POLICY PAPER

A Score Is Not Proof

Protecting Virginia Students from Unreliable AI Detection

Submitted to the Virginia General Assembly
for Consideration in the 2027 Regular Session

MANGO (Modern Advocates for New Global Opinions)
Richmond, Virginia
September 30, 2026

I. Executive Summary

Virginia should prohibit school divisions from disciplining or academically penalizing a student based solely on the output of AI detection software. These tools cannot reliably determine whether a student used artificial intelligence, and their errors fall hardest on English learners, Black students, and students with disabilities.

AI detectors produce a probability score, not proof. The research shows they misclassify human writing at meaningful rates, are easily fooled by students who actually cheat, and disproportionately flag students whose writing is simpler or more formulaic. Even detection vendors caution against using scores as the sole basis for action. OpenAI withdrew its own detector for low accuracy, and universities including Vanderbilt have disabled detection features entirely.

Virginia is well positioned to act. Under SB 394 and HB 1186 (2026), the Department of Education must issue AI guidance and every school board must adopt an AI policy. MANGO asks the General Assembly to require that those policies include the following protections:

1. **No penalty based on a detector alone.** A detection score may prompt a conversation, but no grade reduction, zero, integrity charge, or discipline may rest on it without independent evidence.
2. **Notice and a chance to respond.** A student accused of unauthorized AI use is told the basis for the concern and given a meaningful opportunity to explain, including through drafts, version history, or discussing the work.
3. **Transparency.** Divisions disclose which detection tools they use and how scores are used.
4. **Protection of student work.** Vendors may not keep or use student writing to train products without consent.
5. **Reporting.** Divisions report annual, de-identified counts of AI-related academic integrity cases, so bias can be measured.

These measures protect honest students without tying teachers' hands. Teachers keep every tool they need to address cheating except one: treating an unreliable number as a verdict.

II. The Problem

Generative AI has put teachers in a hard position, and many have turned to detection software for help. In a national survey, the Center for Democracy & Technology found that student discipline related to suspected AI use rose from 48 percent to 64 percent of teachers reporting it between the 2022–23 and 2023–24 school years, alongside growing use of detection tools (K-12 Dive, reporting CDT).

The result is a growing number of students accused on the basis of a number. A typical case looks like this: a student submits an essay, the teacher runs it through a detector, the detector reports a high percentage of "AI-generated" text, and the student receives a zero or an integrity referral. The student, who wrote the essay, has no way to prove a negative.

A. Who Gets Flagged

- **English learners.** Stanford researchers tested seven widely used detectors on essays written by non-native English speakers for the TOEFL exam. The detectors misclassified more than 61 percent of them as AI-generated, while classifying essays by U.S. eighth graders almost perfectly. Nearly 98 percent of the non-native essays were flagged by at least one detector (Liang et al., Patterns, 2023).
- **Black students.** A 2024 nationally representative Common Sense Media survey found that 20 percent of Black teens reported a teacher incorrectly flagging their work as AI-generated, compared with 7 percent of white teens and 10 percent of Latino teens. Of the teens whose work was wrongly flagged, about 79 percent said it had been run through detection software (Education Week).
- **Students with disabilities.** Researchers and advocates have raised concerns that detectors may disproportionately flag writing by autistic students and others whose writing is formulaic or repetitive. Students with dyslexia who rely on approved grammar tools face similar risk (study on autistic writing and AI detection).

B. Why It Matters

An academic integrity finding can follow a student into honors placement, recommendation letters, and college applications. False accusations also erode trust between students and teachers and discourage the very students who most need support from engaging in class. When the tool is wrong, the student pays the price, and the student has almost no way to challenge it.

III. How AI Detectors Work, and Why They Fail

AI detectors do not find hidden markers in a student's essay. Most estimate how "predictable" the text is. Writing that uses common words and simple sentence structures looks more machine-like to the software. That design explains both of the tools' central weaknesses.

A. They Penalize Simple, Clear Writing

Because detectors associate predictable language with AI, students with a smaller vocabulary or a more structured style are flagged more often. In the Stanford study, the researchers showed that enriching the vocabulary of non-native essays reduced false flags, while simplifying native essays increased them (Liang et al.). The detector was measuring language proficiency, not honesty.

B. They Are Easy to Fool

Students who actually cheat can evade detection by paraphrasing AI output or running it through "humanizer" tools, some offered by detection companies themselves. The result is a tool that catches honest students who write simply while missing dishonest students who know how to avoid it.

