arXiv:2603.20254v1 [cs.CY] 11 Mar 2026

AI检测器无法适用于多元化的学生群体：对检测结构性局限的数学阐释

N.A. Garland昆士兰量子与先进技术研究所，格里菲斯大学，昆士兰州，澳大利亚；格里菲斯大学环境与科学学院，昆士兰州，澳大利亚

2026年3月24日

摘要

学生的亲身经历和实证研究表明，“黑箱”人工智能文本检测器会产生较高的假阳性率，并对某些学生群体造成不成比例的误判；然而，理论分析通常将检测建模为针对人类与AI生成文本这两种已知分布的检验。这种框架忽略了大学评估中的结构性特征：评估者通常并不知道每位学生各自的写作分布，因此零假设是复合的。将全变差距离的变分刻画标准地应用于这一复合原假设，可得出权衡界限：任何具备实用检验功效的纯文本单次检测器，其错误指控率必然受学生写作与人工智能输出之间的分布重叠程度制约。这是由群体多样性导致的限制，在逻辑上独立于人工智能模型质量，无法通过改进检测器工程或技术来克服。子群体混合界将这些量与可观测人口群体联系起来，为实证研究记录的差异化影响模式提供了理论依据。我们提出了改进政策与实践的建议，并认为检测分数不应作为学术不端程序中的唯一证据。

1 引言

在大语言模型（LLMs）时代，出于对学术诚信的合理担忧，AI文本检测工具已广泛应用于大学评估中（参见 Tight（2024年）；Luo（2024年））。然而，越来越多的实证研究记录了这些工具存在的严重问题。例如，Weber-Wulff 等人（2023年）测试了14种检测工具，发现没有一种工具达到80%的准确率，且所有工具都会产生假阳性（将人类撰写的文本标记为人工智能生成文本）和假阴性（将人工智能生成文本判定为人类撰写文本）。Liang 等人（2023年）证明，检测器会不成比例地标记非英语母语者的文本，引发了对公平性的担忧。Elkhatat 等人（2023年）发现，将检测工具应用于人类撰写的对照样本时，不同工具之间存在不一致性。最近，Hadra 等人（2026年）评估了服务于英语作为外语的学习者的机构所使用的检测器的准确性，得出结论认为，当前工具“不适合作为作者身份的权威裁决者”，机构应“谨慎且审慎地”使用检测。教育测量学界已经认识到大语言模型对评估效度和公平性的更广泛影响 Hao等（2024年），但迄今尚未建立起一个将总体多样性与结构性检测局限联系起来的正式统计框架。

这些实证发现描绘出一致的图景：当前的人工智能文本检测器并不可靠，尤其是在多元化的学生群体中。但现有文献大多将各项研究的发现视为特定 检测器的属性，认为可以通过更好的工程设计加以改进。在本文中，我们提出了一个不同的问题：是否存在结构性原因，使得任何仅基于文本、进行单次检测的 检测器都必须在检测功效与错误指控之间权衡，而这一权衡与检测器质量无关?

我们证明，答案是肯定的，而这一结构性障碍的根源在于群体多样性——学生群体在写作风格、能力和语言背景方面存在的自然差异。这一点不同于、但又补充了广泛讨论的另一个问题：人工智能文本质量逐渐趋近于人类文本（Sadasivan 等，2023；Chakraborty 等人，2024）。

现有的理论分析（Sadasivan 等，2023；Chakraborty 等人，2024）将人工智能检测建模为两个已知分布之间的二元检验： p_H （“人类”分布）与 p_M （“人工智能”分布）。这些分析表明，随着这两个分布趋于一致，检测结果将接近随机猜测。

但在大学中，并不存在单一的“人类分布”。每位学生的写作方式都不同，评估者在评估一份提交文本时，通常并不了解某位学生个人的写作特征。零假设并非“该文本来自 p_H”，而是：

“该文档由学生 i 撰写，其写作分布 p_{θ_i} 未知。”

用统计假设检验的语言来说（Lehmann & Romano，2022），这是一种复合零假设——零假设分布并非单个已知实体，而是一个由学生特征索引的分布族 $\{p_\theta : \theta \in \Theta\}$ 。这种看似简单的重新框定，会对任何检测器所能实现的性能产生具体且可量化的影响。

在本文中，我们提出将“黑箱”式大学一次性AI检测建模为具有复合原假设的复合假设检验，从而明确揭示先前理论工作所忽视的一项结构性特征。我们将群体多样性识别为一种在逻辑上独立于AI质量的检测困难来源，并精确说明了哪些政策杠杆可以应对每种情形（第4.1节）。我们给出了一个界限，帮助将理论重叠量与可观测的真实学生子群体联系起来，从而为 Liang 等人（2023）、Weber-Wulff 等人（2023）、Elkhatat 等人（2023）和 Hadra 等人（2026）所记录的偏差与审计证据提供桥梁。最后，我们提出了一套具体的分层假阳性率审计程序，机构可利用现有数据实施该程序（第4.5节）。

证明技巧是对总变差（TV）距离变分不等式的标准应用。我们并未声称在假设检验理论中发现了新结果。本文的贡献在于应用层面的形式化、对失效模式进行概念区分，以及通过运用基本统计学定义的相对简单的数学练习，将分析与机构实践联系起来。

