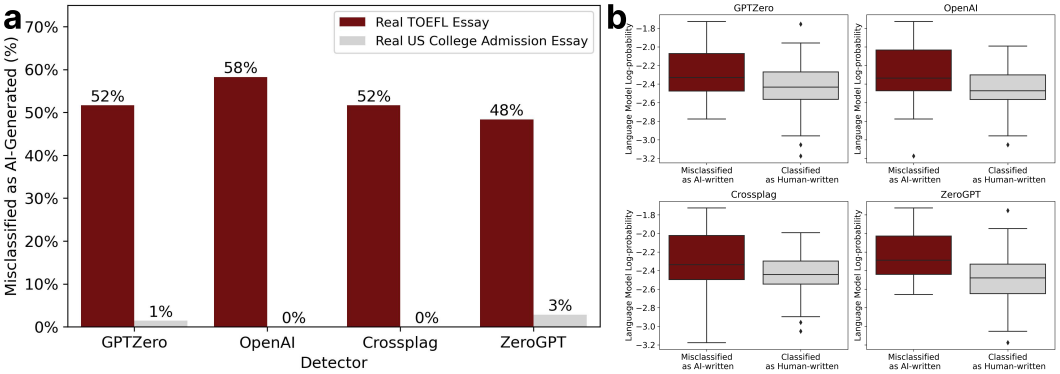

Figure 1: **GPT detectors demonstrate biases against non-native English writing samples.** (**a**) Comparison of performance among four widely-utilized GPT detectors using a corpus of 91 non-native-authored TOEFL essays (TOEFL: Test of English as a Foreign Language) and 70 US college admission essays by native speakers. Over half of the TOEFL essays are incorrectly classified as "AI-generated," while detectors display near-flawless accuracy for college essays with only 0-2 false positives. (**b**) Analysis of the correlation between language model log-probability (assessed on text-davinci-002, a GPT-3.5 model) and the probability of TOEFL essay misclassification as AI-generated. Essays that are misidentified tend to exhibit higher log-probability (i.e., lower perplexity), implying that GPT detectors might penalize authors with restricted linguistic expressions.

We assessed the performance of four widely-employed GPT detectors using a corpus of 91 human-authored TOEFL essays obtained from a Chinese educational forum and 70 US college admission essays sourced from PrepScholar (**Fig. 1**a). Notably, while detectors display near-flawless accuracy for US college admission essays (0-2 false positives out of 70), over half of the TOEFL essays were misclassified as "AI-generated" (average false positive rate: 52.47%). All four detectors unanimously identified 22 of the 91 essays (24.18%) as AI-authored, while at least one detector flagged 72 of the 91 essays (79.12%) as AI-generated.

To further investigate this phenomenon, we analyzed the correlation between language model log-probability (measured on text-davinci-002, a GPT-3.5 model) and the likelihood of TOEFL essay misclassification as AI-generated (**Fig. 1**b). We found that misidentified essays tend to have higher log-probability (i.e., lower perplexity), suggesting that GPT detectors may penalize non-native writers with limited linguistic expressions.

MITIGATING BIAS THROUGH LINGUISTIC DIVERSITY ENHANCEMENT OF NON-NATIVE SAMPLES

We conjectured that the restricted linguistic variability and word choices characteristic of non-native English writers contribute to the observed bias. To investigate this hypothesis, we employed GPT-4 to enrich the language in the essays, with the aim of emulating the vocabulary usage of native speakers (Prompt: *"Enhance the word choices to sound more like that of a native speaker."*). Remarkably, this intervention led to a substantial reduction in misclassification, with the average false positive rate decreasing by 41.76% (from 52.47% to 10.71%). Post-intervention, only 1 out of 91 essays (1.10%) was unanimously detected as AI-written.

Our hypothesis was further corroborated by the notable increase in the essays' average perplexity after the GPT-4 intervention, consistently observed across four language models (Davinci (GPT-3 base model) (p-value=1.3123E-07), text-davinci-002 (GPT-3.5 model) (p-value=8.4286E-10), GPT2-XL (p-value=9.3612E-05), GPT2 (p-value=3.3135E-10)). This observation indicates that essays authored by non-native writers inherently exhibit reduced linguistic variability compared to those penned by native speakers, which results in their misclassification as AI-generated text. Thus, our findings underscore the critical need to account for potential biases against non-native writers when employing perplexity-based detection methods. In light of these results, practitioners are urged to exercise caution when using low perplexity as an indicator of AI-generated text, as this approach might inadvertently perpetuate systematic biases against non-native authors.

