Patterns

CellPress
OPEN ACCESS

Opinion

GPT detectors are biased against non-native English writers

Weixin Liang,[1,4] Mert Yuksekgonul,[1,4] Yining Mao,[2,4] Eric Wu,[2,4] and James Zou[1,2,3,*]
[1]Department of Computer Science, Stanford University, Stanford, CA, USA
[2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[4]These authors contributed equally
*Correspondence: jamesz@stanford.edu
https://doi.org/10.1016/j.patter.2023.100779

GPT detectors frequently misclassify non-native English writing as AI generated, raising concerns about fairness and robustness. Addressing the biases in these detectors is crucial to prevent the marginalization of non-native English speakers in evaluative and educational settings and to create a more equitable digital landscape.

Introduction

Generative language models based on GPT, such as ChatGPT, have gained significant attention in recent times. Within a mere 2 months of its launch, ChatGPT amassed over 100 million monthly active users, marking its place as one of the fastest-growing consumer internet applications in history.[1] Despite their immense potential for enhancing productivity and fostering creativity, these powerful models also pose risks, such as the proliferation of AI-generated content masquerading as human written, which may lead to the spread of fake content and exam cheating.

Educators, in particular, are increasingly concerned about determining when and where students have used AI and AI writing tools in their work. However, multiple studies have demonstrated the difficulty humans face in detecting AI-generated content with the naked eye,[2] thus creating an urgent and pressing demand for effective detection methods. While several GPT detectors have been developed and implemented to mitigate the risks associated with AI-generated content, their accuracy, reliability, and effectiveness remain uncertain due to limited evaluation.[3] This knowledge gap is especially worrisome given the potentially harmful consequences of mistakenly flagging an innocent student's work as AI generated.[4]

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 our recent preprint,[5,6] we exposed an alarming bias in GPT detectors against non-native English speakers: over half of the non-native English writing samples were misclassified as AI generated, while the accuracy for native samples remained near perfect. Our analysis further revealed a trend where more literary language was classified as more "human": enhancement of word choice in non-native English writing samples reduced misclassification, while simplifying native writing samples increased it, suggesting that GPT detectors are inadvertently penalizing individuals with limited linguistic proficiency. On the other hand, we found that GPT detectors be easily bypassed by better ChatGPT prompt design. This raises a pivotal question: if AI-generated content can easily evade detection while human text is frequently misclassified, how effective are these detectors truly?

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 is among the first to systematically examine the biases present in GPT detectors and advocates 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.

GPT detectors exhibit bias against non-native English authors

GPT detectors exhibit significant bias against non-native English authors, as demonstrated by their high misclassification of TOEFL essays written by non-native speakers. In our study, we evaluated the performance of seven widely used GPT detectors on 91 TOEFL (Test of English as a Foreign Language) essays from a Chinese forum and 88 US eighth-grade essays from the Hewlett Foundation's ASAP dataset. While the detectors accurately classified the US student essays, they incorrectly labeled more than half of the TOEFL essays as "AI-generated" (average false-positive rate: 61.3%). All detectors unanimously identified 19.8% of the human-written TOEFL essays as AI authored, and at least one detector flagged 97.8% of TOEFL essays as AI generated. Upon closer inspection, the unanimously identified TOEFL essays exhibited significantly lower text perplexity. Here text perplexity is a measure of how "surprised" or "confused" a generative language model is when trying to guess the next word in a sentence. If a generative language model can predict the next word easily, the text perplexity is low. On the other hand, if the next word is hard to predict, the text perplexity is high. Most GPT detectors use text perplexity to detect AI-generated text, which might inadvertently penalize non-native writers who use a more limited range of linguistic expressions.

Mitigating bias through linguistic diversity enhancement of non-native samples

Addressing limitations in linguistic variability in non-native English writing could