C. Even Their Makers Urge Caution

Source	What It Found or Said
OpenAI (2023)	Withdrew its own AI text classifier, which correctly identified only 26 percent of

Source	What It Found or Said
	AI-written text while falsely flagging 9 percent of human writing.
Turnitin	Reports a document-level false positive rate under 1 percent, suppresses low scores because of false positive risk, and advises that results should not be the sole basis for adverse action.
Vanderbilt University (2023)	Disabled Turnitin's AI detector. Applying a 1 percent false positive rate to its roughly 75,000 annual submissions meant about 750 papers could be wrongly labeled each year.
Federal Trade Commission (2025)	Ordered Workado, a detection company advertising 98 percent accuracy, to stop unsupported claims after independent testing found about 53 percent accuracy on general content (FTC).

A 1 percent error rate sounds small until it is applied at scale. A Virginia division where students submit tens of thousands of essays a year could, even by a vendor's own estimate, wrongly flag hundreds of students annually.

IV. Current Law and Policy

A. Virginia Has Started, but Not Finished

Va. Code § 22.1-20.2:1 addresses artificial intelligence systems used to support teaching, learning, assessment, or school operations, and calls for guidance on student data privacy in the context of those systems. In 2026, SB 394 and HB 1186 directed the Department of Education to develop guidance for the safe, ethical, and equitable use of AI in public schools and required local school boards to adopt and enforce AI policies consistent with that guidance (Williams Mullen summary). The Governor's Guidelines for AI Integration Throughout Education, issued under Executive Order 30 (2024), state that AI should never replace human judgment.

None of these sources sets a rule on AI detectors. Divisions are now writing their AI policies. Richmond Public Schools, for example, has posted a draft AI policy that addresses academic integrity and disciplinary consequences and prohibits discriminatory profiling (RPS draft policy). Without a statewide floor, each division will decide for itself whether a detector score can justify a zero.

B. Other States Have Warned Against Detectors

Several state education agencies have already advised caution. West Virginia's Department of Education recommends that teachers not use AI content detectors because of reliability concerns, and North Carolina's guidance notes that detection tools often create false positives that penalize non-native speakers and creative writing styles (CDT review of state guidance). Hawaii and Washington have also cautioned educators against relying on these tools. Virginia can go further by turning that caution into a clear rule.

C. Students Have a Right to Be Heard

Under *Goss v. Lopez*, 419 U.S. 565 (1975), public school students facing suspension are entitled to notice of the charges and an opportunity to respond. Academic integrity penalties based on detection software rarely trigger those formal protections, yet the consequences can be lasting. The principle behind *Goss* (that a student must be told what they are accused of and allowed to answer) should apply when the accuser is an algorithm.

D. Civil Rights Exposure

Federal law prohibits discrimination based on race and national origin in federally funded programs (Title VI of the Civil Rights Act of 1964) and on disability (Section 504 and the Americans with Disabilities Act). A tool that flags English learners and Black students at markedly higher rates exposes divisions to civil rights complaints. Clear statewide rules protect divisions as well as students.

V. Answering the Case for Detectors

Argument	Response
Teachers need a way to catch AI cheating	This proposal does not ban detectors. A score may still prompt a conversation. It cannot be the only evidence.
It is unfair to honest students if cheaters get away with it	Detectors are easily evaded by students who cheat, so they often fail honest students twice: falsely flagging them while missing actual misconduct.
Newer detectors are more accurate	Some vendors report improved results, and accuracy may continue to improve. A rule requiring independent evidence remains appropriate as long as any false positive rate exists, and no vendor claims zero.
Teachers do not have time for extra steps	Reviewing drafts, version history, or briefly discussing the work with the student takes minutes and produces far stronger evidence than a percentage.
This interferes with local control	Local boards are already required to adopt AI policies. This sets a minimum protection while leaving everything else to divisions.

VI. Recommendations

1. **Prohibit adverse action based solely on detection software.** No grade penalty, academic integrity finding, or discipline may rest solely or primarily on the output of an AI detection tool.
2. **Require notice and an opportunity to respond.** Before any penalty for unauthorized AI use, the student (and parent, for a minor) is told the basis for the concern and given a meaningful opportunity to respond.
3. **Recognize better evidence.** Guidance should direct teachers toward evidence of the writing process, such as drafts, document version history, in-class writing samples, or a conversation about the work.

