Proof for the existence of bias and potential causes

Weixin Liang et al. (2023) from Stanford University made an experiment try to prove the existence of the bias in AI detectors towards NNS. In their experiment, they tested seven widely used GPT detectors on 91 TOEFL essays written by Chinese students and 88 U.S. eighth-grade essays. TOEFL is a standardized English assessment for NNS seeking admission to English-speaking universities (*About the TOEFL IBT Test*, n.d.). The detectors correctly classified American students' papers, but falsely labeled the TOEFL essays as AI-generated with an average false-positive rate of 61.3%. The false-positive rate reflects human-written content being inaccurately detected as AI-generated text. In addition, 97.8% of TOEFL essays were incorrectly flagged as AI-generated by at least one detector. These results indicate that current detectors are biased towards recognizing grammar and sentence structures typically used by NNS, which is likely caused by training detectors with biased datasets which are mainly consisted of native English writing (Liang et al., 2023).

Another experiment made by Weixin Liang et al. shows that this bias can also originate from the way these detectors were developed. As Figure 1 shows, two ways to modify the word choices were introduced in this study. After enhancing the vocabulary of TOEFL essays to resemble native-speaker language, false-positive rate significantly reduced from 61.3% to 11.6% (Liang et al., 2023). This improvement highlights how linguistic complexity affects the detection accuracy, as more varied vocabulary led to fewer misclassifications. Conversely, the misclassification rate as AI-generated text increased after simplifying the vocabulary of US eighth-grade essays to mimic non-native writing, demonstrating the bias toward simpler language structures. After analyzing the results, Liang and his team came up with a key factor that underpins this outcome is text perplexity, which is one of the most important features uses in AI detectors. Text perplexity measures how difficult it is for the model to predict the next word in a sequence. A lower perplexity score indicates highly predictable text, often lacking the variation and complexity typical of human writing (Liang et al., 2023). What's more, as text perplexity can provide clearer and more straightforward signal, which refers to complexity of structure and vocabulary use of a text in the case of AI detectors, it plays a significant role in classification results. Under this circumstance, once a detector easily predicts the upcoming words in a text, that text will be assigned a low perplexity score, and it is more likely to be flagged as AI-generated (Liang et al., 2023). Therefore, Liang and his group believe that the usage of text perplexity will make the misclassification more likely to happen on the non-native writers who use a limited range of linguistic expressions.

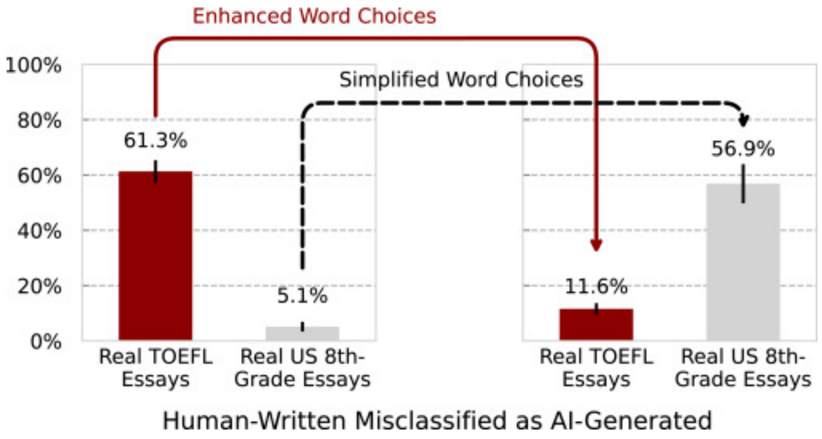

Figure 1: Result analysis after modifying the word choices (Liang et al., 2023)

The idea that the process of development for current AI detectors reinforced the bias was also proved by Yang Jiang and his team. In their study, Jiang et al. (2024a) used another commonly used feature E-rater® to assess the writing quality based on structure, grammar and vocabulary. As E-rater® serves as an overall assessment for the quality of article, it can also provides clear signal to model and affects model's decision significantly. Therefore, texts with lower E-rater® score are more likely to be misclassified as AI-generated, as these texts inlude simpler word choices or sentence structures that typical of AI writing. Jiang et al. (2024a) calculated E-rater® for 111,375 submissions (20.8% were submitted by NNS) from the GRE exam, which is a standardized test designed to assess verbal and quantitative reasoning for prospective graduate students (Learn about GRE, n.d.). As Figure 2 shows, the results revealed a significant disparity in writing scores between NS and NNS. While the majority of NNS' score is 3, nearly half of the NS scored a 4. Jiang et al. (2024a) thought this distribution highlighted the fact that though grammatically correct, NNS' texts often exhibited simpler sentence structures and vocabulary. Thus, due to the simplicity of NNS' creations, the texts created by them always have a lower average E-rater® score and are more likely to be flagged as AI work incorrectly.

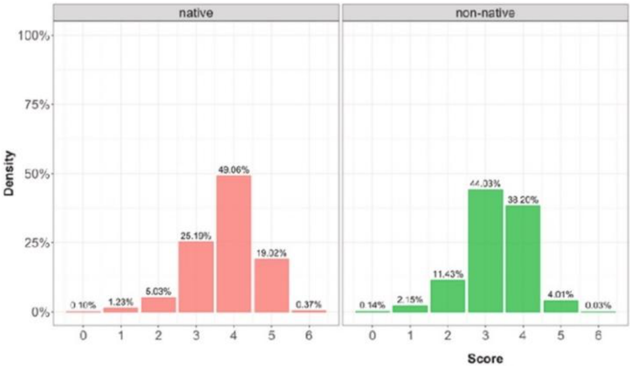

Figure 2: Distribution for the E-rater® score (Jiang et al., 2024)

Reference

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

Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024a). Detecting ChatGPT-Generated Essays in a Large-Scale Writing Assessment: Is There a Bias Against Non-Native English Speakers? *Computers and Education/Computers & Education*, 105070–105070. https://doi.org/10.1016/j.compedu.2024.105070