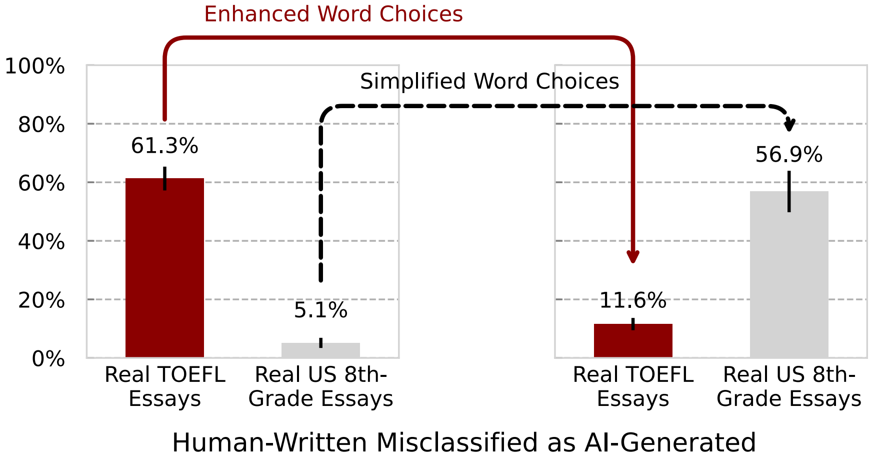

Figure 1. Bias in GPT detectors against non-native English writing samples
High misclassification of TOEFL essays written by non-native English authors as AI generated, with near-perfect accuracy for US eighth-grade essays. Improved word choice in TOEFL essays reduces misclassification (prompt: "Enhance the word choices to sound more like that of a native speaker"), while simplification of US eighth-grade essays increases misclassification (prompt: "Simplify word choices as if written by a non-native speaker"). Performance averaged across seven widely used GPT detectors. The error bars represent the standard deviation across the seven detectors.

help mitigate the GPT detectors' bias. We used ChatGPT to enhance the vocabulary of TOEFL essays, aiming to emulate native-speaker language use. This intervention significantly reduced misclassification, with the average false-positive rate dropping by 49.7% (from 61.3% to 11.6%). After this modification, the essays' text perplexity increased significantly, and only one TOEFL essay was unanimously identified as AI generated. In contrast, simplifying the vocabulary in US eighth-grade essays to mirror non-native writing led to a substantial increase in misclassification as AI-generated text (Figure 1).

Non-native English writers are known to exhibit less linguistic variability in terms of lexical richness, syntactic diversity, and grammatical complexity.[7] Analyzing academic research papers from ICLR 2023 (International Conference on Learning Representations), we found that papers by first authors from countries whose native language is not English showed lower text perplexity compared to their native English-speaking counterparts, indicating that their language use is more predictable by generative language models. This trend remained after accounting for review ratings. Therefore, practitioners should exercise caution when using low perplexity as an indicator of AI-generated text, as such an approach could unintentionally exacerbate systemic biases against non-native authors within the academic community.

Bypassing GPT detectors through linguistic diversity enhancement in prompts

On the other hand, we found that current GPT detectors are not as adept at catching AI plagiarism as one might assume. As a proof-of-concept, we asked ChatGPT to generate responses for the 2022–2023 US Common App college admission essay prompts. Initially, detectors were effective in spotting these AI-generated essays. However, upon prompting ChatGPT to self-edit its text with more literary language (prompt: "Elevate the provided text by employing literary language"), detection rates plummeted to near zero (Figure 2). A parallel experiment with scientific abstracts yielded similar results. In both cases, the text perplexity increased significantly after the self-edit. These findings underscore the vulnerabilities of current detection techniques, indicating that a simple manipulation in prompt design can easily bypass current GPT detectors.

Discussion

Many teachers consider GPT detection as a critical countermeasure to deter "a 21st-century form of cheating,"[4] but most GPT detectors are not transparent. Claims of GPT detectors' "99% accuracy" are often taken at face value by a broader audience, which is misleading at best, given the lack of access to a publicly available test dataset, information on model specifics, and details on training data. The commercial and closed-source nature of most GPT detectors introduces additional challenges and unnecessary obstacles to independently verify and validate their effectiveness. In this paper, we show that the hype about GPT detectors hides an under-discussed risk: GPT detectors are biased against non-native English writers. This is illustrated by the high rate of misclassification of TOEFL essays written by non-native English authors, which stands in sharp contrast to the nearly nonexistent misclassification rate of essays written by native English speakers.

The design of many GPT detectors inherently discriminates against non-native authors, particularly those exhibiting restricted linguistic diversity and word choice. The crux of the issue lies in the reliance of these detectors on specific statistical measures to identify AI-crafted writing, measures that also unintentionally distinguish non-native- and native-written samples. Text perplexity, a widely adopted statistical measure in numerous GPT detectors, typifies this issue.[8] Essentially, text perplexity gauges the degree of "surprise" a generative language model experiences when predicting the subsequent word in a sentence. If a generative language model can predict the next word easily, the perplexity is low. On the other hand, if the next word is hard to predict, the perplexity is high. Conceptually, this approach appears effective, considering generative language models such as ChatGPT work essentially like a sophisticated version of auto-complete, looking for the most probable word to write next, which often results in low text perplexity. Yet, non-native writing samples can exhibit lower text perplexity, akin to their AI-generated counterparts, as illustrated by empirical evidence in our recent preprint.[5] The predictability of non-native writing, stemming from a limited vocabulary and grammar range, can result in lower text perplexity. An interesting finding from our research was that, by introducing an intervention to diversify the word choice in non-native essays, we noticed a significant elevation in text perplexity, coupled with a substantial decrease in the misclassification of these texts as AI generated.

