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/.

Accusé par Perplexity : la détection par IA comme système de décision automatisé non validé dans l'enseignement supérieur

Heather Grizzle · May 22, 2025

Les détecteurs d'IA mesurent le caractère inhabituel d'un texte, pas qui l'a écrit. Cela pèse le plus lourdement sur les auteurs qui écrivent dans une langue seconde et les étudiants utilisant des technologies d'assistance approuvées.

Sommaire

1. Ce que mesure le détecteur
2. Les taux d'erreur, et qui les subit
3. La population que cela touche, et pourquoi ce n'est pas la bonne
4. La norme de preuve est inférieure à celle appliquée à tout élément comparable
5. Ce qu'un étudiant de troisième cycle devrait réellement faire
6. Ce qu'un établissement devrait établir
7. Conclusion
8. Références

Le conseil que tout le monde donne se trompe sur le mécanisme

Le conseil standard donné aux étudiants de troisième cycle sur l'intelligence artificielle se résume à peu près ainsi : utilisez-la comme un outil plutôt que comme un substitut, réfléchissez par vous-même, et sachez que le corps professoral est perspicace et finira par remarquer un travail qui manque de profondeur. Les deux premières

propositions sont justes. La troisième est fausse, et sa fausseté a de l'importance, car elle décrit un processus de jugement humain qui a largement été remplacé.

La plupart des établissements ne détectent pas l'usage de l'intelligence artificielle grâce au discernement de lecteurs expérimentés. Ils le détectent, ou croient le détecter, au moyen de classificateurs automatisés intégrés aux plateformes de soumission, qui attribuent à un document un score de probabilité avant de le transmettre à une procédure disciplinaire d'intégrité académique. La question à laquelle une étudiante de troisième cycle est confrontée n'est donc pas de savoir si un professeur remarquera une pensée superficielle. C'est de savoir si un classificateur statistique signalera son écrit, et ce qu'il lui en coûtera si c'est le cas.

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.

Ce que mesure le détecteur

Un détecteur d'écriture par IA n'identifie pas la paternité d'un texte. Il estime dans quelle mesure un texte est statistiquement inhabituel, principalement via la perplexité, qui mesure la prévisibilité de chaque mot compte tenu du texte qui précède, et la « burstiness » (irrégularité), qui mesure la variation de la structure et de la longueur des phrases au sein d'un document. Un texte hautement prévisible et structurellement uniforme obtient un score indiquant une plus forte probabilité qu'il soit généré par une machine.

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.

Cela produit une population spécifique et prévisible d'accusations infondées, et celle-ci n'a rien d'aléatoire.

Les taux d'erreur, et qui les subit

Une recherche publiée dans *Patterns* en juillet 2023 par Liang et ses collègues de Stanford a révélé que les détecteurs signalaient environ soixante et un pour cent des essais TOEFL rédigés par des locuteurs non natifs de l'anglais comme générés par IA, contre des taux de faux positifs proches de zéro pour des écrits comparables rédigés par des locuteurs natifs de l'anglais, et ce pour les sept détecteurs testés. Il s'agissait d'essais rédigés par des êtres humains ayant passé un examen normalisé pour démontrer leur maîtrise de l'anglais. Les auteurs ont identifié directement le mécanisme en cause : les détecteurs se fondaient sur la perplexité du texte, et écrire dans une langue seconde produit une perplexité plus faible.

Des tests indépendants n'ont cessé de contredire les affirmations des fournisseurs sur l'exactitude de leurs outils. Weber-Wulff et ses collègues, publiant dans l'*International Journal for Educational Integrity* en décembre 2023, ont évalué un ensemble d'outils de détection sur plusieurs types de documents et ont conclu qu'ils n'étaient ni précis ni fiables. Les fournisseurs, pour leur part, refusent généralement de publier un taux de faux positifs pour leurs produits, ce qui signifie qu'un établissement en adoptant un ne peut pas indiquer le taux d'erreur d'un instrument qu'il utilise dans le cadre de procédures disciplinaires.

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 fait de marché supplémentaire mérite sa place dans toute évaluation institutionnelle. Plusieurs fournisseurs d'outils de détection commercialisent également des produits « humanisateurs » conçus pour déjouer la détection, y compris la leur propre. Dans tout autre contexte d'assurance qualité, un fournisseur vendant à la fois le test et le moyen de le réussir serait disqualifié de la fourniture de l'un comme de l'autre.

Dans tout autre

La population que cela touche, et pourquoi ce n'est pas la bonne

contexte d'assurance qualité, un fournisseur vendant à la fois le test et le moyen de le réussir serait disqualifié de la fourniture de l'un comme de l'autre.

L'écart documenté concerne les auteurs non natifs de l'anglais, et l'explication est mécanique plutôt que mystérieuse. La détection fondée sur la perplexité pénalise un vocabulaire plus simple et une construction grammaticale plus régulière. Une personne écrivant dans une langue seconde produit exactement ce profil, non parce que l'écrit est généré par une machine, mais parce que construire une formulation inhabituelle dans une langue seconde est plus difficile que de construire une formulation correcte.

Le même mécanisme touche une population qui n'a presque pas reçu d'attention, et c'est précisément celle pour laquelle ce cabinet existe.

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.

