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/.

# Perplexity 的指控:研究生教育中未经验证的自动化决策系统——AI检测

Heather Grizzle · May 22, 2025

*AI*检测器衡量的是文字有多不寻常，而不是谁写的。这对使用第二语言写作的人和使用经批准辅助技术的学生打击最大。

**目录**

1. 检测器所测量的对象
2. 错误率,以及谁在承受它
3. 这一机制波及的人群,以及为何这是错误的对象
4. 其证据标准低于任何可比事物
5. 研究生实际应该怎么做
6. 院校应当建立的机制
7. 结语
8. 参考文献

## 所有人给出的建议在机制层面上都是错误的

针对研究生使用人工智能的标准建议大致如下:把它当作工具而非替代品,自己动脑思考,并且要明白教师是敏锐的,最终会察觉到缺乏深度的作品。前两条主张是成立的。第三条是错误的,而这个错误至关重要,因为它描述的是一种人类判断的过程,而这种过程如今在很大程度上已被取代。

大多数院校检测人工智能使用,靠的并非经验丰富的阅读者的辨识力。它们依靠的——或者说自认为依靠的——是嵌入在提交平台中的自动分类器,这些分类器会为文档赋予一个概率分数,并将其纳入学术诚信审查流程。因此,研究生面临的问题不是教授是否会察觉

到肤浅的思考,而是统计分类器是否会标记她的写作,以及一旦被标记,她将面临什么后果。

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.

## 检测器所测量的对象

人工智能写作检测器并不能识别作者身份。它估算的是一段文本在统计意义上有多“不寻常”,主要通过两个指标:困惑度(perplexity),衡量在给定前文的情况下每个词的可预测程度;以及突发性(burstiness),衡量文档中句子结构和长度的变化程度。高度可预测且结构统一的文本,会被判定为更可能是机器生成的。

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.

这会产生一群特定且可预见的被误判者,而这绝非随机分布。

## 错误率,以及谁在承受它

2023年7月发表于*Patterns*期刊的一项研究,由斯坦福大学的Liang及其同事完成,发现在所测试的七种检测器中,非母语英语使用者撰写的TOEFL作文,约有百分之六十一被标记为人工智能生成,而母语为英语者所写的可比文本,其误报率则接近于零。这些作文是由通过标准化考试来证明其英语水平的真实人类所写。作者直接指出了其中的机制:检测器是根据文本困惑度进行判断的,而使用第二语言写作会产生更低的困惑度。

Independent testing has repeatedly failed to support vendor accuracy claims. Weber-Wulff and colleagues, publishing in the *International Journal for Educational Integrity* in December 2023, evaluated a range of detection tools across multiple document types and concluded that they were neither accurate nor reliable. Vendors,

for their part, generally decline to publish a false positive rate for their products, which means an institution adopting one cannot state the error rate of an instrument it is using in disciplinary proceedings.

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.

还有一个市场事实,理应纳入任何机构层面的评估之中。数家营销检测工具的供应商,同时也在营销旨在规避检测(包括规避它们自家检测工具)的“人性化”产品。在任何其他保障性场景中,一个既卖测试又卖通过测试之手段的供应商,理应被取消提供其中任何一项的资格。

## 这一机制波及的人群,以及为何这是错误的对象

已有文献记录的这种差异,涉及的是非母语英语写作者,其原因是机械性的,而非神秘莫测的。基于困惑度的检测,会惩罚更简单的词汇和更规则的语法构造。使用第二语言写作的人,恰恰会产生这样的特征,这并非因为其写作是机器生成的,而是因为用第二语言构造出不寻常的措辞,远比构造出正确的措辞更难。

在任何其他保障性场景中,一个既卖测试又卖通过测试之手段的供应商,理应被取消提供其中任何一项的资格。

同样的机制还波及了一个几乎从未受到关注的群体,而这正是本公司存在的意义所在——去关注这一群体。

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.