2 背景

我们简要回顾文献中的两项主要理论结果，并着重指出各自所依赖的结构性假设。关键的数学工具是两个概率分布 P 和 Q之间的全变差 （TV）距离，定义如下

$$\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)| = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)|\, d\mu(x), \tag{1}$$

其中，sup 表示上确界， A 是所有可能文本文件构成的集合（样本空间） $\mathcal{X}$中的任一可能事件，其参考测度为 $\mu(x)$。

对于任意检验函数（我们稍后会将其形式化为检测器） $\phi : \mathcal{X} \to [0,1]$，变分不等式给出

$$|\mathbb{E}_P[\phi] - \mathbb{E}_Q[\phi]| \le \mathrm{TV}(P,Q), \tag{2}$$

其中， $\mathbb{E}_P[\phi] = \int_{\mathcal{X}} \phi(x)p(x)d\mu(x)$表示在给定分布 P下 ϕ的期望。这一不等式是我们分析中唯一使用的数学工具。

就分布收敛而言，Sadasivan 等人（2023）将 (2) 应用于已知的人类分布 P_H 和人工智能分布 P_M，并证明任何检测器的检验功效超过其假阳性率的幅度最多为全变差距离$(P_H,\ P_M)$。随着人工智能模型不断进步，且全变差距离$(P_M,\ P_H) \to 0$，所有检测指标都会退化。一个关键假设是， P_H 和 P_M 都是单一且已知的分布。

对于多样本检测，Chakraborty 等人（2024）的研究表明，要实现可靠检测，需要$O(1/\delta^2)$个独立样本，条件是全变差距离$(p_M,\ p_H) = \delta > 0$。在单次设定（$n = 1$）下，当 δ 很小时，这无法提供有用的保证。同样，他们的研究假设两个分布均为已知，且有多个独立同分布的样本可供使用。

结构性差距在于，这两项研究都没有讨论这样的情形：零假设分布随总体而变化，且对每个个体而言都是未知的。这恰恰就是大学评估中的情形。

单次人工智能检测问题在结构上也与计算文体学中无注册的作者身份验证相关（Koppel 和 Schler，2004年；Stamatatos，2008年；Dawson 等人，2019年）。在这种情形下，

必须判断一份待鉴定文档是否由某位声称的作者撰写，而几乎没有（或仅有非常有限的）该作者的参考材料。

这两个问题都有一个复合原假设，因为作者分布未知。在没有注册数据的情况下进行验证存在根本性困难，这一点在文体计量学界已得到充分确立（Koppel & Schler，2004年；Stamatatos，2008年）。

关键的结构性不对称在于，备择分布 p_M 原则上是已知的，因为检测器软件设计者可以查询人工智能系统、刻画其输出分布并制定规则。传统的作者身份验证面临双重复合 问题（声称的作者未知，且 备择作者也未知）。我们的结果1–3利用了备择分布已知这一结构。具体而言，结果 3 利用了p_M 可供采样这一事实，来构建子群体与人工智能分类器，并生成审计协议所需的人工智能侧语料库。这种不对称性并不会让检测变得容易；复合原假设仍决定着假阳性率权衡，但它使我们提出的一种特定形式的审计成为可能。

3 方法

3.1 大学检测问题的设定

大学收到学生 i 提交的一份文档 x，用于评估任务 τ。评估者希望判断学生 i是否写了 x，或者该文档是否主要由人工智能系统生成。每名学生的写作特点由一个未知参数 $\theta_i \in \Theta$ 刻画，该参数编码了写作能力、领域知识、风格、努力程度和语言背景。学生从分布$p_{\theta_i,\ \tau}$中生成文档，而机器生成文档则从$p_{M,\ \tau}$中生成。本文中的所有界限均针对单个任务，适用于固定任务τ 及其相关分布。为简化符号，除非重点讨论任务依赖性（第 4.3 节），否则我们省略 τ。

3.2 复合原假设

零假设为

$$H_0 : x \sim p_\theta \quad \text{for some unknown } \theta \in \Theta.$$

检测器是一个可测函数 $\phi : \mathcal{X} \to [0,1]$。对于文档 x，$\phi(x)$ 表示检测器将 x 判定为人工智能生成的概率。我们定义单个学生的假阳性率（FPR）$\alpha(\theta;\phi) = \mathbb{E}_{p_\theta}[\phi(x)],$，而检测器功效（真阳性率 / TPR）定义为 $\beta(\phi) = \mathbb{E}_{p_M}[\phi(x)]$。令 π 为 Θ 上的概率测度，对于任意可测子集 $S \subseteq \Theta$，$\pi(S)$ 表示类型为 θ 的学生中落入 S 的比例。

对于某个全变差距离阈值 δ，我们定义重叠集合 $\Theta^*(\delta) = \{\theta \in \Theta : \mathrm{TV}(p_\theta,\ p_M) \le \delta\}$，即写作与人工智能输出之间的全变差距离不超过 δ 的学生集合。通常，所有重叠量都视为特定于任务，因此 Θ^*、δ^* 和子群混合总变差距离取决于 τ。图1展示了复合原假设的结构，以及它与先前研究中通常假设的简单对简单假设情形有何不同。

