GPT DETECTORS ARE BIASED AGAINST NON-NATIVE ENGLISH WRITERS

Weixin Liang[*], **Mert Yuksekgonul**[*], **Yining Mao**[*], **Eric Wu**[*], **James Zou**[†]
Department of Computer Science
Stanford University
Stanford, CA, USA
jamesz@stanford.edu

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

[*]these authors contributed equally to this work
[†]Correspondence should be addressed to: jamesz@stanford.edu

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

We evaluated the performance of seven widely-used GPT detectors on a corpus of 91 human-authored TOEFL essays obtained from a Chinese educational forum and 88 US 8-th grade essays sourced from the Hewlett Foundation's Automated Student Assessment Prize (ASAP) dataset Kaggle (2012) (**Fig. 1***a*). The detectors demonstrated near-perfect accuracy for US 8-th grade essays. However, they misclassified over half of the TOEFL essays as "AI-generated" (average false positive rate: 61.22%). All seven detectors unanimously identified 18 of the 91 TOEFL essays (19.78%) as AI-authored, while 89 of the 91 TOEFL essays (97.80%) are flagged as AI-generated by at least one detector. For the TOEFL essays that were unanimously identified (**Fig. 1***b*), we observed that they had significantly lower perplexity compared to the others (P-value: 9.74E-05). This suggests that GPT detectors may penalize non-native writers with limited linguistic expressions.

MITIGATING BIAS THROUGH LINGUISTIC DIVERSITY ENHANCEMENT OF NON-NATIVE SAMPLES

To explore the hypothesis that the restricted linguistic variability and word choices characteristic of non-native English writers contribute to the observed bias, we employed ChatGPT to enrich the language in the TOEFL essays, aiming to emulate the vocabulary usage of native speakers (Prompt: *"Enhance the word choices to sound more like that of a native speaker."*) (**Fig. 1***c*). Remarkably, this intervention led to a substantial reduction in misclassification, with the average false positive rate decreasing by 49.45% (from 61.22% to 11.77%). Post-intervention, the TOEFL essays' perplexity significantly increased (P-value=9.36E-05), and only 1 out of 91 essays (1.10%) was unanimously detected as AI-written. In contrast, applying ChatGPT to adjust the word choices in US 8th-grade essays to mimic non-native speaker writing (Prompt: *"Simplify word choices as if written by a non-native speaker."*) led to a significant increase in the misclassification rate as AI-generated text, from an average of 5.19% across detectors to 56.65% (**Fig. 1***ac*). This word choice adjustment also resulted in significantly lower text perplexity (**Fig. 1***d*).

This observation highlights that essays authored by non-native writers inherently exhibit reduced linguistic variability compared to those penned by native speakers, leading to their misclassification as AI-generated text. Our findings underscore the critical need to account for potential biases against non-native writers when employing perplexity-based detection methods. Practitioners should 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. Non-native English writers have been shown to exhibit reduced linguistic variability in terms of lexical richness (Laufer & Nation, 1995), lexical diversity (Jarvis, 2002; Daller et al., 2003), syntactic complexity (Lu, 2011; Crossley & McNamara, 2014; Ortega, 2003), and grammatical complexity (Biber et al., 2011). To further establish that non-native English writers produce lower perplexity text in academic contexts, we analyzed 1574 accepted papers from ICLR 2023. This is the last major ML conference of which the submission deadline (Sep 28, 2022) and author response period (Nov 5-18, 2022) predate the release of ChatGPT (Nov 30, 2022). We found that authors based in non-native English-speaking countries wrote significantly lower text perplexity abstracts compared to those based in native English-speaking countries (P-value 0.035). After controlling for average review ratings, the

difference in perplexity between native and non-native authors remained significant (P-value 0.033). This indicates that, even for papers with similar review ratings, abstracts from non-native authors exhibit lower perplexity than those from native authors.

