

PROOF POINTS: It's easy to fool ChatGPT detectors

H hechingerreport.org/proof-points-its-easy-to-fool-chatgpt-detectors

September 4, 2023



Get important education news and analysis delivered straight to your inbox

Choose from our newsletters

A high school English teacher recently explained to me how she's coping with the latest challenge to education in America: ChatGPT. She runs every student essay through five different generative AI detectors. She thought the extra effort would catch the cheaters in her classroom.

A clever series of experiments by computer scientists and engineers at Stanford University indicate that her labors to vet each essay five ways might be in vain. The researchers demonstrated how seven commonly used GPT detectors are so primitive that they are both easily fooled by machine generated essays and improperly flagging innocent students. Layering several detectors on top of each other does little to solve the problem of false negatives and positives.

"If AI-generated content can easily evade detection while human text is frequently misclassified, how effective are these detectors truly?" the Stanford scientists wrote in a [July 2023 paper](#), published under the banner, "opinion," in the peer-reviewed data science journal Patterns. "Claims of GPT detectors' '99% accuracy' are often taken at face value by a broader audience, which is misleading at best."

The scientists began by generating 31 counterfeit college admissions essays using [ChatGPT 3.5](#), the free version that any student can use. GPT detectors were pretty good at flagging them. Two of the seven detectors they tested caught all 31 counterfeits.

But all seven GPT detectors could be easily tricked with a simple tweak. The scientists asked ChatGPT to rewrite the same fake essays with this prompt: "Elevate the provided text by employing literary language."

Detection rates plummeted to near zero (3 percent, on average).

I wondered what constitutes literary language in the ChatGPT universe. Instead of college essays, I asked ChatGPT to write a paragraph about the perils of plagiarism. In ChatGPT's first version, it wrote: "Plagiarism presents a grave threat not only to academic integrity but also to the development of critical thinking and originality among students." In the second, "elevated" version, plagiarism is "a lurking specter" that "casts a formidable shadow over the realm of academia, threatening not only the sanctity of scholastic honesty but also the very essence of intellectual maturation." If I were a teacher, the preposterous magniloquence would have been a red flag. But when I ran both drafts through several AI detectors, the boring first one was flagged by all of them. The flamboyant second draft was flagged by none. Compare the two drafts side by side for yourself.

Standard vs. "Elevated" ChatGPT

First prompt: Write one paragraph about the perils of plagiarism to submit to my high school teacher

