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/.

Acusado por Perplexity: la detección de IA como sistema de decisión automatizada sin validar en la educación de posgrado

Heather Grizzle · May 22, 2025

Los detectores de IA miden cuán inusual es un texto, no quién lo escribió. Eso perjudica más a los escritores que hablan un segundo idioma y a los estudiantes que utilizan tecnología de asistencia aprobada.

Contenidos

1. Lo que mide el detector
2. Las tasas de error, y quién las soporta
3. La población a la que esto afecta, y por qué es la equivocada
4. El estándar probatorio es más bajo que el de cualquier cosa comparable
5. Lo que un estudiante de posgrado debería realmente hacer
6. Lo que una institución debería establecer
7. Conclusión
8. Referencias

El consejo que todos dan está equivocado respecto al mecanismo

La orientación habitual que se da a los estudiantes de posgrado sobre la inteligencia artificial es, más o menos, la siguiente: úsala como herramienta y no como sustituto, piensa por ti mismo, y entiende que el profesorado es perceptivo y con el tiempo notará un trabajo que carece de profundidad. Las dos primeras afirmaciones son acertadas. La tercera es falsa, y su falsedad importa, porque describe un proceso de juicio humano que en gran medida ha sido reemplazado.

La mayoría de las instituciones no detectan el uso de inteligencia artificial mediante el criterio de lectores experimentados. Lo detectan, o creen detectarlo, mediante clasificadores automatizados integrados en las plataformas de entrega de trabajos, que asignan una puntuación de probabilidad a un documento y la remiten a un proceso de integridad académica. La pregunta que enfrenta una estudiante de posgrado no es, por tanto, si un profesor notará un pensamiento superficial. Es si un clasificador estadístico marcará su escrito, y qué le ocurrirá si eso sucede.

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.

Lo que mide el detector

Un detector de escritura por IA no identifica la autoría. Estima cuán estadísticamente inusual es un texto, principalmente mediante la perplejidad, que mide cuán predecible es cada palabra dado el texto que la precede, y la variabilidad (burstiness), que mide la variación en la estructura y longitud de las oraciones a lo largo de un documento. Un texto altamente predecible y estructuralmente uniforme recibe una puntuación que indica mayor probabilidad de haber sido generado por una máquina.

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.

Esto produce una población específica y predecible de acusaciones falsas, y no es una población aleatoria.

Las tasas de error, y quién las soporta

Una investigación publicada en *Patterns* en julio de 2023 por Liang y colegas de Stanford encontró que los detectores marcaron aproximadamente el sesenta y uno por ciento de los ensayos de TOEFL escritos por hablantes no nativos de inglés como ge-

nerados por IA, frente a tasas de falsos positivos cercanas a cero para escritos comparables de hablantes nativos de inglés, en los siete detectores evaluados. Se trataba de ensayos escritos por seres humanos que habían presentado un examen estandarizado para demostrar competencia en inglés. Los autores identificaron el mecanismo directamente: los detectores se basaban en la perplejidad del texto, y escribir en una segunda lengua produce una perplejidad más baja.

Las pruebas independientes han fracasado reiteradamente en respaldar las afirmaciones de precisión de los proveedores. Weber-Wulff y colegas, publicando en el *International Journal for Educational Integrity* en diciembre de 2023, evaluaron una variedad de herramientas de detección en múltiples tipos de documentos y concluyeron que no eran ni precisas ni fiables. Los proveedores, por su parte, generalmente se niegan a publicar una tasa de falsos positivos para sus productos, lo que significa que una institución que adopte uno no puede indicar la tasa de error de un instrumento que está utilizando en procedimientos disciplinarios.

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.

Un hecho adicional del mercado merece formar parte de cualquier evaluación institucional. Varios proveedores que comercializan herramientas de detección también comercializan productos «humanizadores» diseñados para burlar la detección, incluida la de sus propias herramientas. En cualquier otro contexto de garantía de calidad, un proveedor que vende tanto la prueba como el medio para superarla quedaría descalificado para suministrar cualquiera de las dos.

La población a la que esto afecta, y por qué es la equivocada

La disparidad documentada afecta a los escritores no nativos de inglés, y la explicación es

En cualquier otro contexto de garantía de calidad, un proveedor que vende tanto la

mecánica más que misteriosa. La detección basada en perplejidad penaliza el vocabulario más simple y la construcción gramatical más regular. Un escritor que trabaja en una segunda lengua produce exactamente ese perfil, no porque el texto haya sido generado por una máquina, sino porque construir un fraseo inusual en una segunda lengua es más difícil que construir un fraseo correcto.

> prueba como el medio para superarla quedaría descalificado para suministrar cualquiera de las dos.

El mismo mecanismo afecta a una población que casi no ha recibido atención, y es precisamente aquella que esta firma existe para considerar.

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.

