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 pelo Perplexity: A Deteção por IA como um Sistema de Decisão Automatizado Não Validado na Educação de Pós-Graduação

Heather Grizzle · May 22, 2025

Os detetores de IA medem o quão invulgar é um texto, não quem o escreveu. Isso afeta mais gravemente os escritores de segunda língua e os estudantes que utilizam tecnologia assistiva aprovada.

Conteúdo

1. O que o detetor mede
2. As taxas de erro, e quem as carrega
3. A população atingida, e por que razão é a errada
4. O padrão probatório é inferior ao de qualquer situação comparável
5. O que uma estudante de pós-graduação deveria realmente fazer
6. O que uma instituição deveria estabelecer
7. Conclusão
8. Referências

O conselho que todos estão a dar está errado quanto ao mecanismo

O aconselhamento padrão dado a estudantes de pós-graduação sobre inteligência artificial é, grosso modo, o seguinte: usem-na como ferramenta e não como substituto, pensem por vocês mesmos, e compreendam que os docentes são perspicazes e acabarão por notar um trabalho que carece de profundidade. As duas primeiras proposições são válidas. A terceira é falsa, e essa falsidade importa, porque descreve um processo de julgamento humano que foi, em grande medida, substituído.

A maioria das instituições não deteta o uso de inteligência artificial através do discernimento de leitores experientes. Deteta-o, ou acredita que o deteta, através de classificadores automatizados integrados nas plataformas de submissão, que atribuem uma pontuação de probabilidade a um documento e a transmitem a um processo de integridade académica. A questão que se coloca a uma estudante de pós-graduação não é, portanto, se um professor irá notar um pensamento superficial. É se um classificador estatístico assinalará o seu texto, e o que lhe acontecerá se isso suceder.

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.

O que o detetor mede

Um detetor de escrita por IA não identifica autoria. Estima o grau de inusitado estatístico de um texto, principalmente através da perplexidade, que mede quão previsível é cada palavra dado o texto precedente, e da explosividade (burstiness), que mede a variação na estrutura e no comprimento das frases ao longo de um documento. Um texto altamente previsível e estruturalmente uniforme recebe uma pontuação indicando maior probabilidade de ter sido gerado por 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.

Isto produz uma população específica e previsível de acusações falsas, e não se trata de uma população aleatória.

As taxas de erro, e quem as carrega

Uma investigação publicada na revista *Patterns* em julho de 2023 por Liang e colegas de Stanford verificou que os detetores assinalaram aproximadamente sessenta e um por cento dos ensaios do TOEFL escritos por falantes não nativos de inglês como gera-

dos por IA, contra taxas de falsos positivos próximas de zero para escrita comparável de falantes nativos de inglês, nos sete detetores testados. Tratava-se de ensaios escritos por seres humanos que tinham realizado um exame padronizado para demonstrar proficiência em inglês. Os autores identificaram diretamente o mecanismo: os detetores baseavam-se na perplexidade do texto, e escrever numa segunda língua produz menor perplexidade.

Testes independentes têm repetidamente falhado em corroborar as alegações de precisão dos fornecedores. Weber-Wulff e colegas, num artigo publicado no *International Journal for Educational Integrity* em dezembro de 2023, avaliaram uma série de ferramentas de deteção em vários tipos de documentos e concluíram que não eram nem precisas nem fiáveis. Os fornecedores, por seu lado, geralmente recusam-se a publicar uma taxa de falsos positivos dos seus produtos, o que significa que uma instituição que adote uma destas ferramentas não consegue declarar a taxa de erro de um instrumento que está a usar em processos disciplinares.

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.

Há ainda um outro facto de mercado que deve constar de qualquer avaliação institucional. Vários fornecedores que comercializam ferramentas de deteção também comercializam produtos "humanizadores", concebidos para iludir a deteção, incluindo a sua própria. Em qualquer outro contexto de garantia de qualidade, um fornecedor que vendesse simultaneamente o teste e os meios de o passar seria desqualificado de fornecer qualquer um dos dois.

A população atingida, e por que razão é a errada

A disparidade documentada diz respeito a escreventes de inglês não nativos, e a explicação é mecânica, não misteriosa. A deteção basea-

Em qualquer outro contexto de garantia de qualidade, um fornecedor que vendesse si-

da na perplexidade penaliza vocabulário mais simples e construções gramaticais mais regulares. Uma pessoa que escreve numa segunda língua produz exatamente esse perfil, não porque o texto seja gerado por máquina, mas porque construir frases invulgares numa segunda língua é mais difícil do que construir frases corretas.

multaneamente o teste e os meios de o passar seria desqualificado de fornecer qualquer um dos dois.

O mesmo mecanismo atinge uma população que tem recebido quase nenhuma atenção, e é precisamente essa a razão de ser desta empresa.

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.