3.3 检测器必然面临的理论结果

现在我们总结三项主要的数学结果。每项结果都是将变分不等式 (2) 应用于复合原假设结构所得出的推论。有兴趣的读者可在附录A中查阅形式化证明。

3.3.1 平均情形权衡

结果1：平均情形大小–功效权衡

假设有比例为 $\pi(\Theta^*)$ 的学生，其写作分布与人工智能输出之间的全变差距离为 δ。那么，**任何** 检验功效（真阳性率）为 β_0 的检测器，其总体平均假阳性率 $\bar{\alpha}$ 至少为

$$\bar{\alpha} \ge \pi(\Theta^*) \cdot (\beta_0 - \delta).$$

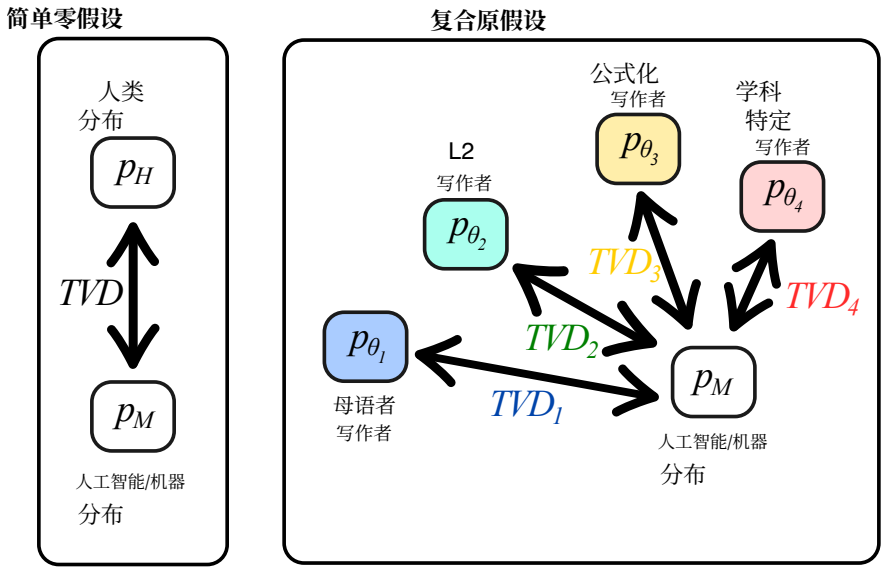


图1：检测结构的示意比较。**左侧：** 先前的理论研究假设存在一个已知的人类分布 p_H 和一个已知的人工智能分布 p_M。**右侧：** 大学环境涉及一个复合原假设——每位学生的写作分布各不相同 p_θ，而检测器不知道适用的是哪一种θ 。一些学生（例如英语为第二语言的学生，或撰写采用结构化规范的学科特定评估的学生）的分布可能与 p_M 非常接近，全变差距离也很小。

为了说明这一结果的含义，我们给出一个概念示例。假设有这样一种情况：10% 的学生写作方式与人工智能输出接近，即与预期 AI 分布的名义全变差距离为 0.05（$\pi(\Theta^*) = 0.10, \delta = 0.05$），那么，检验功效（$\beta_0 = 0.8$）为 80% 的检测器平均至少会导致 $\bar{\alpha} \geq 0.10 \times (0.80 - 0.05) \geq 0.075$ 起错误指控（按每名受测学生计）。在一所有 10,000 名学生的机构中，这可能意味着约 750 起错误指控；这并不是因为检测器设计不佳，而是由这一问题的数学结构所决定的。

我们指出，这些参数仅用于本例示意，目的是复现文献中经实证展示的约10%–20%的假阳性率。 π（Θ^*）和 δ 的实际取值尚无实证结论，并且会因机构和学科而异，无法仅凭理论估计。我们强调，本例展示的是这种权衡的形式，而非可用于特定机构的定量预测工具；这一结果最好被视为一种阐释视角，借此解释实证结果以及高等教育界反馈的高假阳性率案例。

为帮助直观呈现这一概念性权衡，以及假阳性率数值显著上升的情况，图2展示了结果1中的下界在整个参数空间中的变化。当 π（Θ^*）超过约5%，且 δ 处于中等水平时，该下界便会超过机构通常设定的1–5%容忍阈值。尽管 π（Θ^*）和 δ 都尚未以分布意义上的方式直接测量，实证文献已将它们限定在一个使该下界具有参考意义的范围内。就重叠比例 π（Θ^*）而言：在不同类型的机构中，国际学生占注册人数的20–33%（例如，高等教育统计局，2025年；澳大利亚教育部，2025年）；Liang 等人（2023年）发现，七种广泛使用的检测器将61.3%的非英语母语者托福作文误判为人工智能生成，这意味着相当一部分第二语言写作者所写的文本，现有检测器无法将其与人工智能输出区分开来。即使采取保守估计，不同类型大学的有效重叠比例也在5–20%之间。就全变差阈值δ而言，Sadasivan 等人（2023年）估计，使用 GPT-3 模型生成的300词段落，其人类文本与人工智能文本之间的总体全变差距离约为0.2–0.4；对于能力更强的模型和限制更多的领域，这一数值更低。由于总体全变差距离是重叠集合中学生个体全变差距离的上界，因此，对于在受限评估任务中写作的第二语言写作者，个体 δ 值低于0.1是合理的。这个示例

