NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2025/05/22/accused-by-perplexity-ai-detection-as-an-unvalidated-automated-decision-system-in-graduate-education/.

Von Perplexity beschuldigt: KI-Erkennung als unvalidiertes automatisiertes Entscheidungssystem in der Graduiertenausbildung

Heather Grizzle · May 22, 2025

KI-Detektoren messen, wie ungewöhnlich ein Text ist, nicht, wer ihn geschrieben hat. Das trifft Zweitsprachenschreiber und Studierende, die zugelassene unterstützende Technologien nutzen, am härtesten.

Inhalt

1. Was der Detektor misst
2. Die Fehlerquoten, und wer sie trägt
3. Welche Gruppe dies trifft, und warum es die falsche ist
4. Der Beweisstandard ist niedriger als bei allem Vergleichbaren
5. Was eine Doktorandin tatsächlich tun sollte
6. Was eine Institution festlegen sollte
7. Fazit
8. Quellen

Der Rat, den alle geben, ist beim Mechanismus falsch

Der übliche Ratschlag an Doktoranden zum Thema künstliche Intelligenz lautet ungefähr so: Nutzen Sie sie als Werkzeug und nicht als Ersatz, denken Sie selbst, und verstehen Sie, dass Lehrkräfte aufmerksam sind und Arbeiten, denen es an Tiefe mangelt, letztlich auffallen. Die ersten beiden Aussagen sind richtig. Die dritte ist falsch, und ihre Falschheit ist von Bedeutung, denn sie beschreibt einen Prozess menschlichen Urteilsvermögens, der weitgehend ersetzt worden ist.

Die meisten Institutionen erkennen den Einsatz künstlicher Intelligenz nicht durch das Urteilsvermögen erfahrener Leser. Sie erkennen ihn, oder glauben es zu tun, durch automatisierte Klassifikatoren, die in Einreichungsplattformen eingebettet sind und einem Dokument einen Wahrscheinlichkeitswert zuweisen, der dann in ein Verfahren zur akademischen Integrität einfließt. Die Frage, der sich eine Doktorandin gegenübersieht, lautet daher nicht, ob eine Professorin oberflächliches Denken bemerken wird. Sie lautet, ob ein statistischer Klassifikator ihre Arbeit markiert und was mit ihr geschieht, wenn das der Fall ist.

Novara Consulting Group's position is that AI detection in higher education is an automated decision system producing consequential decisions about individuals, that it is deployed with less validation than institutions require of almost any other instrument affecting student standing, and that its error distribution falls hardest on multilingual and disabled writers. The governance problem is institutional. The consequences are individual.

Was der Detektor misst

Ein KI-Schreiberkennungstool identifiziert keine Urheberschaft. Es schätzt, wie statistisch ungewöhnlich ein Text ist, hauptsächlich anhand von Perplexität, die misst, wie vorhersehbar jedes Wort angesichts des vorangehenden Textes ist, und Burstiness, die die Variation in Satzstruktur und -länge über ein Dokument hinweg misst. Text, der hochgradig vorhersehbar und strukturell einheitlich ist, wird als eher maschinell erzeugt eingestuft.

That is the entire mechanism, and stating it plainly makes the problem visible. The detector is not measuring whether a person wrote something. It is measuring whether the writing is unusual, and treating unusualness as evidence of humanity. Writing that is careful, plain, structurally consistent, and lexically conservative is scored as machine-like, because that is what the model's proxy for machine-likeness is.

Dies erzeugt eine spezifische und vorhersehbare Gruppe von Fehlanschuldigungen, und diese ist nicht zufällig.