Em cada caso, é a adaptação que gera o alerta. A estudante que utiliza o apoio aprovado pela sua instituição tem mais probabilidade de ser acusada do que a estudante que não precisou dele, e a acusação chega acompanhada de um número, sem qualquer explicação sobre o que esse número mede. É-lhe então exigido que prove que escreveu o seu próprio trabalho, uma exigência de prova negativa, feita precisamente à pessoa menos capaz de a satisfazer.

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.

O padrão probatório é inferior ao de qualquer situação comparável

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

Um teste de admissão exigiria evidência de validade publicada, fiabilidade estabelecida, desempenho documentado entre grupos demográficos e uma pontuação de corte defensável. Uma ferramenta de vigilância eletrónica que desencadeasse um encaminhamento por questões de integridade exigiria, no mínimo, que a gravação pudesse ser revista. Uma determinação de deficiência exigiria documentação e um responsável de decisão identificado. Entender-se-ia, normalmente, que uma instituição que tomasse uma decisão consequente com base numa pontuação numérica não explicada, produzida por um instrumento cuja taxa de erro o fornecedor se recusa a divulgar, teria agido de forma arbitrária.

As pontuações de deteção de IA são tratadas de forma diferente, e a razão parece ser de género, e não de princípio. Chegam inseridas numa ferramenta antiplágio em que as instituições já confiam, parecem-se com o índice de similaridade ao lado do qual são apresentadas, e foram adotadas por via de procurement, e não através da governação académica. O índice de similaridade pelo menos aponta para um documento-fonte que a estudante pode examinar e contestar. A pontuação de IA não aponta para nada. Não há fonte, não há passagem, não há comparação, nem qualquer mecanismo através do qual a estudante possa examinar o fundamento da alegação.

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.

O que uma estudante de pós-graduação deveria realmente fazer

O conselho útil não é esforçar-se mais por soar humano, o que é simultaneamente degradante e inatingível. É manter o registo, pela mesma razão pela qual qualquer instituição confrontada com uma determinação automatizada não explicada deveria manter o registo: a acusação é infalsificável sem prova processual, e a prova processual tem de ser criada antes de ser necessária.

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 disto diz respeito a engano. É o que qualquer pessoa deveria fazer perante um sistema que não pode ser inspecionado e que pode fazer sobre ela uma afirmação que lhe será pedido que refute.

O que uma instituição deveria estabelecer

Quatro perguntas às quais um gabinete de integridade académica deveria conseguir responder antes de o resultado de um detetor chegar a um estudante.

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 é o desempenho desagregado entre as populações que a instituição serve, incluindo estudantes multilingues, estudantes internacionais e estudantes que utilizam tecnologia assistiva aprovada. Uma instituição que serve estas populações e não consegue responder não avaliou a ferramenta para o seu próprio contexto.

3. Que peso probatório é atribuído ao resultado, declarado por escrito, e alguma conclusão desfavorável assenta nele. Se uma pontuação, por si só, puder dar início a um processo, a instituição adotou um instrumento não validado como fator desencadeador de medidas disciplinares.
4. Quem revê a interação entre as adaptações aprovadas e os resultados de deteção. Se ninguém o faz, a instituição deve presumir que a interação é desfavorável, porque o mecanismo assim o garante.

Conclusão

O conselho de que os estudantes de pós-graduação devem pensar por si mesmos, defender as suas próprias ideias e desenvolver conhecimento especializado em vez de atalhos é correto, e seria correto em qualquer época. Também não responde ao risco que os estudantes realmente enfrentam, que não é o de serem considerados superficiais, mas o de serem assinalados por um classificador que mede o grau de inusitado estatístico e lhes ser pedido que provem uma negativa.

Esse risco está distribuído de forma desigual, e a sua distribuição é o oposto daquela que a política de integridade pretende produzir. Os estudantes com maior probabilidade de serem acusados injustamente são os que escrevem numa segunda língua, os que gesticulam numa primeira e os que utilizam as adaptações formalmente aprovadas pelas suas instituições. São investigados com mais frequência não porque sejam mais propensos a fraudar, mas porque uma aproximação para escrita de máquina acaba por ser uma aproximação para escrita sob constrangimento.

As instituições não escolheram este resultado. Adquiriram-no, numa decisão de procurement tomada sem evidência de validação, e continuarão a produzi-lo até que alguém faça as perguntas que deveriam ter sido feitas antes da implementação.

Referências

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 de 1990 e Section 504 do Rehabilitation Act de 1973, conforme aplicados a instituições pós-secundárias e a 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.

Como citar

Grizzle, H. M. (2025, May 22). Acusado pelo Perplexity: A Deteção por IA como um Sistema de Decisão Automatizado Não Validado na Educação de Pós-Graduação. Novara Consulting Group. https://www.novaracg.com/pt/2025/05/22/accused-by-perplexity-ai-detection-as-an-unvalidated-automated-decision-system-in-graduate-education/