（π（Θ^*） $= 0.10$， $\delta = 0.05$）因此处于经验上合理的范围内，不过我们强调，这些是一致性论证，而非直接估算。图中最重要的观察在于其结构性：在参数空间中一个广阔且符合实证情况的区域内，该下界都达到不可忽略的程度，而不仅仅是在极端或人为设定的参数值下如此。

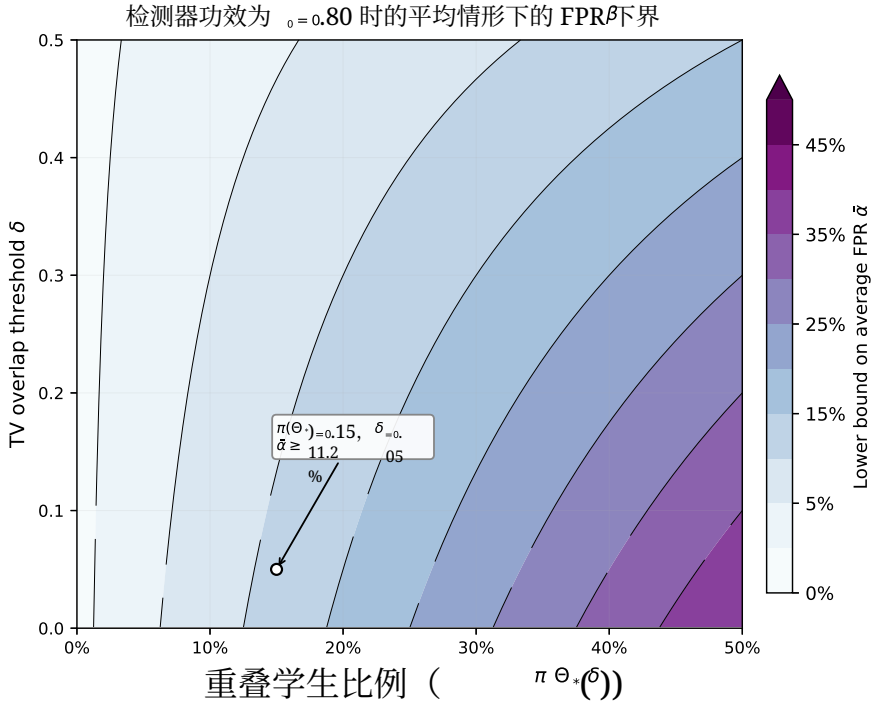


图2：结果1给出的总体平均假阳性率下界，在检测器功效为 $\beta_0 = 0.80$时，展示于（$\pi(\Theta^*(\delta))$ δ）参数空间中。请注意， $\pi(\Theta^*(\delta))$按定义随 δ 单调不减，因此，给定的学生群体会在该空间中描绘出一条单调曲线，而非任意占据其中的位置。该图展示了每一组数学上有效的参数对应的界值；具体总体和任务下实际取值如何，则是一个经验问题。第3.3.1节中的示例已标出。

3.3.2 最坏情况保护

结果 2：最坏情况规模–功效界

如果大学要求**不得有任何单个学生**面临超过 α_0的假阳性率，那么检测器的检验功效上限为

$$\beta(\phi) \;\le\; \alpha_0 + \delta^*,$$

其中， $\delta^* = \inf_\theta \mathrm{TV}(p_\theta, p_M)$是人工智能输出与**最近的**人类写作者之间的全变差距离。

与前文一样，我们再举一个概念示例来说明上述结果的含义。如果某个学生群体中最近的人类写作者与AI生成文本模型之间的全变差距离为 $\delta^* = 0.05$ ，而大学要求每位学生的假阳性率低于1%（$\alpha_0 = 0.01$），那么检测器的检验功效不能超过 $0.01 + 0.05 = 0.06$；它最多只能检出6%的人工智能生成的提交内容。这并非任何特定检测器的局限，而是问题本身在数学上的必然结果。

我们谨此说明，结果 2 是任何有效检验都必须满足的条件，而不是 Huber（1965年）和 Huber & Strassen（1973年）意义上的极小极大定理。极小极大结果还需构造最不利分布，并证明鞍点存在（参见 Faus 等人，2021年）。我们的界给出了检验大小约束下可达到的检验功效上限，但并未确定最优检验。

3.3.3 子群体混合界

重叠量 $\pi(\Theta^*)$ 和 δ^* 涉及个体学生分布与 AI生成文本模型之间的全变差距离，这些量无法直接观测。结果 3 将可解释的概念理论与可观测的 子群体层面量联系起来。

假设学生按人口统计特征或学科划分为若干子群体 G_1，……，G_K（例如按第一语言、学位课程或学习年级划分）。子群体混合分布 $\bar{p}_{G_k}$ 是指从子群体 G_k中随机选取一名学生所写文档的分布，可根据汇总提交文本进行估计。

结果 3：子群体混合界

Fo对于任意检测器 ϕ 和任意子群 G_k，子群体平均假阳性率满足 s

$$\bar{\alpha}_{G_k}(\phi) \;\geq\; \beta(\phi) - \mathrm{TV}(\bar{p}_{G_k}, p_M).$$

