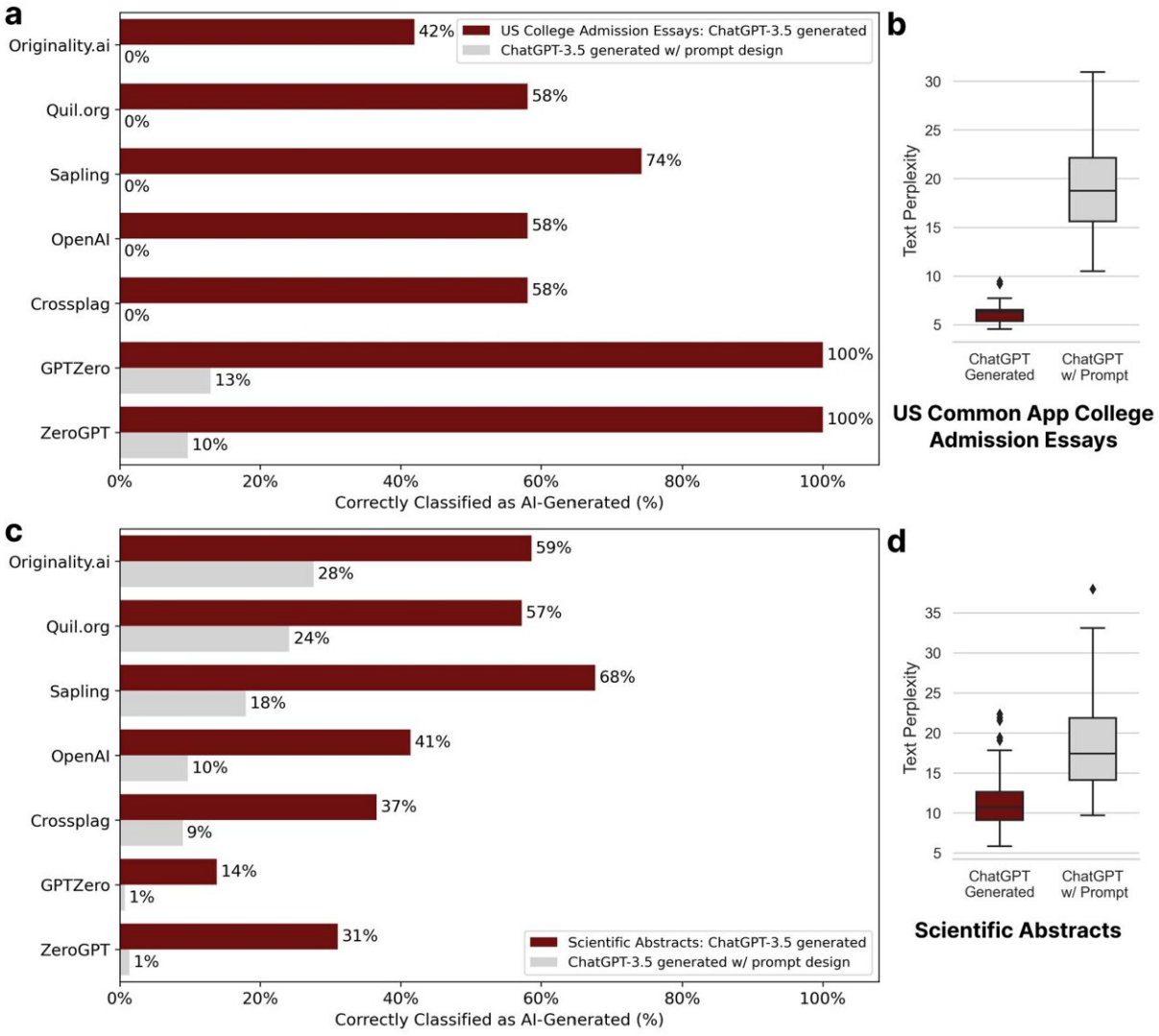
Why GPT detectors aren't a solution to the AI cheating problem

May 18 2023, by Andrew Myers



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. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2304.02819

In the wake of the high-profile launch of ChatGPT, no fewer than seven developers or companies have countered with AI detectors. That is, AI they say is able to tell when content was written by another AI. These new algorithms are pitched to educators, journalists, and others as tools to flag cheating, plagiarism, and mis- or disinformation.

It's all very meta, but according to a new paper from Stanford scholars, there's just one (very big) problem: The detectors are not particularly reliable. Worse yet, they are especially unreliable when the real author (a human) is not a native English speaker.

The numbers are grim. While the detectors were "near-perfect" in evaluating essays written by U.S.-born eighth-graders, they classified more than half of TOEFL essays (61.22%) written by non-native English students as AI-generated (TOEFL is an acronym for the Test of English as a Foreign Language).

It gets worse. According to the study, all seven AI detectors unanimously identified 18 of the 91 TOEFL student essays (19%) as AI-generated and a remarkable 89 of the 91 TOEFL essays (97%) were flagged by at

least one of the detectors.

"It comes down to how detectors detect AI," says James Zou, a professor of biomedical data science at Stanford University, a Stanford Institute for Human-Centered AI affiliate, and the senior author of the study. "They typically score based on a metric known as 'perplexity,' which correlates with the sophistication of the writing—something in which non-native speakers are naturally going to trail their U.S.-born counterparts."

Zou and co-authors point out that non-native speakers typically score lower on common perplexity measures such as lexical richness, lexical diversity, syntactic complexity, and grammatical complexity.

"These numbers pose serious questions about the objectivity of AI detectors and raise the potential that foreign-born students and workers might be unfairly accused of or, worse, penalized for cheating," Zou says, highlighting the team's ethical concerns.

Zou also notes that such detectors are easily subverted by what is known as "prompt engineering." That term of art in the AI field simply means asking generative AI to "rewrite" essays, for example, to include more sophisticated language, Zou says. He provides an example of just how easy bypassing the detectors is. A student wishing to use ChatGPT to cheat might simply plug in the AI-generated text with the prompt: "Elevate the provided text by employing literary language."

"Current detectors are clearly unreliable and easily gamed, which means we should be very cautious about using them as a solution to the AI cheating problem," Zou says.

The question then turns to what to do about it. Zou offers a few suggestions. In the immediate future, he says we need to avoid relying on

detectors in educational settings, especially where there are high numbers of non-native English speakers. Second, developers must move past using perplexity as their main metric to find more sophisticated techniques or, perhaps, applying watermarks in which the generative AI embeds subtle clues about its identity into the content it creates. Finally, they need to make their models less vulnerable to circumvention.

"The detectors are just too unreliable at this time, and the stakes are too high for the students, to put our faith in these technologies without rigorous evaluation and significant refinements," Zou says.

The findings are published on the *arXiv* preprint server.

More information: Weixin Liang et al, GPT detectors are biased against non-native English writers, *arXiv* (2023). DOI: 10.48550/arxiv.2304.02819

Provided by Stanford University