在每一种情形中,正是这项辅助措施引发了标记。使用了院校批准的辅助手段的学生,反而比不需要这种手段的学生更可能被指控,而这项指控附带一个数字,却没有任何关于该数字所衡量内容的解释。她随后被要求证明自己独立完成了作品——这是要求她证明一件不存在的事,而她恰恰是最没有能力满足这一要求的人。

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.

## 其证据标准低于任何可比事物

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

一项入学考试需要公开的效度证据、既定的信度、跨人口群体表现的详实记录,以及站得住脚的分数线。一项引发诚信转介程序的监考工具,至少需要录像可供复查。一项残障认定需要相关文件记录以及一位具名的决策者。而一个院校若仅凭一个无从解释的数值分数——该分数由一种供应商拒绝披露其错误率的工具生成——就做出具有实质后果的决定,通常会被认为是武断行事。

人工智能检测分数却被区别对待,而原因似乎在于其“外观归属”而非原则本身。它们是嵌入在院校已经信任的查重工具内部出现的,看起来就像旁边那个相似度指数一样,而且它们是通过采购流程被采用的,而非通过学术治理程序。相似度指数至少指向一份学生可以查阅并质疑的原始文献。而人工智能分数则什么都不指向。没有来源,没有段落,没有比对,学生也没有任何途径去审查这项指控的依据。

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.

## 研究生实际应该怎么做

有用的建议不是要更努力地让自己“听起来像人”,这种建议既是一种贬低,也是不可能实现的。真正有用的做法,是保留记录——原因和任何面对无法解释的自动化裁定的机构应当保留记录的原因一样:这类指控在没有过程证据的情况下是无法被证伪的,而过程证据必须在需要之前就被创造出来。

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.

这一切都与欺骗无关。这只是任何人在面对一个无法被审查、却可能对自己做出指控、并要求自己反驳的系统时,理应采取的做法。

## 院校应当建立的机制

有四个问题,学术诚信办公室应当在检测器的输出结果送达学生之前就能够回答。

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. 该工具在院校所服务的各类群体中的表现分别如何,包括多语言学生、国际学生以及使用经批准的辅助技术的学生。如果一所院校服务着这些群体,却无法回答这一问题,那就说明它尚未针对自身的具体情境评估过这一工具。
3. 该输出结果被赋予多大的证据权重,这一点是否已书面明确规定,以及是否有任何不利认定完全建立在其基础之上。如果仅凭一个分数就能启动一项处分程序,那么该院校就是把一种未经验证的工具当作了触发纪律处分的依据。
4. 谁来审查经批准的辅助措施与检测输出结果之间的相互作用。如果无人负责此事,该院校就应当假定这种相互作用是不利的,因为这一机制必然导致这种结果。

## 结语

主张研究生应当独立思考、捍卫自己的观点、培养真才实学而非走捷径的建议是正确的,而且在任何时代都会是正确的。但这一建议并未回应学生实际面临的风险——他们所面临的风险,不是被发现思想肤浅,而是被一个衡量统计学“不寻常性”的分类器标记,进而被要求证明一件不存在的事。

这种风险的分布是不均衡的,而其分布方式恰恰与诚信政策本应产生的效果背道而驰。最有可能被错误指控的学生,是那些用第二语言写作的人、那些以第一语言进行手语表达的人,以及那些使用了院校正式批准的辅助措施的人。他们之所以更频繁地被调查,并非因为他们更可能作弊,而是因为一种本应代表“机器写作”的替代指标,结果却变成了代表“受限制状态下写作”的替代指标。

院校并非有意选择了这一结果。它们是在一次未经验证的采购决策中获得了这一结果,而且只要没有人提出那些本应在部署之前就提出的问题,它们就将继续产生这一结果。

## 参考文献

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.

1990年《美国残疾人法案》和1973年《康复法案》第504条,适用于高等院校及学术调整事宜。

**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.

---

## 引用方式

Grizzle, H. M. (2025, May 22). Perplexity 的指控:研究生教育中未经验证的自动化决策系统——AI检测. Novara Consulting Group. https://www.novaracg.com/zh/2025/05/22/accused-by-perplexity-ai-detection-as-an-unvalidated-automated-decision-system-in-graduate-education/