如果某个子群的汇总写作分布与人工智能输出接近（全变差距离 TV($\bar{p}_{G_k}$， p_M) 较小），那么任何具有实用检验功效的检测器都必然会对该子群产生较高的假阳性率。这为 Liang 等人（2023）的实证发现提供了理论解释：他们记录了非英语母语者的假阳性率偏高；Hadra 等人（2026）也发现了英语作为外语学习者的类似模式。

关键的数学步骤是注意到，在子群体混合分布下，检测器的期望响应等于 子群体平均假阳性率。对各子群进行汇总，可得全机构界

$$\text{Institution-wide average FPR} \;\geq\; \sum_{k=1}^{K} \pi_k \cdot \big(\beta(\phi) - \mathrm{TV}(\bar{p}_{G_k}, p_M)\big),$$

其中 π_k 是学生子群 G_k所占的比例。

4 讨论

4.1 检测困难的两种不同来源

本文提出的一个关键概念，是将导致人工智能检测困难的两种逻辑上相互独立的机制区分开来；这两种机制各自具有不同的政策影响。

来源 1 是 AI质量趋同：随着大语言模型不断改进，人工智能输出与人类写作之间的统计距离会对所有人类作者同时缩小。Sadasivan 等人（2023）分析的正是这一机制，可采取的干预措施包括降低人工智能输出质量、采用水印技术，或要求人工智能服务提供商嵌入可检测的信号。

来源 2 是本文讨论的群体多样性。即使人工智能模型固定且不再改进，如果学生群体足够多样，以至于部分学生的写作与人工智能输出重叠，那么任何仅基于文本的一次性检测器都会面临根本性的规模—功效权衡。机构层面的干预策略包括重新设计评估（开放式任务、个人反思）、采用过程性评估（草稿、口头答辩），以及根据多次提交文本建立学生写作档案。

这些来源互为补充且相互独立。关键在于，无法通过构建更好的检测器来解决来源 2：只要存在分布重叠，任何纯文本单次检测器都无法摆脱规模—功效权衡。这是复合原假设导致的结果，并非任何特定检测器的缺陷。要解决来源 2，必须改变评估规程，也就是收集比单份提交的文档更丰富的证据。

4.2 理论联系实践：估计与审计

结果 3 的实用性取决于能否将混合分布总变差距离 TV($\bar{p}_{G_k}$， p_M) 与可观测量联系起来。下面我们将讨论这一步的前景与重大困难，旨在为机构实践提供审慎的指导。

一种自然的方法是训练一个分类器，以区分合并的子群体文档和人工智能生成的文档。若该分类器在类别均衡的条件下达到准确率 $\hat{a}$ ，则贝叶斯最优准确率满足 $a^* =$

$\frac{1}{2}(1 + \mathrm{TV})$, so

$$\mathrm{TV}(\bar{p}_{G_k}, p_M) \;\geq\; 2\hat{a} - 1.$$

这是 TV 的下界。然而，要使结果 3 得出有关子群体假阳性率的有效下界，我们需要 TV较小 ——也就是说，我们需要 TV 的一个上界，方向恰好相反。

这意味着，如果分类器难以区分某个子群与人工智能输出（准确率较低），这表明TV可能较小，但这并非确凿证据——分类器本身也可能并非最优。具体而言，某个分类器的准确率未能超过50%，并不能证明TV较小，除非能够论证该分类器对于相关分布而言接近贝叶斯最优。反之，如果分类器能够轻易区分某个子群与人工智能，那么TV就较大，该子群的界限也就失去了实际意义；但这本身仍有参考价值，因为它指出了可能实现检测的子群。

4.3 任务依赖性

所有界限都取决于评估任务 τ。重叠参数 δ^*_τ, $\pi(\Theta^*_\tau)$ 以及 $\mathrm{TV}(\bar{p}_{G_k,\,\tau},\; p_{M,\,\tau})$ 可能因任务而有显著差异，例如格式受限任务，如“写一篇500字的文章摘要”，这类任务会降低学生答复之间的差异，同时限制人工智能输出。这两种效应都会增大人类与人工智能分布之间的重叠，使检测在结构上更加困难。

开放式或反思性任务，例如“讨论你的个人经历与这一理论有何关联”，会增加学生之间的风格变异性，但也会增大大多数学生与人工智能之间的有效全变差距离，使检测在结构上对大多数学生而言更容易（尽管未必对所有学生都如此）。

实际意义在于，即使是在同一门课程中，同一个检测器在结构上可能适用于某些任务类型，却不适用于另一些类型。这进一步支持一项审计协议建议：应按任务组评估检测可靠性，而非采用单一的机构阈值。

4.4 局限性

我们呈现的结果可作为一种解读视角，帮助我们理解和解释实证研究中观察到的结果，以及高等教育领域各处报告的轶闻和社群观察。为简化人工智能检测问题、使统计分析与考量得以开展而采用的假设是合理的，因为全球高等教育机构正在实施的一些最简单的“黑箱”人工智能文本检测器正是如此。不过，未来的分析迭代仍可能进一步改进方法论并放宽某些假设，从而将效度范围扩展至更多检测器情境。下文将简要概述未来工作中应考虑的一些潜在局限或改进方向。