SIMPLE PROMPT CAN EASILY BYPASS CURRENT GPT DETECTORS

Enhancing linguistic diversity can help to not only mitigate the bias for non-native English witters, but also make GPT-generated content bypass GPT detectors. As a proof of concept, we prompted ChatGPT-3.5 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 ChatGPT-3.5 significantly reduced detection rates from 100% to 13% (**Fig. 2***a*). Although ChatGPT-3.5 generated essays initially exhibit notably low perplexity, applying the self-edit prompt leads to a significant increase in perplexity (**Fig. 2***b*) (P-value 1.94E-15). In a parallel experiment, we prompted ChatGPT-3.5 to generate scientific abstracts using 145 Stanford CS224n final project report titles (**Fig. 2***c*). Detectors were less effective in this context, partly because the generated abstracts have slightly higher perplexity than their essays counterpart (**Figs. 2***bd*), but still identified up to 68% of fake abstracts. However, applying a second-round self-edit prompt (*"Elevate the provided text by employing advanced technical language"*) lowered detection rates to up to 28%. Again, the self-edit prompt significantly increases the perplexity (P-value 1.06E-31). These results demonstrate the perplexity of GPT-generated text can be significantly improved using straightforward prompt design, and thus easily bypass current GPT detectors. revealing the vulnerability of perplexity-based approaches. A lot of Room of improvement, 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. Non-native English writers have been shown to exhibit reduced linguistic variability in terms of lexical richness (Laufer & Nation, 1995), lexical diversity (Jarvis, 2002; Daller et al., 2003), syntactic complexity (Lu, 2011; Crossley & McNamara, 2014; Ortega, 2003), and grammatical complexity (Biber et al., 2011). 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 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, along with recent studies on paraphrasing attacks (Krishna et al., 2023; Sadasivan et al., 2023), raises significant concerns about the reliability of current detection methods. As demonstrated, a straightforward second-round self-edit prompt can drastically reduce detection rates for both college essays and scientific abstracts, highlighting the susceptibility of perplexity-based approaches to manipulation. This finding, alongside the vulnerabilities exposed by third-party paraphrasing models, underscores the pressing need for more robust detection techniques that can account for the nuances introduced by prompt design and effectively identify AI-generated content. Ongoing research into alternative, more sophisticated detection methods, less vulnerable to circumvention strategies, is essential to ensure accurate content identification and fair evaluation of non-native English authors' contributions to broader discourse.

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

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *Tesol Quarterly*, 45(1):5–35, 2011.

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.

Scott A Crossley and Danielle S McNamara. Does writing development equal writing quality? a computational investigation of syntactic complexity in l2 learners. *Journal of Second Language Writing*, 26:66–79, 2014.

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.

Helmut Daller, Roeland Van Hout, and Jeanine Treffers-Daller. Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics*, 24(2):197–222, 2003.

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.

Scott Jarvis. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1):57–84, 2002.

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

Kaggle. The hewlett foundation: Automated essay scoring. https://www.kaggle.com/c/asap-aes, 2012. Accessed: 2023-03-15.

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.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*, 2023.

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.

Batia Laufer and Paul Nation. Vocabulary size and use: Lexical richness in l2 written production. *Applied linguistics*, 16(3):307–322, 1995.

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.

Xiaofei Lu. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL quarterly*, 45(1):36–62, 2011.

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.

Lourdes Ortega. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied linguistics*, 24(4):492–518, 2003.

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.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 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.