4. **Require disclosure.** Divisions publish which detection tools they use and how results may be used.
5. **Protect student work.** Contracts must bar vendors from retaining or using student submissions to train or improve products without consent.
6. **Collect data on disparities.** Divisions report annual, de-identified counts of AI-related academic integrity cases by outcome and student group, and the Department publishes a statewide summary.
7. **Train educators.** The Department's AI guidance should include training on detector limitations and fair procedures.

Appendix A: Draft Legislation

The following is a working draft offered for a patron's consideration. Official text would be prepared by the Division of Legislative Services, and the section number below is a placeholder.

COMMONWEALTH OF VIRGINIA
2027 SESSION

A BILL to amend the Code of Virginia by adding in Chapter 14 of Title 22.1 a section numbered 22.1-[____], relating to use of artificial intelligence detection software in public schools; student protections.

Be it enacted by the General Assembly of Virginia:

1. That the Code of Virginia is amended by adding in Chapter 14 of Title 22.1 a section numbered 22.1-[____] as follows:

§ 22.1-[____]. Artificial intelligence detection software; limitations; student protections.

A. As used in this section, "artificial intelligence detection software" means any software, tool, or service that analyzes student work and generates a score, probability, classification, or other output indicating whether the work was produced in whole or in part by an artificial intelligence system.

B. No school board, school, or employee thereof shall impose any grade reduction, failing grade, academic integrity finding, disciplinary action, or other adverse consequence on a student based solely or primarily on the output of artificial intelligence detection software.

C. Before imposing any adverse consequence on a student for the unauthorized use of an artificial intelligence system, the school shall (i) inform the student and, if the student is a minor, his parent of the basis for the concern, including any output of artificial intelligence detection software relied upon, and (ii) provide the student a meaningful opportunity to respond, which may include presenting drafts, document version history, or other evidence of the student's work process, or discussing the work with the teacher.

D. Each school board shall include the requirements of this section in the policy on artificial intelligence systems required by law and shall make publicly available (i) the name of each artificial intelligence detection software product used in the division and (ii) the circumstances under which its output may be considered.

E. Any contract between a school board and a provider of artificial intelligence detection software shall prohibit the provider from retaining, using, or disclosing student work submitted for analysis for any purpose other than providing the contracted service, including the training or improvement of any product, unless the student or, if the student is a minor, his parent consents in writing.

F. Each school board shall annually report to the Department of Education, in de-identified and aggregate form, the number of cases in which a student was alleged to have engaged in unauthorized use of an artificial intelligence system, the outcomes of such cases, and the number of cases in which artificial intelligence detection software was used, disaggregated by race, ethnicity, English learner status, and disability status, subject to suppression necessary to protect student privacy. The Department shall publish an annual statewide summary.

G. The Department of Education shall include in its guidance on the use of artificial intelligence systems in public schools (i) information on the limitations and error rates of artificial intelligence detection software, (ii) recommended procedures for addressing suspected unauthorized use of artificial intelligence consistent with this section, and (iii) professional development resources for educators.

2. That the provisions of this act shall become effective on July 1, 2027.

Sources

- Liang, Yuksekgonul, Mao, Wu, and Zou, GPT Detectors Are Biased Against Non-Native English Writers, Patterns 4(7) (2023)
- Education Week, Black Students Are More Likely to Be Falsely Accused of Using AI to Cheat (September 2024), reporting Common Sense Media, The Dawn of the AI Era
- Federal Trade Commission, FTC Order Requires Workado to Back Up Artificial Intelligence Detection Claims (April 2025)
- Center for Democracy & Technology, Looking Back at AI Guidance Across State Education Agencies
- K-12 Dive, As Teacher Use of AI Detection Grows, Discipline Guidance a Mixed Bag, reporting CDT research
- Va. Code § 22.1-20.2:1, Artificial intelligence system use and safety in instructional settings
- Williams Mullen, Virginia to Establish Guidance for Artificial Intelligence in Schools (SB 394 and HB 1186, 2026)
- Richmond Public Schools, Draft Policy: Artificial Intelligence (BoardDocs)
- Virginia Secretary of Education, Guidelines for AI Integration Throughout Education
- Ballotpedia, AI Guidance Issued by State Departments of Education
- UCLA HumTech, The Imperfection of AI Detection Tools (OpenAI classifier results)
- University of San Diego Legal Research Center, The Problems with AI Detectors (Turnitin false positive and false negative statements)
- The Misclassification of Autistic Writing as AI-Generated (preprint)
- *Goss v. Lopez*, 419 U.S. 565 (1975)