1. **估计难度**。 结果 3 给出了以全变差距离 $\mathrm{TV}(\bar{p}_{G_k}, p_M)$ 表示的界。要严谨地估计这一量，需要使用近乎贝叶斯最优的分类器或进行密度估计；分类准确率只能得出 TV 的下界，而这对于使假阳性率界具有信息量而言方向恰好相反（第 4.2 节）。因此，我们建议进行直接假阳性率审计（第 4.5 节），而非总变差估计。在考试舞弊的教育测量文献中，人们早已充分认识到检测分类中真实标签难以确立这一问题（Meng & Ma，2023）：缺少经核实的标签，限制了对假阳性率和假阴性率的评估，而复合零假设框架则将这一挑战形式化，用于人工智能文本检测。

2. **单文档模型**。 我们将每份提交文本视为从 p_θ 中独立抽取的一次样本，并忽略文档内部结构。对于目前大学普遍使用的“黑箱”单次、基于文本的人工智能检测器而言，这是一个合理的假设。当然，也存在其他检测技术，例如可以依据更复杂的文档特征或学生纵向数据作出判断，但这些不在本文的讨论范围内。

3. **固定人工智能备择假设**。 此外，“人工智能生成”与“人工智能辅助的人类写作”之间并没有清晰的界限：学生在独立写作前使用人工智能系统进行头脑风暴或获取结构建议，其所处的位置是一个连续区间中的某一点，而不能明确归入任一假设。我们的二元表述（人类写作与人工智能生成）是一种简化，反映了部分大学目前所作的二元决策；但复合备择问题表明，检测问题在其边界处可能存在一定程度的定义不充分。

4. **关于“结构性障碍”主张的适用范围**。 我们证明的是，只要学生群体与人工智能输出之间存在分布重叠，任何纯文本单次检测器都无法摆脱规模—功效权衡。改变信息结构的其他机构设计——例如为每位学生收集多个样本或建立写作档案、重新设计任务，或要求提供过程证据——或许能够降低实际的分布重叠。我们的结果并未涉及这些备择方案；它们只说明，在目前许多高等教育环境中普遍采用的单次、纯文本情境下，单凭一份提交的文档能够实现什么。

4.5 一条实用途径：直接假阳性率审计

在实践中，最具操作实用性的方法是完全绕过全变差距离估计。如果一所大学将检测器应用于来自子群 G_k 的已知由人类撰写的文档语料，并观察到较高的假阳性率，这就直接证实，对于该检测器而言， $\bar{\alpha}_{G_k}$ 很大，而无需估计底层的全变差距离。Liang 等人（2023）、Weber-Wulff 等人（2023）、Elkhatat 等人（2023）和 Hadra 等人（2026）采用的正是这种方法，而结果 3 给出了理论解释：当子群体混合分布接近 p_M时，观察到较高的假阳性率是规模—功效权衡所导致的必然结果。分层审计方法符合教育测量领域长期确立的自动评分评估框架；这些框架要求在实际部署前提供证据，证明各子群的表现一致（Williamson 等人，2012年）。

拟议的机构审计规程

在部署任何人工智能文本检测器之前，机构应：

1. **组建分层人工撰写语料库**。收集经确认由人类撰写的文档，并按照机构中可能与人工智能输出呈现不同重叠程度的常见学生群体所对应的可观测子群进行分层：第一语言（第一语言与第二语言使用者）、不同的学位课程、学习年级、评估任务类型等。

2. **在每个分层上运行检测器**。对每个子群语料库应用候选检测器，并记录每个子群 G_k 的假阳性率 $\hat{\bar{\alpha}}_{G_k}$。

3. **报告分层假阳性率和检验功效**。对于每个子群，报告观测到的假阳性率以及检测器估计的检验功效（根据针对相同任务类型的人工智能生成文本平行语料库得出）。结果 3 预测，重叠程度较高的子群通常预期会表现出较高的假阳性率；如果观测到的权衡关系过于严重，则该检测器在该学生群体与评估任务类型组合上存在结构性不足。

4. **按任务进行评估**。针对每种评估任务类型重复步骤 1–3。与开放式任务或个人反思任务相比，受限任务（例如“用 500 字总结这篇文章”）可能产生更高的重叠程度（第 4.3 节）。

5. **按分层设定部署阈值，而非全机构统一设定**。 如果检测器通过了某些子群/任务组合的审计，但未通过其他组合的审计，就应据此限制部署。不要采用单一的机构级检测阈值，因为这会对规模较大的学生子群造成不公平的偏差。

该方案检验结果 3 的经验性预测，无需估算抽象的全变差距离（第 4.2 节）。它还会生成公平导向政策所需的各子群证据基础，正如 Hadra 等人（2026）和 Liang 等人（2023）所建议的那样。

4.6 政策与实践的启示

我们的研究结果凸显了人工智能检测技术研究及实际应用中已逐渐形成的若干观点。下面，我们将重点阐述当前研究对大学应如何

开展人工智能文本检测的若干具体启示。

尽管 AI 检测因学术诚信方面的顾虑而便捷且得到广泛采用（Luo，2024），但检测分数不应作为唯一证据。结果 1 和 2 表明，任何具有实用检测功效的纯文本单次检测器，都不可避免地会错误指控写作内容与人工智能输出相重叠的学生。因此，在学术不端程序中将检测分数作为唯一或主要证据，会使机构面临在数学上无法避免的错误判定风险。这为 Weber-Wulff 等人（2023）、Elkhatat 等人（2023）和 Hadra 等人（2026）基于实证依据提出的政策建议提供了严谨的理论基础。