The implications of GPT detectors for non-native writers are serious, and we

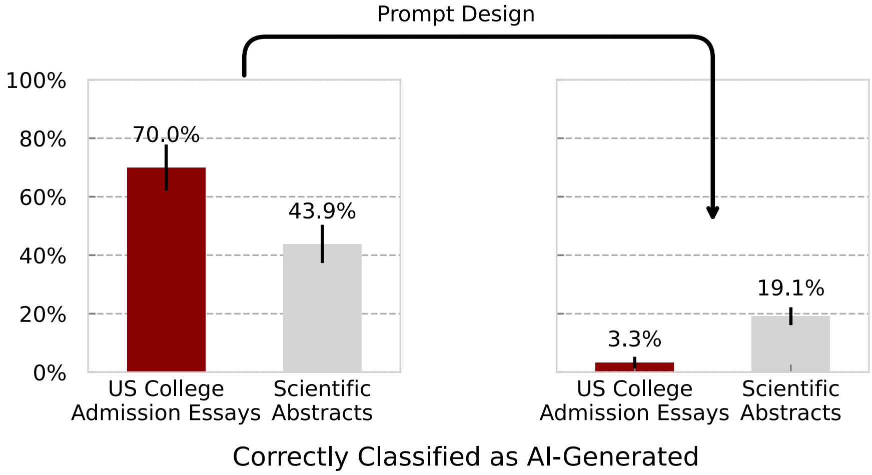

Figure 2. Simple prompts effectively bypass GPT detectors
Detection rates for ChatGPT-3.5-generated college essays and scientific abstracts drop significantly with a self-edit prompt (e.g., “Elevate the provided text by employing literary language”). Performance averaged across seven widely used GPT detectors. The error bars represent the standard deviation across the seven detectors.

need to think through them to avoid situations of discrimination. Within social media, GPT detectors could spuriously flag non-native authors’ content as AI plagiarism, paving the way for undue harassment of specific non-native communities. Internet search engines, such as Google, that implement mechanisms to devalue AI-generated content may inadvertently restrict the visibility of non-native communities, potentially silencing diverse perspectives. Academic conferences or journals prohibiting use of GPT may penalize researchers from non-English-speaking countries. In education, arguably the most significant market for GPT detectors, non-native students bear more risks of false accusations of cheating, which can be detrimental to a student’s academic career and psychological well-being. Even if the accusation is revoked later, the student’s reputation is already damaged. The use of these tools also ushers in an atmosphere of "presumption of guilt," where students are assumed to be dishonest until proven otherwise. Given the potential for mistrust and anxiety provoked by the deployment of GPT detectors, it raises questions about whether the negative impact on the learning environment outweighs the perceived benefits. If the purpose of these tools is to foster integrity in academic writing, it is crucial to enhance trust and ensure the maintenance of a supportive, inclusive educational climate.

Paradoxically, GPT detectors might compel non-native writers to use GPT more to evade detection. As GPT text-generation models advance and detection thresholds tighten, the risk of non-native authors being inadvertently caught in the GPT detection net increases. If non-native writing is more consistently caught as GPT, this may create an unintended consequence of ironically causing non-native writers to use GPT to refine their vocabulary and linguistic diversity to sound more native. Also, non-native speakers may increasingly use GPT legitimately as a way to improve their English and adopt certain grammatical structures common in GPT models. This could trigger an unintended cycle wherein non-native writers are forced to use GPT more extensively to enhance their vocabulary and diversify their linguistic usage to sound more “native.” Moreover, as non-native speakers increasingly rely on GPT to legitimately improve their English, they may begin to incorporate grammatical structures typical of GPT models. This phenomenon raises crucial questions about the ethical use of AI tools and the necessity for transparent guidelines that respect the rights of non-native authors while maintaining academic and professional integrity.

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. Our study’s identified high false-positive rate for non-native English writing underscores the potential for unwarranted consequences and the exacerbation of existing biases against these individuals. Even for native English speakers, linguistic variation across different socioeconomic backgrounds could potentially subject certain groups to a disproportionately higher risk of false accusations. Our second recommendation is for a more comprehensive evaluation of GPT detectors. To mitigate unjust outcomes stemming from biased detection, it is crucial to benchmark GPT detectors with diverse writing samples that reflect the heterogeneity of users. These evaluation strategies will catalyze the development of future detection algorithms that are more fairness-aware and inclusive. Third, the design and use of GPT detectors should not follow a one-size-fits-all approach. Rather, they should be designed by domain experts and used in collaboration with users. They should undergo rigorous evaluation in the intended domain and should communicate the relevant risks. A potential low-risk application of GPT detectors could be their use as educational aids rather than assessment tools. Proficient at recognizing clichéd expressions and repetitive patterns, GPT detectors can serve as self-check mechanisms for students. By highlighting overused phrases or structures, they may encourage writers to be more original and creative. As a result, these tools could potentially foster not only greater language proficiency but also the development of unique writing styles.