Die Fehlerquoten, und wer sie trägt

Eine im Juli 2023 in *Patterns* veröffentlichte Studie von Liang und Kollegen aus Stanford ergab, dass Detektoren etwa einundsechzig Prozent der von nicht-muttersprach-

lichen Englischsprechern verfassten TOEFL-Aufsätze als KI-generiert markierten, gegenüber nahezu null Prozent Falsch-Positiv-Rate bei vergleichbaren Texten von englischen Muttersprachlern, und zwar über alle sieben getesteten Detektoren hinweg. Dies waren Aufsätze von Menschen, die eine standardisierte Prüfung abgelegt hatten, um ihre Englischkenntnisse nachzuweisen. Die Autoren benannten den Mechanismus direkt: Die Detektoren orientierten sich an der Textperplexität, und das Schreiben in einer Zweitsprache erzeugt eine geringere Perplexität.

Unabhängige Tests haben wiederholt die von den Anbietern behaupteten Genauigkeitswerte nicht bestätigt. Weber-Wulff und Kollegen kamen in einer im Dezember 2023 im *International Journal for Educational Integrity* veröffentlichten Untersuchung, die eine Reihe von Erkennungswerkzeugen über mehrere Dokumenttypen hinweg bewertete, zu dem Schluss, dass diese weder genau noch zuverlässig seien. Anbieter wiederum weigern sich in der Regel, eine Falsch-Positiv-Rate für ihre Produkte zu veröffentlichen, was bedeutet, dass eine Institution, die ein solches Werkzeug einsetzt, die Fehlerquote des in Disziplinarverfahren verwendeten Instruments nicht angeben kann.

Institutions have begun to act on this. Vanderbilt University disabled Turnitin's AI detection feature in August 2023, publishing the arithmetic that made the decision straightforward: a one percent false positive rate applied to seventy-five thousand submitted papers produces roughly seven hundred and fifty wrongful accusations a year. Other institutions have followed, and the direction of published university guidance since has been toward treating a detector score as a prompt for inquiry rather than as a finding.

Eine weitere Markttatsache gehört in jede institutionelle Bewertung. Mehrere Anbieter, die Erkennungswerkzeuge vermarkten, vermarkten auch sogenannte Humanizer-Produkte, die darauf ausgelegt sind, die Erkennung zu umgehen – einschließlich ihrer eigenen. In jedem anderen Kontext der Qualitätssicherung wäre ein Anbieter, der sowohl den Test als auch die Mittel zu seinem Bestehen verkauft, von der Lieferung beider ausgeschlossen.

Welche Gruppe dies trifft, und warum es die falsche ist

In jedem anderen Kontext der Qualitätssicherung wäre ein Anbieter, der sowohl den Test als auch die Mittel zu seinem Bestehen verkauft, von der Lieferung beider ausgeschlossen.

Die dokumentierte Diskrepanz betrifft nichtmuttersprachliche englische Schreibende, und die Erklärung ist mechanisch, nicht mysteriös. Perplexitätsbasierte Erkennung bestraft einfacheren Wortschatz und regelmäßigere grammatikalische Konstruktionen. Ein in einer Zweitsprache schreibender Mensch erzeugt genau dieses Profil, nicht weil der Text maschinell erzeugt ist, sondern weil es schwieriger ist, in einer Zweitsprache ungewöhnliche Formulierungen zu konstruieren, als korrekte Formulierungen zu konstruieren.

Derselbe Mechanismus betrifft eine Gruppe, der bisher kaum Aufmerksamkeit geschenkt wurde, und genau diese zu betrachten ist der Grund, warum es dieses Unternehmen gibt.

A Deaf writer whose first language is American Sign Language is writing English as a second language, and the syntactic conservatism that follows is precisely what the detector scores as machine-like. A student with dyslexia using text prediction, grammar assistance, or structured writing support, all of which are ordinary and often formally approved accommodations, produces text that has been regularized by the assistive tool. A student with a motor impairment composing by speech-to-text produces text shaped by dictation software's own language model. A student using translation assistance produces output carrying the statistical signature of the translation system.

In jedem Fall führt die Anpassung zur Markierung. Die Studentin, die die von ihrer Institution genehmigte Unterstützung nutzt, wird mit höherer Wahrscheinlichkeit beschuldigt als die Studentin, die sie nicht benötigte, und die Anschuldigung kommt mit einer Zahl versehen und ohne Erklärung, was diese Zahl eigentlich misst. Sie wird dann aufgefordert zu beweisen, dass sie selbst geschrieben hat – eine Forderung nach einem Negativbeweis, gerichtet an die Person, die am wenigsten in der Lage ist, ihr nachzukommen.