群体多样性是一项独立于检测器质量的限制，无法通过提升检测器质量来消除。对于学生群体多样的机构，例如拥有英语二语使用者、来自不同学科传统的学生，或写作经验各异的学生的机构，群体多样性带来的结构性障碍可能超过人工智能质量趋同带来的障碍。在纯文本单次设定下，投资开发更好的检测器无法消除这一结构性障碍。鉴于此，值得考虑加大对评估设计的投入，这或许会有所裨益。

分层审计应成为标准做法。在部署任何检测器之前，机构应开展第 4.5 节所述的分层假阳性率审计。这既能利用现有资源切实可行地完成，也能直接说明该检测器的错误特征对于预期的常见子群和任务类型是否可以接受。

重新设计评估是一种持久的解决方案，还可能带来额外价值，并有助于确保评估目标的效度（Dawson 等人，2024；Corbin 等人，2025）。鉴于群体多样性是复合原假设的属性，而非任何检测器的属性，机构有效的应对方式是改变信息结构：采用过程性评估、加入口试环节、设置结合个人经历与已形成理解的反思性任务，并建立纵向写作档案，从而将复合原假设转变为一个（部分）已知的原假设。

5 结论

我们将大学中的“黑箱”单次文本AI检测问题形式化为复合假设检验，并基于数理统计的基本原理提出三个定量结果：平均情形与最坏情形下的权衡界（结果 1-2），以及一个将理论框架与可观测的子群层面量相联系的子群体混合界（结果 3）。核心观点是，群体多样性会对纯文本单次AI检测形成一种结构性障碍，而这一障碍在逻辑上独立于人工智能模型质量：只要学生群体与人工智能输出之间存在分布重叠，任何纯文本单次检测器都无法摆脱检验大小与功效之间的权衡。

这一结构性障碍为不断增长的关于人工智能检测工具的实证文献中所记录的不可靠性与偏差模式提供了理论解释（Weber-Wulff 等人，2023年；Liang等人，2023年；Elkhatat 等人，2023年；Hadra 等人，2026年）。这也阐明了可能改进的范围：更好的检测器工程无法消除因受测人群的多样性而产生的障碍。持久有效的机构应对之策是重新设计评估，改变信息结构，使决策不必仅仅依据单份提交的文档。

参考文献

澳大利亚教育部（2025）。高等教育精选统计数据——*2024*年学生数据。https://www.education.gov.au/higher-education-statistics/student-data/selected-higher-education-statistics-2024-student-data。访问于2026年 Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F.（2024）。观点：关于人工智能生成文本检测的可能性。第*41*届国际机器学习会议论文集，ICML’24。Corbin, T., Dawson, P., & Liu, D.（2025）。空谈无益：为什么生成式人工智能时代需要结构性评估变革。《高等教育评估与评价》，50(7)，1087–1097。https://doi.org/10. 1080/02602938.2025.2503964

Dawson, P., Bearman, M., Dollinger, M., & Boud, D.（2024）。效度比作弊更重要。《高等教育评估与评价》，49(7)，1005–1016。https://doi.org/10.1080/02602938.2024. 2386662 Dawson, P., Sutherland-Smith, W., & Ricksen, M.（2019）。软件能否提高评分者检测合同作弊的准确性？一项关于 Turnitin Authorship Investigate Alpha 的试点研究。《高等教育评估与评价》，45(4)，473–482。https://doi.org/10.1080/02602938.2019.1662884 Elkhatat, A. M., Elsaid, K., & Almeer, S.（2023）。评估人工智能内容检测工具区分人类撰写文本与人工智能生成文本的效能。《国际教育诚信期刊》，19(1)。https://doi.org/10.1007/s40979-023-00140-5 Faus, M., Zoubir, A., & Poor, H. V.（2021）。极小极大鲁棒检测：经典成果与最新进展。《*IEEE*信号处理汇刊》，69，2252–2283。https://doi.org/10.1109/tsp.2021.3061298 Hadra, M., Cambridge, K., & Mesbah, M.（2026）。评估学术情境中人工智能内容检测器的准确性与可靠性。《国际教育诚信期刊》，22(1)。https://doi.org/10.1007/s40979-026-00213-1 Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J.（2024）。变革评估：大语言模型与生成式人工智能的影响及其意义。《教育测量：问题与实践》，43(2)，16–29。https://doi.org/10.1111/emip.12602 高等教育统计局（2025）。高等教育学生统计：英国，*2023/24*。https: // www.hesa.ac.uk/news/20-03-2025/sb271-higher-education-student-statistics/location。访问于2026年 Huber, P. J.（1965）。概率比检验的一个鲁棒版本。《数学统计年刊》，36(6)，1753–1758。https://doi.org/10.1214/aoms/1177699803 Huber, P. J. & Strassen, V.（1973）。容量下的极小极大检验与奈曼–皮尔逊引理。《统计年刊》，1(2)。https://doi.org/10.1214/aos/1176342363 Koppel, M. & Schler, J.（2004）。将作者身份验证作为单类分类问题。第二十一届国际机器学习会议——*ICML '04*，ICML '04，62。https://doi.org/10.1145/ 1015330.1015448 Lehmann, E. & Romano, J. P.（2022）。《统计假设检验》。施普林格国际出版。https://doi.org/10.1007/978-3-030-70578-7 Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J.（2023）。GPT检测器对非英语母语写作者存在偏差。《*Patterns*》，4(7)，100779。https://doi.org/10.1016/j.patter.2023.100779 Luo, J.（2024）。高等教育评估中生成式人工智能政策的批判性综述：重新思考学生作品"原创性"的呼吁。《高等教育评估与评价》，49(5)，651–664。https: //doi.org/10.1080/02602938.2024.2309963 Meng, H. & Ma, Y.（2023）。考试作弊检测中的基于机器学习的画像分析。《教育测量：问题与实践》，42(1)，59–75。https://doi.org/10.1111/emip.12541 Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S.（2023）。人工智能生成文本能否被可靠地检测出来？*arXiv*预印本 *arXiv:2303.11156*。https://doi.org/10.48550/arXiv.2303.11156。第4版，2025年 Stamatatos, E.（2008）。现代作者归属判定方法综述。《美国信息科学与技术学会会刊》，60(3)，538–556。https://doi.org/10.1002/asi.21001 Tight, M.（2024）。应对高等教育中的作弊：研究与实践综述。《高等教育评估与评价》，49(7)，911–923。https://doi.org/10.1080/02602938.2023.2300104