En cada caso, la adaptación es lo que provoca la alerta. La estudiante que utiliza el apoyo aprobado por su institución tiene más probabilidades de ser acusada que la estudiante que no lo necesitó, y la acusación llega acompañada de un número, sin explicación alguna sobre lo que ese número mide. Después se le exige demostrar que escribió su propio trabajo, lo cual es exigir una prueba negativa, precisamente a la persona menos capacitada para satisfacerla.

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.

El estándar probatorio es más bajo que el de cualquier cosa comparable

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

Un examen de admisión necesitaría evidencia publicada de validez, fiabilidad establecida, rendimiento documentado en distintos grupos demográficos y un punto de corte defendible. Una herramienta de supervisión de exámenes que active una remisión por integridad requeriría, como mínimo, que la grabación pudiera revisarse. Una determinación de discapacidad requeriría documentación y un responsable de decisión identificado. Se entendería, por lo general, que una institución que toma una decisión con consecuencias basándose en una puntuación numérica no explicada, producida por un instrumento cuya tasa de error el proveedor se niega a revelar, ha actuado de manera arbitraria.

Las puntuaciones de detección de IA se tratan de manera diferente, y la razón parece ser de género antes que de principio. Llegan integradas en una herramienta antiplagio en la que las instituciones ya confían, se parecen al índice de similitud junto al que aparecen, y fueron adoptadas mediante un proceso de contratación en lugar de mediante gobernanza académica. El índice de similitud al menos señala un documento fuente que el estudiante puede inspeccionar y refutar. La puntuación de IA no señala nada. No hay fuente, ni pasaje, ni comparación, ni mecanismo alguno mediante el cual el estudiante pueda examinar el fundamento de la acusación.

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.

Lo que un estudiante de posgrado debería realmente hacer

El consejo útil no es esforzarse más por sonar humano, algo tan denigrante como inalcanzable. Es conservar el registro, por la misma razón por la que cualquier institución

que enfrente una determinación automatizada no explicada debería conservar el registro: la acusación es infalsable sin evidencia de proceso, y la evidencia de proceso debe crearse antes de que sea necesaria.

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.

Nada de esto tiene que ver con el engaño. Es lo que cualquier persona debería hacer cuando un sistema que no puede ser inspeccionado puede formular una afirmación sobre ella que se le pedirá que refute.

Lo que una institución debería establecer

Cuatro preguntas que una oficina de integridad académica debería poder responder antes de que el resultado de un detector llegue a un estudiante.

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. ¿Cuál es el rendimiento desagregado en las poblaciones a las que sirve la institución, incluidos los estudiantes multilingües, los estudiantes internacionales y los estudiantes que utilizan tecnología asistiva aprobada? Una institución que atiende a estas poblaciones y no puede responder no ha evaluado la herramienta para su propio contexto.
3. ¿Qué peso probatorio se asigna al resultado, expresado por escrito, y depende de él algún hallazgo adverso? Si una puntuación por sí sola puede iniciar un procedimiento, la institución ha adoptado un instrumento sin validar como desencadenante de una medida disciplinaria.

4. ¿Quién revisa la interacción entre las adaptaciones aprobadas y los resultados de detección? Si nadie lo hace, la institución debería asumir que la interacción es adversa, porque el mecanismo lo garantiza.

Conclusión

El consejo de que los estudiantes de posgrado deben pensar por sí mismos, defender sus propias ideas y desarrollar conocimiento especializado en lugar de atajos es correcto, y lo sería en cualquier época. Sin embargo, no responde al riesgo que los estudiantes realmente enfrentan, que no es que se les considere superficiales, sino que un clasificador que mide la inusualidad estadística los marque y se les pida demostrar una negativa.

Ese riesgo se distribuye de manera desigual, y su distribución es lo contrario de lo que la política de integridad debería producir. Los estudiantes con mayor probabilidad de ser acusados injustamente son quienes escriben en una segunda lengua, quienes se comunican mediante lengua de señas como primera lengua, y quienes utilizan las adaptaciones que sus instituciones aprobaron formalmente. Se les investiga con mayor frecuencia no porque sean más propensos a hacer trampa, sino porque un indicador sustitutivo de la escritura maquinal resulta ser, en realidad, un indicador de la escritura bajo restricción.

Las instituciones no eligieron este resultado. Lo adquirieron, mediante una decisión de contratación tomada sin evidencia de validación, y seguirán produciéndolo hasta que alguien formule las preguntas que deberían haberse formulado antes de su despliegue.

Referencias

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.

Ley para Estadounidenses con Discapacidades de 1990 y Sección 504 de la Ley de Rehabilitación de 1973, según se aplican a las instituciones postsecundarias y los ajustes académicos.

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.

Cómo citar

Grizzle, H. M. (2025, May 22). Acusado por Perplexity: la detección de IA como sistema de decisión automatizada sin validar en la educación de posgrado. Novara Consulting Group. https://www.novaracg.com/es/2025/05/22/accused-by-perplexity-ai-detection-as-an-unvalidated-automated-decision-system-in-graduate-education/