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/.

Accusato da Perplexity: il rilevamento tramite IA come sistema decisionale automatizzato non validato nell'istruzione post-laurea

Heather Grizzle · May 22, 2025

I rilevatori di IA misurano quanto la scrittura è insolita, non chi l'ha scritta. Questo penalizza più duramente chi scrive in una seconda lingua e gli studenti che utilizzano tecnologie assistive approvate.

Contenuti

1. Cosa misura il rilevatore
2. I tassi di errore, e chi li subisce
3. La popolazione che ne è colpita, e perché è quella sbagliata
4. Lo standard probatorio è inferiore rispetto a qualsiasi cosa comparabile
5. Cosa dovrebbe effettivamente fare una dottoranda
6. Cosa dovrebbe stabilire un'istituzione
7. Conclusione
8. Riferimenti

Il consiglio che tutti danno è sbagliato riguardo al meccanismo

Il consiglio standard rivolto ai dottorandi riguardo all'intelligenza artificiale suona più o meno così: usala come strumento e non come sostituto, pensa con la tua testa, e tieni presente che i docenti sono persone attente e prima o poi noteranno un lavoro privo di profondità. Le prime due affermazioni sono fondate. La terza è falsa, e la sua falsità conta, perché descrive un processo di giudizio umano che è stato in gran parte sostituito.

La maggior parte delle istituzioni non rileva l'uso dell'intelligenza artificiale attraverso il discernimento di lettori esperti. Lo rilevano, o credono di rilevarlo, tramite classificatori automatici integrati nelle piattaforme di consegna degli elaborati, che assegnano a un documento un punteggio di probabilità e lo inoltrano a un processo di integrità accademica. La domanda che si pone a una dottoranda non è quindi se un professore noterà un ragionamento superficiale. È se un classificatore statistico segnalerà il suo testo, e cosa le accadrà se ciò avviene.

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.

Cosa misura il rilevatore

Un rilevatore di scrittura IA non identifica l'autorialità. Stima quanto un testo sia statisticamente insolito, principalmente tramite la perplessità, che misura quanto sia prevedibile ogni parola rispetto al testo precedente, e la burstiness, che misura la variazione nella struttura e nella lunghezza delle frasi all'interno di un documento. Un testo molto prevedibile e strutturalmente uniforme viene valutato come più probabilmente generato da una macchina.

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.

Questo produce una popolazione specifica e prevedibile di accuse infondate, e non è casuale.

I tassi di errore, e chi li subisce

Una ricerca pubblicata su *Patterns* nel luglio 2023 da Liang e colleghi a Stanford ha rilevato che i rilevatori segnalavano circa il sessantuno percento dei saggi TOEFL scritti da madrelingua non inglesi come generati dall'IA, contro tassi di falsi positivi

quasi nulli per scritti comparabili di madrelingua inglesi, su tutti i sette rilevatori testati. Si trattava di saggi scritti da esseri umani che avevano sostenuto un esame standardizzato per dimostrare la propria competenza in inglese. Gli autori hanno individuato direttamente il meccanismo: i rilevatori si basavano sulla perplessità del testo, e scrivere in una seconda lingua produce una perplessità inferiore.

Test indipendenti non hanno ripetutamente confermato le dichiarazioni di accuratezza dei fornitori. Weber-Wulff e colleghi, pubblicando sull'*International Journal for Educational Integrity* nel dicembre 2023, hanno valutato una serie di strumenti di rilevamento su più tipi di documenti e hanno concluso che non erano né accurati né affidabili. I fornitori, da parte loro, generalmente si rifiutano di pubblicare un tasso di falsi positivi per i loro prodotti, il che significa che un'istituzione che ne adotta uno non può indicare il tasso di errore di uno strumento che utilizza in procedimenti disciplinari.

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 ulteriore dato di mercato merita di essere incluso in qualsiasi valutazione istituzionale. Diversi fornitori che commercializzano strumenti di rilevamento commercializzano anche prodotti "humanizer" progettati per eludere il rilevamento, compreso il proprio. In qualsiasi altro contesto di garanzia, un fornitore che vende sia il test sia il mezzo per superarlo verrebbe escluso dal fornire entrambi.

La popolazione che ne è colpita, e perché è quella sbagliata

La disparità documentata riguarda chi scrive in inglese come seconda lingua, e la spiegazione è meccanica piuttosto che misteriosa. Il rilevamento basato sulla perplessità penalizza

In qualsiasi altro contesto di garanzia, un fornitore che vende sia il test sia il mezzo per superarlo verrebbe

un vocabolario più semplice e una costruzione grammaticale più regolare. Chi scrive in una seconda lingua produce esattamente questo profilo, non perché il testo sia generato da una macchina, ma perché costruire frasi insolite in una seconda lingua è più difficile che costruire frasi corrette.

escluso dal fornire entrambi.

Lo stesso meccanismo colpisce una popolazione che ha ricevuto pochissima attenzione, ed è proprio quella per cui questa società esiste.