Second prompt: Elevate the text by employing literary language

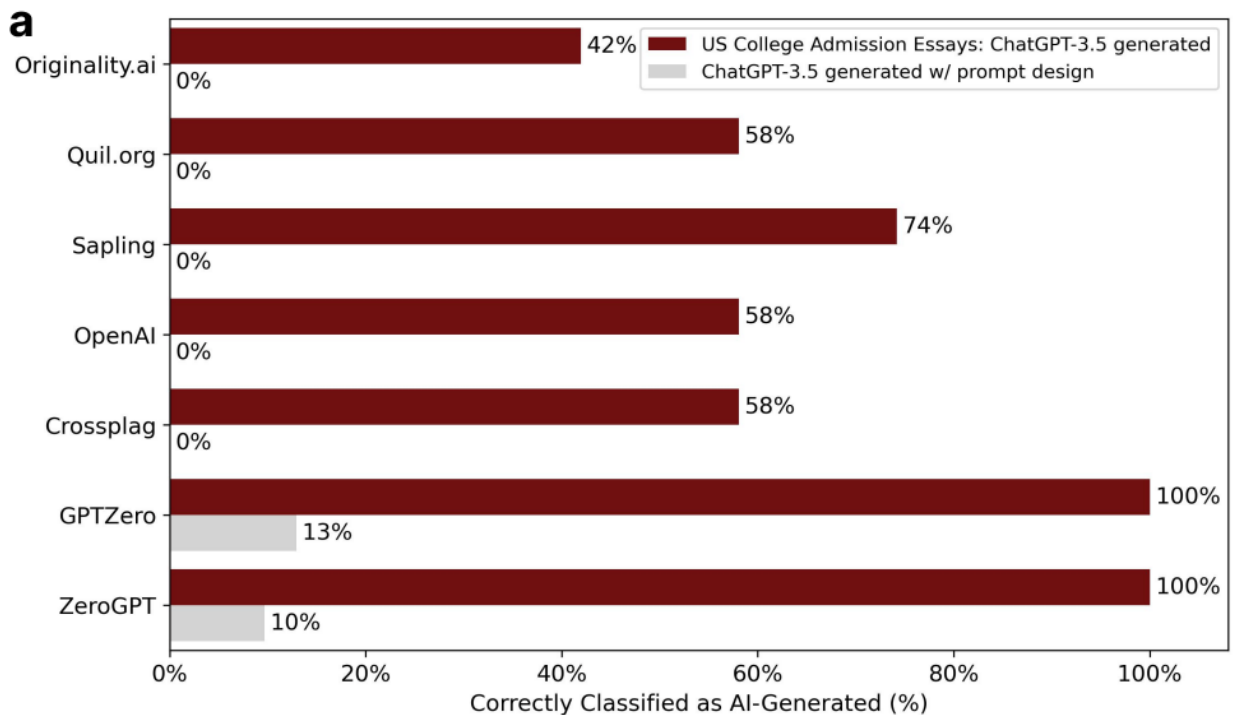
Standard ChatGPT

Plagiarism presents a grave threat not only to academic integrity but also to the development of critical thinking and originality among students. By copying and submitting others' work as their own, students undermine the fundamental principles of learning, hindering their own growth and understanding. Moreover, plagiarism diminishes the value of education, eroding the trust between educators and students while distorting the assessment of true knowledge and skills. Confronting the perils of plagiarism is vital to fostering an environment where creativity, authenticity, and ethical conduct flourish, ultimately preparing students for a future where integrity and innovation are paramount.

Elevated ChatGPT

Plagiarism, a lurking specter, casts a formidable shadow over the realm of academia, threatening not only the sanctity of scholastic honesty but also the very essence of intellectual maturation. In surreptitiously adopting another's words or ideas as one's own, students sow the seeds of complacency, stunting the growth of their cognitive faculties and stifling the sprouting of original thought. This betrayal of learning's sacred tenets not only corrodes the foundation of education but also corrodes the trust that binds mentors and mentees, leaving truth and proficiency in disarray. By unflinchingly addressing the precipice of plagiarism, we summon an atmosphere wherein ingenuity, authenticity, and ethical comportment can thrive, fashioning minds resilient and refined, poised for a future where integrity and inventive ideation stand as unassailable beacons.

Simple prompts bypass ChatGPT detectors. Red bars are AI detection before making the language loftier; gray bars are after.

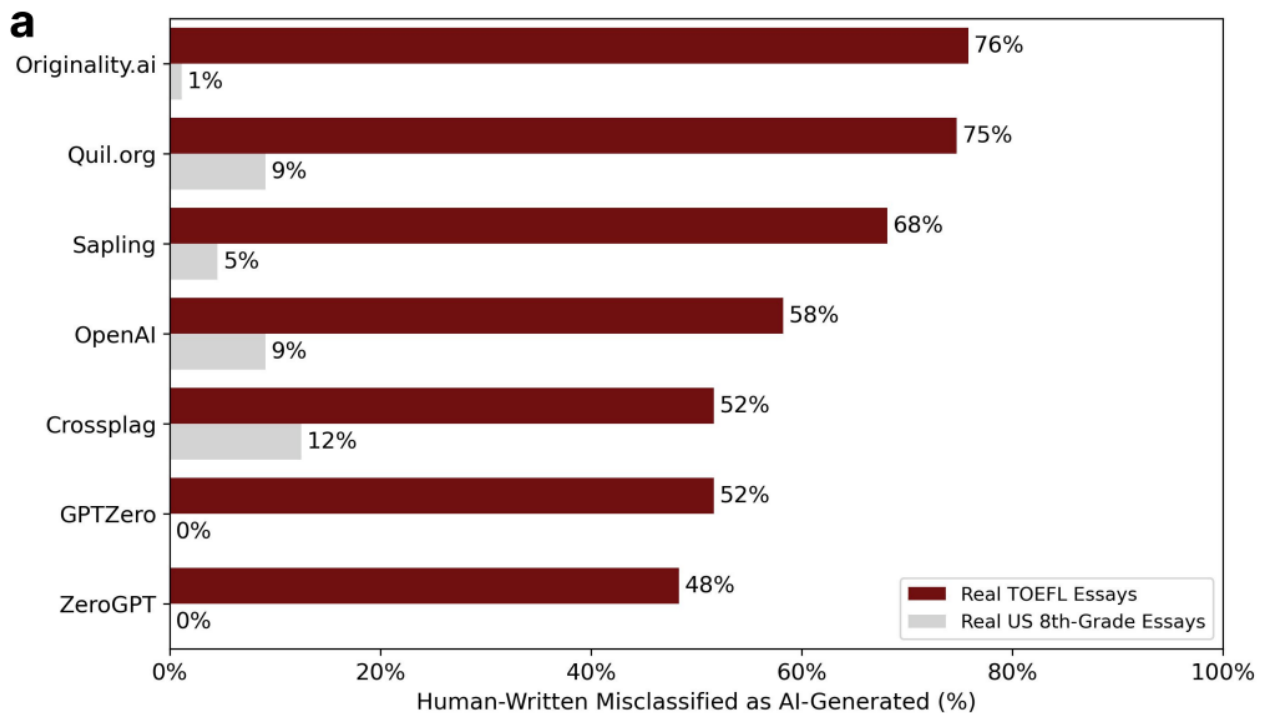


For ChatGPT 3.5 generated college admission essays, the performance of seven widely used ChatGPT detectors declines markedly when a second round self-edit prompt (“Elevate the provided text by employing literary language”) is applied. Source: Liang, W., et al. “[GPT detectors are biased against non-native English writers](#)” (2023)