Dans chaque cas, c'est l'aménagement lui-même qui déclenche le signalement. L'étudiante qui utilise le soutien approuvé par son établissement est plus susceptible d'être accusée que celle qui n'en avait pas besoin, et l'accusation arrive accompagnée d'un chiffre, sans aucune explication de ce que ce chiffre mesure. On exige alors d'elle qu'elle prouve qu'elle a bien écrit son propre travail, ce qui revient à exiger la preuve d'un fait négatif, de la part de la personne la moins à même de la fournir.

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.

La norme de preuve est inférieure à celle appliquée à tout élément comparable

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

Un test d'admission nécessiterait des preuves de validité publiées, une fiabilité établie, des performances documentées selon les groupes démographiques, et un seuil de réussite défendable. Un outil de surveillance déclenchant un signalement pour atteinte à l'intégrité exigerait, au minimum, que l'enregistrement soit consultable. Une détermination de handicap exigerait une documentation et un décideur nommément désigné. Un établissement prenant une décision lourde de conséquences sur la base d'un score numérique inexpliqué, produit par un instrument dont le fournisseur refuse de divulguer le taux d'erreur, serait ordinairement considéré comme ayant agi arbitrairement.

Les scores de détection d'IA sont traités différemment, et la raison semble tenir au genre de l'outil plutôt qu'à un quelconque principe. Ils apparaissent au sein d'un outil anti-plagiat auquel les établissements font déjà confiance, ils ressemblent à l'indice de similarité auquel ils sont juxtaposés, et ils ont été adoptés par la voie des achats plutôt que par la gouvernance académique. L'indice de similarité pointe au moins vers un document source qu'un étudiant peut consulter et contester. Le score d'IA ne pointe vers rien. Il n'y a ni source, ni passage, ni comparaison, ni aucun moyen par lequel l'étudiant pourrait examiner le fondement de l'accusation.

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 distribu-

tion above, it is a mechanism for deciding that multilingual and disabled students get investigated more often.

Ce qu'un étudiant de troisième cycle devrait réellement faire

Le conseil utile n'est pas de s'efforcer davantage de « sonner humain », ce qui est à la fois humiliant et irréalisable. C'est de conserver des traces, pour la même raison que tout établissement confronté à une détermination automatisée inexpliquée devrait conserver des traces : l'accusation est infalsifiable en l'absence de preuves de processus, et ces preuves doivent être constituées avant d'en avoir besoin.

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.

Rien de tout cela ne relève de la tromperie. C'est ce que toute personne devrait faire face à un système impossible à inspecter et susceptible de formuler à son sujet une allégation qu'on lui demandera ensuite de réfuter.

Ce qu'un établissement devrait établir

Quatre questions auxquelles un service d'intégrité académique devrait pouvoir répondre avant que le résultat d'un détecteur n'atteigne un étudiant.

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. Quelle est la performance désagrégée selon les populations que l'établissement dessert, y compris les étudiants plurilingues, les étudiants internationaux et les étudiants utilisant une technologie d'assistance approuvée. Un

établissement desservant ces populations et incapable de répondre n'a pas évalué l'outil pour son propre contexte.

3. Quel poids probant est attribué au résultat, consigné par écrit, et toute conclusion défavorable en dépend-elle. Si un score à lui seul peut déclencher une procédure, l'établissement a adopté un instrument non validé comme élément déclencheur de sanctions disciplinaires.
4. Qui examine l'interaction entre les aménagements approuvés et les résultats de détection. Si personne ne le fait, l'établissement devrait présumer que cette interaction est défavorable, car le mécanisme le garantit.

Conclusion

Le conseil selon lequel les étudiants de troisième cycle devraient réfléchir par eux-mêmes, défendre leurs propres idées et développer une expertise plutôt que de recourir à des raccourcis est juste, et le serait à toute époque. Il ne répond toutefois pas au risque que les étudiants affrontent réellement, qui n'est pas d'être jugés superficiels, mais d'être signalés par un classificateur mesurant une anomalie statistique et sommés de prouver un fait négatif.

Ce risque est réparti de manière inégale, et sa répartition est l'inverse de celle que la politique d'intégrité est censée produire. Les étudiants les plus susceptibles d'être accusés à tort sont ceux qui écrivent dans une langue seconde, ceux qui signent dans une langue première, et ceux qui utilisent les aménagements formellement approuvés par leurs établissements. Ils font l'objet d'enquêtes plus fréquentes non parce qu'ils sont plus susceptibles de tricher, mais parce qu'un indicateur censé révéler l'écriture par machine se révèle en réalité être un indicateur de l'écriture sous contrainte.

Les établissements n'ont pas choisi ce résultat. Ils l'ont acquis, à l'occasion d'une décision d'achat prise sans preuve de validation, et ils continueront à le produire tant que personne ne posera les questions qui auraient dû être posées avant le déploiement.

Références

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 et section 504 du Rehabilitation Act de 1973, tels qu'appliqués aux établissements postsecondaires et aux aménagements académiques.

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.

Comment citer

Grizzle, H. M. (2025, May 22). Accusé par Perplexity : la détection par IA comme système de décision automatisé non validé dans l'enseignement supérieur. Novara Consulting Group. https://www.novaracg.com/fr/2025/05/22/accused-by-perplexity-ai-detection-as-an-unvalidated-automated-decision-system-in-graduate-education/