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L.（2023年）。人工智能生成文本检测工具测试。《国际教育诚信期刊》，19(1)。https://doi.org/10.1007/s40979-023-00146-z

Williamson, D. M., Xi, X., & Breyer, F. J.（2012年）。自动评分的评估与使用框架。《教育测量：问题与实践》，31(1)，2–13。https://doi.org/10.1111/j.1745-3992. 2011.00223.x

A 附录：形式化陈述与证明

为求完整，我们将定理形式化陈述并给出证明。所有证明仅使用变分不等式（2）。

定理 1（平均情形权衡）。 设 $\{p_\theta\}_{\theta\in\Theta}$ 为一个分布族， p_M 为一个固定分布，π 为定义在 Θ上的概率测度。对于 $\delta > 0$，令 $\Theta^* = \{\theta:$ 全变差距离$(p_\theta, p_M) \le \delta\}$ 且 $\pi(\Theta^*) > 0$。那么对于任意$\phi:$ $\mathcal{X} \to [0,1]$:

$$\int_{\Theta} \alpha(\theta;\phi)\, d\pi(\theta) \;\ge\; \pi(\Theta^*)\cdot\big(\beta(\phi) - \delta\big). \tag{3}$$

证明。 对于 $\theta \in \Theta^*$，由 (2) 得： $\alpha(\theta;\ \phi) = \mathbb{E}_{p_\theta}[\phi] \ge \mathbb{E}_{p_M}[\phi] -$全变差距离$(p_\theta,\ p_M) \ge \beta(\phi) - \delta$。对 Θ 积分并限制到 Θ^*： $\int_\Theta \alpha\, d\pi \ge \int_{\Theta^*}(\beta - \delta)\, d\pi = \pi(\Theta^*)(\beta - \delta)$。 □

定理 2（最坏情况规模–功效界）。 在相同设定下，若 上确界$_\theta\, \alpha(\theta;\ \phi) \le \alpha_0$，则：

$$\beta(\phi) \;\le\; \alpha_0 + \delta^*, \quad \delta^* = \inf_{\theta} \mathrm{TV}(p_\theta, p_M). \tag{4}$$

证明。 令 θ^*达到（或趋近于）下确界。根据 (2)： $\beta(\phi) - \alpha(\theta^*;\ \phi) \le \delta^*$。因此 $\beta(\phi) \le \alpha(\theta^*;\ \phi) + \delta^* \le \alpha_0 + \delta^*$。 □

定理 3 (子群体混合界**)**。 对于具有混合分布 $\bar{p}_{G_k} = \int_{G_k} p_\theta\, d\pi(\theta|G_k)$ 且子群平均假阳性率为 $\bar{\alpha}_{G_k}(\phi) = \int_{G_k} \alpha(\theta;\ \phi)\, d\pi(\theta|G_k)$的子群 G_k：

$$\bar{\alpha}_{G_k}(\phi) \;\ge\; \beta(\phi) - \mathrm{TV}(\bar{p}_{G_k}, p_M). \tag{5}$$

证明。 根据定义： $\mathbb{E}_{\bar{p}_{G_k}}[\phi] = \int_{G_k} \mathbb{E}_{p_\theta}[\phi]\, d\pi(\theta|G_k) = \bar{\alpha}_{G_k}(\phi)$。将 (2) 应用于 $(\bar{p}_{G_k}, p_M)$： $\beta(\phi) - \bar{\alpha}_{G_k}(\phi) \le$ 全变差距离$(\bar{p}_{G_k}, p_M)$。整理后即得结论。 □

注释 1（凸性）。 根据全变差距离的联合凸性：全变差距离$(\bar{p}_{G_k}, p_M) \le \int_{G_k}$全变差距离$(p_\theta, p_M)\, d\pi\ (\theta|G_k)$。混合总变差距离是子群内学生平均总变差距离的下界。较小的混合总变差距离并不意味着所有学生分别都接近人工智能。