Meanwhile, these same GPT detectors incorrectly flagged essays written by real humans as AI generated more than half the time when the students were not native English speakers. The researchers collected a batch of 91 practice English TOEFL essays that Chinese students had voluntarily uploaded to a [test-prep forum](#) before ChatGPT was invented. (TOEFL is the acronym for the Test of English as a Foreign Language, which is taken by international students who are applying to U.S. universities.) After running the 91 essays through all seven ChatGPT detectors, 89 essays were identified by one or more detectors as possibly AI-generated. All seven detectors unanimously marked one out of five essays as AI authored. By contrast, the researchers found that GPT detectors accurately categorized a separate batch of 88 eighth grade essays, submitted by real American students.

My former colleague Tara García Mathewson brought this research to my attention in her first story for The Markup, which highlighted how international college students are facing [unjust accusations of cheating](#) and need to prove their innocence. The Stanford scientists are warning not only about unfair bias but also about the futility of using the current generation of AI detectors.

Bias in ChatGPT detectors. Leading detectors incorrectly flag a majority of essays written by international students, but accurately classify writing of American eighth graders.



More than half of the TOEFL (Test of English as a Foreign Language) essays written by non-native English speakers were incorrectly classified as “AI-generated,” while detectors exhibit near-perfect accuracy for U.S. eighth graders’ essays. Source: Liang, W., et al. “GPT detectors are biased against non-native English writers” (2023)

The reason that the AI detectors are failing in both cases – with a bot’s fancy language and with foreign students’ real writing – is the same. And it has to do with how the AI detectors work. Detectors are a machine learning model that analyzes vocabulary choices, syntax and grammar. A widely adopted measure inside numerous GPT detectors is something called “text perplexity,” a calculation of how predictable or banal the writing is. It gauges the degree of “surprise” in how words are strung together in an essay. If the model can predict the next word in a sentence easily, the perplexity is low. If the next word is hard to predict, the perplexity is high.

Low perplexity is a symptom of an AI generated text, while high perplexity is a sign of human writing. My intentional use of the word “banal” above, for example, is a lexical choice that might “surprise” the detector and put this column squarely in the non-AI generated bucket.

Because text perplexity is a key measure inside the GPT detectors, it becomes easy to game with loftier language. Non-native speakers get flagged because they are likely to exhibit less linguistic variability and syntactic complexity.

The seven detectors were created by [originality.ai](#), Quill.org, [Sapling](#), [Crossplag](#), [GPTZero](#), [ZeroGPT](#) and OpenAI (the creator of ChatGPT). During the summer of 2023, Quill and [OpenAI](#) both decommissioned their free AI checkers because of inaccuracies. Open AI's website says it's planning to launch a [new one](#).

“We have taken down AI Writing Check,” Quill.org wrote on its website, “because the new versions of Generative AI tools are too sophisticated for detection by AI.”

The site blamed newer generative AI tools that have come out since ChatGPT launched last year. For example, [Undetectable AI](#) promises to turn any AI-generated essay into one that can evade detectors ... for a fee.

Quill recommends a clever workaround: check students' Google doc version history, which Google captures and saves every few minutes. A normal document history should show every typo and sentence change as a student is writing. But someone who had an essay written for them – either by a robot or a ghostwriter – will simply copy and paste the entire essay at once into a blank screen. “No human writes that way,” the Quill site says. A more detailed explanation of how to check a document's version history is [here](#).

Checking revision histories might be more effective, but this level of detective work is ridiculously time consuming for a high school English teacher who is grading dozens of essays. AI was supposed to save us time, but right now, it's adding to the workload of time-pressed teachers!

This story about [ChatGPT detectors](#) was written by Jill Barshay and produced by [The Hechinger Report](#), a nonprofit, independent news organization focused on inequality and innovation in education. Sign up for [Proof Points](#) and other [Hechinger newsletters](#).

The Hechinger Report provides in-depth, fact-based, unbiased reporting on education that is free to all readers. But that doesn't mean it's free to produce. Our work keeps educators and the public informed about pressing issues at schools and on campuses throughout the country. We tell the whole story, even when the details are inconvenient. Help us keep doing that.

[Join us today.](#)