An institution that has approved an assistive technology accommodation and simultaneously runs a detector penalizing its output has built a contradiction into its own disciplinary process. It is unlikely to have noticed, because the disability services office and the academic integrity office do not typically review each other's instruments. Postsecondary institutions carry obligations under the Americans with Disabilities Act and Section 504 of the Rehabilitation Act to provide academic adjustments and to avoid practices that screen out students with disabilities. A detection practice that systematically flags the products of approved accommodations sits uncomfortably against both.

Der Beweisstandard ist niedriger als bei allem Vergleichbaren

Consider what an institution would require before acting on any other automated instrument affecting a student's standing.

Ein Aufnahmetest würde veröffentlichte Validitätsnachweise, erwiesene Zuverlässigkeit, dokumentierte Leistung über demografische Gruppen hinweg und einen vertretbaren Grenzwert erfordern. Ein Überwachungstool, das eine Meldung wegen Integritätsverstoßes auslöst, würde mindestens voraussetzen, dass die Aufzeichnung überprüfbar ist. Eine Behindertenfeststellung würde Dokumentation und eine benannte entscheidungsbefugte Person erfordern. Eine Institution, die eine folgenreiche Entscheidung auf Grundlage eines unerklärten numerischen Wertes trifft, der von einem Instrument erzeugt wird, dessen Fehlerquote der Anbieter nicht offenlegt, würde gewöhnlich als willkürlich handelnd verstanden werden.

KI-Erkennungswerte werden anders behandelt, und der Grund scheint eher im Genre als im Prinzip zu liegen. Sie erscheinen innerhalb eines Plagiatstools, dem Institutionen bereits vertrauen, sie sehen aus wie der danebenstehende Ähnlichkeitsindex, und sie wurden über Beschaffung statt über akademische Governance eingeführt. Der Ähnlichkeitsindex verweist zumindest auf ein Quelldokument, das eine Studierende einsehen und anfechten kann. Der KI-Wert verweist auf nichts. Es gibt keine Quelle, keine Textstelle, keinen Vergleich und keinen Mechanismus, mit dem die Studierende die Grundlage der Anschuldigung prüfen könnte.

Better institutional practice treats a detector output as triage rather than evidence: a prompt to open a conversation in which the student explains her argument, walks

through her drafting, and discusses her sources. That is the correct posture, and it is worth noting what it concedes. If the detector's output is only a reason to look more closely, and the looking is what actually determines the outcome, then the detector is a mechanism for deciding which students get investigated. Given the error distribution above, it is a mechanism for deciding that multilingual and disabled students get investigated more often.

Was eine Doktorandin tatsächlich tun sollte

Der nützliche Rat lautet nicht, härter daran zu arbeiten, menschlich zu klingen, was sowohl herabwürdigend als auch unerreichbar ist. Er lautet, Nachweise zu führen – aus demselben Grund, aus dem jede Institution, die mit einer unerklärten automatisierten Entscheidung konfrontiert ist, Nachweise führen sollte: Die Anschuldigung ist ohne Prozessnachweise nicht widerlegbar, und Prozessnachweise müssen erzeugt werden, bevor sie benötigt werden.

Draft in an environment that retains version history, and do not disable it. A document with a two-month revision trail is dispositive in a way no argument about writing style will ever be. Keep research notes, annotated sources, and outlines, and keep them dated. Where an assistive or language support tool is used, disclose it in advance under the course's stated policy rather than after a flag, because prior disclosure converts a defense into a record. Where an accommodation involves writing support, ensure the disability services office has documented it, and understand that this documentation is now also integrity documentation. And be able to explain the argument aloud, not because a professor's perception is the safeguard, but because the structured conversation is where these matters are actually resolved.

