CENTER FOR DEMOCRACY & TECHNOLOGY

A SERIES ON GENERATIVE AI IN EDUCATION

CLASSROOM TRUST

The Shortcomings of Generative AI Detection: How Schools Should Approach Declining Teacher Trust In Students

Generative AI – systems that use machine learning to produce new content (e.g., text or images) in response to user prompts – has infiltrated the education system and fundamentally shifted the relationships between teachers and their students.

Across the country, educators have expressed high levels of anxiety about students using generative AI tools, like ChatGPT, to cheat on assignments, exams, and essays in addition to fears of students losing critical thinking skills. One professor even described it as having "infected [the education system] like a deathwatch beetle, hollowing out sound structures from the inside until the imminent collapse." In response to these fears, school districts, like New York City and Los Angeles, quickly imposed bans on its use by both educators and students. Schools have turned to tools like generative AI detectors to attempt to restore educator control and trust; however, detection efforts have fallen short in both their implementation and efficacy.

CDT Research Affirms Declining Trust…

One significant finding from our polling research of teachers, parents, and students is that teacher perception of widespread generative AI use for cheating appears to be largely unfounded. Forty percent of teachers who say that their students have used generative AI for school think their students have used it to write and submit a paper. But only **19 percent** of students who report having used generative AI say they have used it to write and submit a paper – a finding that is supported by other survey research.

Even despite the reality that a large majority of students are not using generative AI for nefarious academic purposes, teachers have still become more mistrustful of students' work – perhaps due to the widespread, fear-stoking coverage of cheating instances. **Sixty two percent** of teachers agreed with the statement that "[g]enerative AI has made me more distrustful of whether my students' work is actually theirs." And this mistrust is bleeding into certain groups of students being disciplined at disproportionate rates for using or being accused of using generative AI – Title I and licensed special education teachers report higher rates of disciplinary actions for generative AI use among their students.

These high levels of mistrust among teachers and subsequent disciplinary action have led to frustration among students and parents about erroneous accusations of cheating, which can cause an even further rift between teachers and students. This erosion of trust is potentially damaging to school communities where strong relationships between educators and their students are imperative in providing a safe, quality learning environment.

…And Insufficient Detection Tools And Training

Tools designed to detect when generative AI was used to produce content are the only technological solutions currently available to help teachers attempt to combat generative AI-based cheating; however, they fall short of solving existing trust issues. To begin, school policies on content detection tool use is spotty – only **17 percent** of teachers say that their school provides a content detection tool as part of its larger technology platform, and **26 percent** say their school recommends their use, but leaves it up to the educator to choose one and implement it. Without strong guidance on the use and implementation of content detection tools, teachers appear uneasy about utilizing them as a defense mechanism for cheating. Only **38 percent** of teachers report using a generative AI content detection tool regularly, and just **18 percent** of teachers strongly agree that these tools "are an accurate and effective way to determine whether a student is using AI-generated content." Teachers' lack of confidence is well-founded as, at least at this point, these tools are not consistently effective at differentiating between AI-generated and human-written text.

Beyond using tools for detection, teacher confidence in their own effectiveness at detecting generative AI created writing is low – **22 percent** say they are very effective and **43 percent** say they are somewhat effective. This is particularly concerning given that most teachers have not received guidance on how to detect cheating. Only **23 percent** of teachers who have received training on their schools' policies and procedures regarding generative AI have gotten guidance on how to detect student use of ChatGPT (or another generative AI tool) when submitting school assignments.

How Should Schools Approach Declining Teacher Trust?

Given our research and what we know about generative AI content detection tools, they are not the answer, at least for now. These tools suffer from accuracy issues, and may disproportionately flag non-native speakers. Instead, schools need to:

- **Offer teacher training on how to assess student work in light of generative AI.** To help teachers feel like they have more control over academic integrity in the classroom, schools must properly equip them to deal with the new reality of generative AI. This means providing them with training on the limitations of detectors and how to respond if they reasonably suspect that a student is cheating.

- **Craft and implement clear policies about which uses are allowed and prohibited.** Our polling from this past summer shows that schools are failing to provide guidance on what is defined as "improper use" of generative AI, with **37 percent** of teachers reporting that their school has no policy or they are not sure if there is a policy in place on generative AI. It is imperative for both teachers and students to know this, so that everyone is on the same page about responsible generative AI use.

- **Encourage teachers to modify assignments to minimize the effectiveness of generative AI.** Understanding what generative AI systems are *not* good at can help teachers design assignments where using generative AI will not be helpful to students. For instance, generative AI systems are often ineffective at providing accurate sources for their claims. Requiring students to provide citations for any claims they make will likely require students to go far beyond a generated response.