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 ciascun caso, è l'accomodamento a produrre la segnalazione. Lo studente che usa il supporto approvato dalla propria istituzione ha più probabilità di essere accusato rispetto allo studente che non ne aveva bisogno, e l'accusa arriva con un numero allegato e nessuna spiegazione di cosa quel numero misuri. Alla studentessa viene quindi chiesto di dimostrare di aver scritto il proprio lavoro, il che equivale a richiedere la prova di un negativo, da parte della persona meno in grado di soddisfarla.

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.

Lo standard probatorio è inferiore rispetto a qualsiasi cosa comparabile

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

Un test di ammissione richiederebbe prove di validità pubblicate, affidabilità accertata, prestazioni documentate tra i diversi gruppi demografici e un punteggio soglia difendibile. Uno strumento di proctoring che attiva un deferimento per integrità richiederebbe quantomeno che la registrazione fosse riesaminabile. Un accertamento di disabilità richiederebbe documentazione e un responsabile della decisione nominato. Un'istituzione che prende una decisione rilevante sulla base di un punteggio numerico non spiegato, prodotto da uno strumento il cui tasso di errore il fornitore non vuole rivelare, sarebbe normalmente considerata aver agito in modo arbitrario.

I punteggi di rilevamento dell'IA sono trattati diversamente, e la ragione sembra essere di genere piuttosto che di principio. Arrivano all'interno di uno strumento antiplagio di cui le istituzioni già si fidano, assomigliano all'indice di somiglianza accanto al quale si trovano, e sono stati adottati tramite procurement piuttosto che tramite governance accademica. L'indice di somiglianza indica almeno un documento fonte che uno studente può ispezionare e contestare. Il punteggio IA non indica nulla. Non c'è fonte, non c'è passaggio, non c'è confronto, e non c'è alcun meccanismo con cui lo studente possa esaminare la base dell'accusa.

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.

Cosa dovrebbe effettivamente fare una dottoranda

Il consiglio utile non è lavorare di più per sembrare umani, cosa che è sia umiliante sia irraggiungibile. È tenere traccia del processo, per lo stesso motivo per cui qualsiasi istituzione di fronte a una determinazione automatizzata non spiegata dovrebbe con-

servare un registro: l'accusa è infalsificabile senza prove di processo, e le prove di processo vanno create prima che siano necessarie.

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.

Niente di tutto questo riguarda l'inganno. È ciò che qualsiasi persona dovrebbe fare quando un sistema che non può essere ispezionato potrebbe avanzare un'affermazione su di lei che le verrà chiesto di confutare.

Cosa dovrebbe stabilire un'istituzione

Quattro domande, a cui un ufficio di integrità accademica dovrebbe essere in grado di rispondere prima che l'output di un rilevatore raggiunga uno studente.

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. Qual è la prestazione disaggregata tra le popolazioni servite dall'istituzione, compresi gli studenti multilingue, gli studenti internazionali e gli studenti che usano tecnologie assistive approvate. Un'istituzione che serve queste popolazioni e non è in grado di rispondere non ha valutato lo strumento per il proprio contesto.
3. Quale peso probatorio viene assegnato all'output, dichiarato per iscritto, e se un eventuale accertamento sfavorevole si basa su di esso. Se un punteggio da solo può avviare un procedimento, l'istituzione ha adottato uno strumento non validato come innesco per un provvedimento disciplinare.

4. Chi esamina l'interazione tra gli accomodamenti approvati e gli output di rilevamento. Se nessuno lo fa, l'istituzione dovrebbe presumere che l'interazione sia sfavorevole, perché il meccanismo lo garantisce.

Conclusione

Il consiglio secondo cui i dottorandi dovrebbero pensare con la propria testa, difendere le proprie idee e sviluppare competenza anziché scorciatoie è corretto e sarebbe corretto in qualsiasi epoca. Non risponde però al rischio che gli studenti effettivamente affrontano, che non è quello di essere giudicati superficiali, ma quello di essere segnalati da un classificatore che misura l'insolitezza statistica e a cui viene chiesto di dimostrare un negativo.

Questo rischio è distribuito in modo disomogeneo, e la sua distribuzione è l'opposto di quella che la politica di integrità dovrebbe produrre. Gli studenti con maggiore probabilità di essere accusati ingiustamente sono quelli che scrivono in una seconda lingua, quelli che si esprimono in lingua dei segni come prima lingua, e quelli che usano gli accomodamenti formalmente approvati dalle loro istituzioni. Vengono indagati più spesso non perché abbiano maggiori probabilità di imbrogliare, ma perché un proxy per la scrittura macchinica si rivela essere un proxy per la scrittura sotto vincolo.

Le istituzioni non hanno scelto questo esito. Lo hanno acquisito, in una decisione di procurement presa senza prove di validazione, e continueranno a produrlo finché qualcuno non porrà le domande che avrebbero dovuto essere poste prima dell'adozione.

Riferimenti

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 del 1990 e Sezione 504 del Rehabilitation Act del 1973, così come applicati agli istituti post-secondari e agli adeguamenti accademici.

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.

Come citare

Grizzle, H. M. (2025, May 22). Accusato da Perplexity: il rilevamento tramite IA come sistema decisionale automatizzato non validato nell'istruzione post-laurea. Novara Consulting Group. https://www.novara-cg.com/it/2025/05/22/accused-by-perplexity-ai-detection-as-an-unvalidated-automated-decision-system-in-graduate-education/