Bei all dem geht es nicht um Täuschung. Es ist das, was jede Person tun sollte, wenn ein System, das nicht überprüft werden kann, möglicherweise eine Behauptung über sie aufstellt, die sie widerlegen soll.

Was eine Institution festlegen sollte

Vier Fragen, die eine Stelle für akademische Integrität beantworten können sollte, bevor ein Detektorergebnis eine Studierende erreicht.

1. What is the instrument's false positive rate, on what population, established by whom. If the vendor will not state it, that is the finding, and it should be recorded as one rather than treated as an absence of information.
2. Wie sieht die aufgeschlüsselte Leistung über die von der Institution betreuten Gruppen hinweg aus, einschließlich mehrsprachiger Studierender, internationaler Studierender und Studierender mit genehmigter assistierender Technologie? Eine Institution, die diese Gruppen betreut und nicht antworten kann, hat das Werkzeug nicht für ihren eigenen Kontext bewertet.
3. Welches Beweisgewicht wird dem Ergebnis zugewiesen, schriftlich festgehalten, und stützt sich irgendein belastender Befund allein darauf? Wenn ein Wert allein ein Verfahren auslösen kann, hat die Institution ein nicht validiertes Instrument als Auslöser für Disziplinarmaßnahmen übernommen.
4. Wer überprüft die Wechselwirkung zwischen genehmigten Anpassungen und Detektorergebnissen? Wenn niemand dies tut, sollte die Institution davon ausgehen, dass die Wechselwirkung nachteilig ist, denn der Mechanismus garantiert dies.

Fazit

Der Rat, dass Doktoranden selbst denken, ihre eigenen Ideen verteidigen und Fachwissen statt Abkürzungen entwickeln sollten, ist richtig und wäre in jeder Epoche richtig. Er geht jedoch nicht auf das Risiko ein, dem Studierende tatsächlich ausgesetzt sind – nämlich nicht, dass sie als oberflächlich befunden werden, sondern dass sie von einem Klassifikator markiert werden, der statistische Ungewöhnlichkeit misst, und aufgefordert werden, einen Negativbeweis zu erbringen.

Dieses Risiko ist ungleich verteilt, und seine Verteilung ist das Gegenteil dessen, was Integritätspolitik eigentlich bewirken soll. Die Studierenden, die am ehesten fälschlich beschuldigt werden, sind jene, die in einer Zweitsprache schreiben, jene, die in ihrer Erstsprache gebärden, und jene, die die von ihren Institutionen formell genehmigten Anpassungen nutzen. Sie werden häufiger untersucht, nicht weil sie eher betrü-

gen, sondern weil ein Stellvertreterwert für maschinelles Schreiben sich als Stellvertreterwert für Schreiben unter Einschränkung erweist.

Institutionen haben dieses Ergebnis nicht bewusst gewählt. Sie haben es sich durch eine Beschaffungsentscheidung ohne Validierungsnachweise eingehandelt, und sie werden es weiterhin erzeugen, bis jemand die Fragen stellt, die vor der Einführung hätten gestellt werden müssen.

Quellen

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. "GPT Detectors Are Biased Against Non-Native English Writers." *Patterns* 4, no. 7 (July 2023).

Weber-Wulff, Debora, et al. "Testing of Detection Tools for AI-Generated Text." *International Journal for Educational Integrity* 19 (December 2023).

Vanderbilt University Center for Teaching, statement on disabling Turnitin's AI detection feature, August 2023.

Americans with Disabilities Act von 1990 und Section 504 des Rehabilitation Act von 1973, angewandt auf postsekundäre Bildungseinrichtungen und akademische Anpassungsmaßnahmen.

> **TIPS**
>
> Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Zitierweise

Grizzle, H. M. (2025, May 22). Von Perplexity beschuldigt: KI-Erkennung als unvalidiertes automatisiertes Entscheidungssystem in der Graduiertenausbildung. Novara Consulting Group. https://www.novaracg.com/de/2025/05/22/accused-by-perplexity-ai-detection-as-an-unvalidated-automated-decision-system-in-graduate-education/