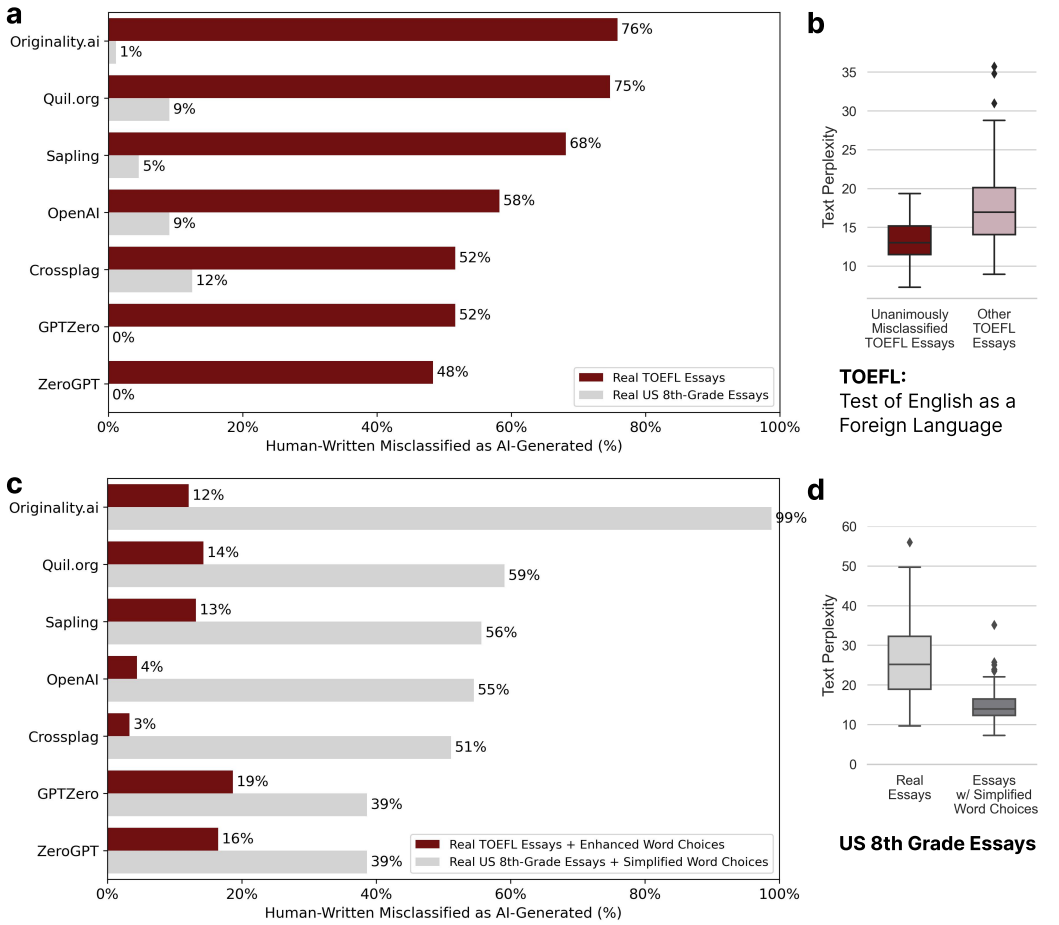

Figure 1: **Bias in GPT detectors against non-native English writing samples.** (**a**) Performance comparison of seven widely-used GPT detectors. More than half of the non-native-authored TOEFL (Test of English as a Foreign Language) essays are incorrectly classified as "AI-generated," while detectors exhibit near-perfect accuracy for US 8-th grade essays. (**b**) TOEFL essays unanimously misclassified as AI-generated show significantly lower perplexity compared to others, suggesting that GPT detectors might penalize authors with limited linguistic expressions. (**c**) Using ChatGPT to improve the word choices in TOEFL essays (Prompt: *"Enhance the word choices to sound more like that of a native speaker."*) significantly reduces misclassification as AI-generated text. Conversely, applying ChatGPT to simplify the word choices in US 8th-grade essays (Prompt: *"Simplify word choices as if written by a non-native speaker."*) significantly increases misclassification as AI-generated text. (**d**) The US 8th-grade essays with simplified word choices demonstrate significantly lower text perplexity.

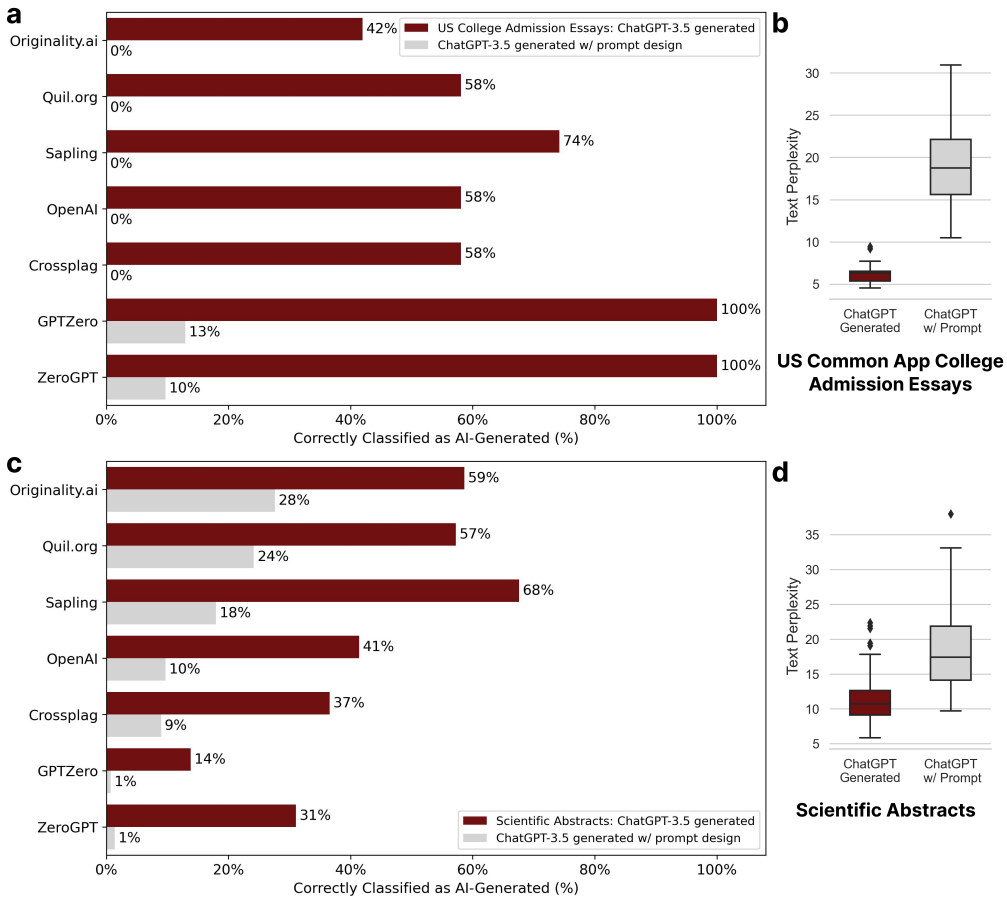

Figure 2: **Simple prompts effectively bypass GPT detectors.** (**a**) For ChatGPT-3.5 generated college admission essays, the performance of seven widely-used GPT detectors declines markedly when a second-round self-edit prompt (*"Elevate the provided text by employing literary language"*) is applied, with detection rates dropping from up to 100% to up to 13%. (**b**) ChatGPT-3.5 generated essays initially exhibit notably low perplexity; however, applying the self-edit prompt leads to a significant increase in perplexity. (**c**) Similarly, in detecting ChatGPT-3.5 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 up to 68% to up to 28%. (**d**) ChatGPT-3.5 generated abstracts have slightly higher perplexity than the generated essays but remain low. Again, the self-edit prompt significantly increases the perplexity.