SIMPLE PROMPT CAN EASILY BYPASS CURRENT GPT DETECTORS

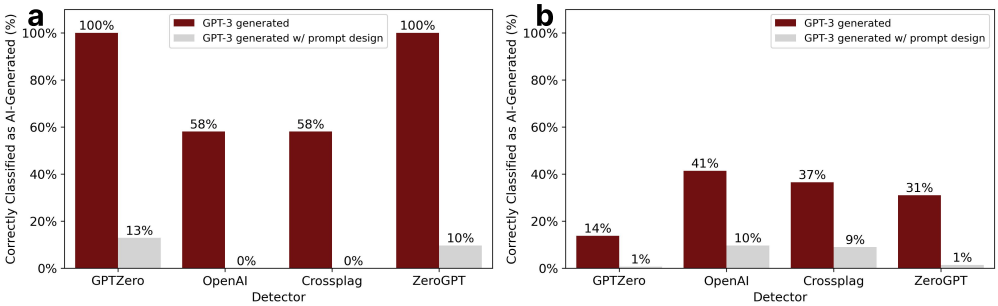

Figure 2: **Simple prompts effectively bypass GPT detectors.** (**a**) For GPT-3.5-turbo generated college admission essays, the performance of four widely-used GPT detectors declines markedly when a second-round self-edit prompt (*"Elevate the provided text by employing literary language"*) is applied, with detection rates dropping from 100% to 13%. (**b**) Analogously, in detecting GPT-3.5-turbo generated scientific abstracts, a second-round self-edit prompt (*"Elevate the provided text by employing advanced technical language"*) leads to a reduction in detection rates from 41% to 10%.

Building upon our insights from mitigating bias, we discovered that similar principles can be applied to make GPT-generated content bypass GPT detectors. As a proof of concept, we prompted GPT-3.5-turbo with the 2022-2023 US Common App college admission essay prompts, generating 31 counterfeit essays after filtering out invalid responses. While detectors were initially effective, a second-round self-edit prompt (*"Elevate the provided text by employing literary language"*) applied to GPT-3.5-turbo significantly reduced detection rates from 100% to 13% (**Fig. 2***a*).

In a parallel experiment, we prompted GPT-3.5-turbo to generate scientific abstracts using 145 Stanford CS224n final project report titles. Detectors were less effective in this context but still identified up to 41% of fake abstracts. However, applying a second-round self-edit prompt (*"Elevate the provided text by employing advanced technical language"*) lowered detection rates to 10% (**Fig. 2***b*). In both cases, the second-round self-edit substantially improved the text's perplexity, as measured on the text-davinci-002 model (a GPT-3.5 model trained without reinforcement learning).

These results demonstrate that simple prompts can easily bypass current GPT detectors, revealing the vulnerability of perplexity-based approaches. As the perplexity of GPT-generated text can be significantly improved using straightforward prompt design, it is crucial to develop more robust detection methods that are less susceptible to such manipulations.

DISCUSSION

This study reveals a notable bias in GPT detectors against non-native English writers, as evidenced by the high misclassification rate of non-native-authored TOEFL essays, in stark contrast to the near zero misclassification rate of college essays, which are presumably authored by native speakers. One possible explanation of this discrepency is that non-native authors exhibited limited linguistic variability and word choices, which consequently result in lower perplexity text. By employing a GPT-4 intervention to enhance the essays' word choice, we observed a substantial reduction in

the misclassification of these texts as AI-generated. This outcome, supported by the significant increase in average perplexity across four language models after the GPT-4 intervention, underscores the inherent limitations in perplexity-based AI content detectors. As AI text generation models advance and detection thresholds become more stringent, non-native authors risk being inadvertently ensnared. Paradoxically, to evade false detection as AI-generated content, these writers may need to rely on AI tools to refine their vocabulary and linguistic diversity. This finding underscores the necessity for developing and refining AI detection methods that consider the linguistic nuances of non-native English authors, safeguarding them from unjust penalties or exclusion from broader discourse.

Our investigation into the effectiveness of simple prompts in bypassing GPT detectors raises significant concerns about the reliability of current detection methods. As we demonstrated, a straightforward second-round self-edit prompt can drastically reduce detection rates for both college essays and scientific abstracts. This finding highlights the susceptibility of perplexity-based approaches to manipulation, indicating that more robust detection techniques are necessary to effectively identify AI-generated content. The ease with which GPT-generated text can be made to elude detection also underscores the importance of ongoing research into alternative, more sophisticated detection methods that are less vulnerable to such circumvention strategies.