Lastly, we emphasize the need for inclusive conversations involving all stakeholders, including developers, students, educators, policymakers, ethicists, and those affected by GPT. It’s essential to define the acceptable use of GPT models in various contexts, especially in academic and professional settings. Consider, for instance, non-native speakers leveraging GPT as a linguistic aid to enhance their writing. Could it be considered as a legitimate use case where GPT augments, not supplants, human efforts, assisting in language construction without undermining the originality of ideas? These dialogues can inform the development of more enlightened and fair policies governing AI usage in writing, so as to maximize benefits and minimize harm. In summary, our joint efforts should strive to foster an atmosphere of trust, understanding, and inclusivity for all writers, regardless of their native language or linguistic capabilities.

ACKNOWLEDGMENTS

We thank B. He and S. Schwartz for discussions. J.Z. is supported by the National Science Foundation (CCF 1763191 and CAREER 1942926), the US National Institutes of Health (P30AG059307 and U01MH098953), and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Mollman, S. (2022). ChatGPT gained 1 million users in under a week. Here's why the AI chatbot is primed to disrupt search as we know it (Yahoo! Finance). https://www.yahoo.com/video/chatgpt-gained-1-million-followers-224523258.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAIYB6YTwTdZ_orPrsDbVfVouswfH7Hm_CgdzVnpIceLQJ8b3FFV4fK9rULMQ8MbFPEqMjVjyofEg3PZ6D_UEip6lNVp20rPOnxXzCz7gKw4orLDmpMAC-pUrdESpZ1tDMziIXneSBmK-UTn8Drgy6jgpjGOnNTvtHcwyeBnbMhBp.
2. Else, H. (2023). Abstracts written by ChatGPT fool scientists. Nature *613*, 423.
3. Heikkilä, M. (2022). How to spot AI-generated text. MIT Technol. Rev. https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/.
4. Fowler, G.A. (2023). We tested a new ChatGPT-detector for teachers. It flagged an innocent student (The Washington Post). https://www.washingtonpost.com/technology/2023/04/01/chatgpt-cheating-detection-turnitin/.
5. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.02819. https://arxiv.org/abs/2304.02819.
6. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). Code and Data for: GPT Detectors Are Biased Against Non-Native English Writers (Zenodo). https://doi.org/10.5281/zenodo.7893958.
7. Laufer, B., and Nation, P. (1995). Vocabulary size and use: Lexical richness in l2 written production. Appl. linguistics *16*, 307–322.
8. Bowman, E. (2023). A college student created an app that can tell whether ai wrote an essay. NPR. https://www.npr.org/2023/01/09/1147549845/gptzero-ai-chatgpt-edward-tian-plagiarism.

About the authors

Weixin Liang is in the second year of his doctorate studies in computer science at Stanford University, working under the supervision of Professor James Zou. Previously, he obtained a master's degree in electrical engineering from Stanford University and a bachelor's degree in computer science from Zhejiang University. His research is primarily focused on the areas of trustworthy AI, data-centric AI, and natural language processing.

Mert Yuksekgonul is a second-year PhD student in computer science at Stanford University, advised by James Zou and Carlos Guestrin. He focuses on enabling safer use and a greater understanding of deep learning, with interests in explaining model behavior, intervention, and multimodal understanding. Mert graduated from Bogazici University with dual bachelors' degrees in computer engineering and industrial engineering.

Yining Mao is a first-year master student at Stanford University, majoring in electrical engineering. She received a BE in computer science from Zhejiang University in 2022. Her research interests currently lie in machine learning and computer vision.

Eric Wu is a PhD candidate in electrical engineering at Stanford University, working with Professors James Zou in biomedical data science and Daniel E. Ho in the law school. Funded by the Stanford Bio-X SIGF Fellowship, Eric's research focuses on health and artificial intelligence, exploring AI regulation in medicine, machine learning for cancer diagnostics, and computational pathology. He has developed AI for cancer detection at DeepHealth and worked in product management at Google. Eric holds a master's degree in computational science from Harvard University and a bachelor's degree from Duke University.

James Zou, PhD, is an assistant professor of biomedical data science, computer science, and electrical engineering at Stanford University. His research focuses on developing reliable, human-compatible, and statistically rigorous machine learning algorithms, with a particular interest in human disease and health applications. He received his PhD from Harvard in 2014 and has held positions at Microsoft Research, Cambridge, and UC Berkeley. At Stanford, he is a two-time Chan-Zuckerberg investigator and faculty director of the Stanford Data4Health hub. His work is supported by the Sloan Fellowship, NSF CAREER Award, and various industry AI awards.