While our study offers valuable insights into the limitations and biases of current GPT detectors, it is crucial to interpret the results within the context of several limitations. Firstly, although our datasets and analysis present novel perspectives as a pilot study, the sample sizes employed in this research are relatively small. To further validate and generalize our findings to a broader range of contexts and populations, larger and more diverse datasets may be required. Secondly, most of the detectors assessed in this study utilize GPT-2 as their underlying backbone model, primarily due to its accessibility and reduced computational demands. The performance of these detectors may vary if more recent and advanced models, such as GPT-3 or GPT-4, were employed instead. Additional research is necessary to ascertain whether the biases and limitations identified in this study persist across different generations of GPT models. Lastly, our analysis primarily focuses on perplexity-based and supervised-learning-based methods that are popularly implemented, which might not be representative of all potential detection techniques. For instance, DetectGPT (Mitchell et al., 2023), based on second-order log probability, has exhibited improved performance in specific tasks but is orders of magnitude more computationally demanding to execute, and thus not widely deployed at scale. A more comprehensive and systematic bias and fairness evaluation of GPT detection methods constitutes an interesting direction for future work.

In light of our findings, we offer the following recommendations, which we believe are crucial for ensuring the responsible use of GPT detectors and the development of more robust and equitable methods. First, we strongly caution against the use of GPT detectors in evaluative or educational settings, particularly when assessing the work of non-native English speakers. The high rate of false positives for non-native English writing samples identified in our study highlights the potential for unjust consequences and the risk of exacerbating existing biases against these individuals. Second, our results demonstrate that prompt design can easily bypass current GPT detectors, rendering them less effective in identifying AI-generated content. Consequently, future detection methods should move beyond solely relying on perplexity measures and consider more advanced techniques, such as second-order perplexity methods (Mitchell et al., 2023) and watermarking techniques (Kirchenbauer et al., 2023; Gu et al., 2022). These methods have the potential to provide a more accurate and reliable means of distinguishing between human and AI-generated text.

REFERENCES

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, 2021.

Evan Crothers, Nathalie Japkowicz, and Herna Viktor. Machine generated text: A comprehensive survey of threat models and detection methods. *arXiv preprint arXiv:2210.07321*, 2022.

Nature Editorial. Tools such as chatgpt threaten transparent science; here are our ground rules for their use. *Nature*, 613(7945):612–612, 2023.

Holly Else. Abstracts written by chatgpt fool scientists. *Nature*, Jan 2023. URL https://www.nature.com/articles/d41586-023-00056-7.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*, pp. 2022–12, 2022.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, 2019.

Chenxi Gu, Chengsong Huang, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. Watermarking pre-trained language models with backdooring. *arXiv preprint arXiv:2210.07543*, 2022.

Melissa Heikkil"a. How to spot ai-generated text. *MIT Technology Review*, Dec 2022. URL https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/.

Krystal Hu. Chatgpt sets record for fastest-growing user base - analyst note. *Reuters*, February 2023. URL https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

ICML. Clarification on large language model policy LLM. https://icml.cc/Conferences/2023/llm-policy, 2023.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*, 2020.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

Sarah Kreps, R McCain, and Miles Brundage. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117, 2022. doi: 10.1017/XPS.2020.37.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*, 2022.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

OpenAI. GPT-2: 1.5B release. https://openai.com/research/gpt-2-1-5b-release, 2019. Accessed: 2019-11-05.

OpenAI. ChatGPT. https://chat.openai.com/, 2022. Accessed: 2022-12-31.

Martine Paris. Chatgpt hits 100 million users, google invests in ai bot and catgpt goes viral. *Forbes*, February 2023. URL https://www.forbes.com/sites/martineparis/2023/02/03/chatgpt-hits-100-million-microsoft-unleashes-ai-bots-and-catgpt-goes-viral/.

Kalhan Rosenblatt. Chatgpt banned from new york city public schools' devices and networks. *NBC News*, Jan 2023. URL https://nbcnews.to/3iTE0t6. Accessed: 22.01.2023.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Christian Terwiesch. Would chat gpt3 get a wharton mba? a prediction based on its performance in the operations management course. *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania*, 2023.