

# A letter to schools regarding AI Detectors

[in.linkedin.com/pulse/letter-schools-regarding-ai-detectors-dan-bowen-tqggc](https://www.linkedin.com/pulse/letter-schools-regarding-ai-detectors-dan-bowen-tqggc)



Image by Bing Image Creator. A boy with VR headset interacting with 3D model



**Dan Bowen**

Technology Strategist @ Microsoft | Artificial Intelligence, Azure, M365, Security |  
Reimagining Education | School Improvement | Always Learning | Speaker | Luthier | AI  
podcast | Mental Health Advocate

December 20, 2023

Dear sir/madam,




I would like to clarify a few things regarding the questioning of academic integrity of students 'caught' by AI detectors. To, effectively, label students as dishonest and not have academic integrity, is something that needs to be addressed for a variety of reasons that I will outline below:

**First and foremost, AI plagiarism checkers of any kind do not work effectively.** Whether a student use AI or not, these detectors have been proven to inaccurately detect content that is in fact human-generated. Turnitin for example, themselves, have actually admitted that data they provided over the accuracy of their own product in detecting AI content should not be relied upon. Interestingly they will not disclose what the amended accuracy rate is: <https://www.turnitin.com/blog/ai-writing-detection-update-from-turnitins-chief-product-officer>. Also on this issue, I urge you to read this academic research on Testing of Detection Tools for AI-Generated Text:

"This paper exposes serious limitations of the state-of-the-art AI-generated text detection tools and their unsuitability for use as evidence of academic misconduct. Our findings do not confirm the claims presented by the systems. They too often present false positives and false negatives." ( [Testing of Detection Tools for AI-Generated Text](#) )



I would also like to share this with you as an example of how a university has communicated with a newsletter announcement that their Teaching Center doesn't endorse any generative AI detection tools: "the Teaching Center will disable the AI detection tool .... effective immediately" ( [Teaching Center doesn't endorse any generative AI detection tools](#) )

You should also read some of the academic research by Associate Professor Ethan Mollick, Stefen Bauschard and others who are worth following in this area (some example posts below):

 **Ethan Mollick**  · 1st  
Associate Professor at The Wharton School  
2d · 

Reminder that there is no current way (and is unlikely to be a future way) to detect AI-generated content. And AI detectors all have high false positive rates.

One thing I have seen teachers do is ask ChatGPT whether AI wrote something. That doesn't work: "ChatGPT tended to classify text generated by LLMs as if it were written by humans, with a misclassification probability of about 50%. While GPT-4 leaned towards labeling human-written text as if it were generated by LLMs, and about 95% of human-written texts are misclassified as LLM-generated texts"

 **Stefan Bauschard** (He/Him) · 1st  
Natural language conversationalist; Cofounder, educating4ai.com; Ow...  
2d · 

There is another paper out today highlighting the limitations of AI-writing detectors.

**Randi Weingarten Ethan Mollick Jason Gulya Dr. Sabba Quidwai**

Full paper --  
<https://lnkd.in/e8VA444R>

<p><b>II. IMPORTANT ISSUES OF LLM-GENERATED TEXT DETECTION</b></p> <p>In this section, we discuss the main issues and limitations of contemporary state-of-the-art techniques designed for detecting text generated by LLMs. It is important to note that this technique has been acknowledged as infallible. The issues outlined herein may pertain specifically to one or multiple types of detectors.</p> <p><i>Out of Distribution Challenges</i></p> <p>Out-of-distribution issues significantly impede the efficacy of</p>	<p>we emphatically draw attention to these issues, as addressing them is pivotal for the accuracy and fairness of detectors for LLM-generated text. Recent research [149] revealed a significant performance drop for state-of-the-art detectors on texts authored by non-native English speakers. Employing effective prompt strategies can inadvertently allow the generated text to evade detection. Consequently, there is a risk of inadvertently penalizing writers who exhibit diverse writing styles or employ limited expressions, which hinders the effectiveness of discrimination within the detection process.</p> <p>c) <i>Prompt Attacks</i>: Prompt attacks pose a significant challenge for current LLM-generated text detectors. The quality of LLM-generated text varies based on the complexity of the prompts that guide the model. As the model size and corpus size increase, LLMs exhibit excellent ICL capabilities for more complex prompts. The large number of effective prompts elicited, such as few-shot [50], combination of few-shot [55], and zero-shot CoT [161], etc., which enhances the quality and capability of LLMs. Existing LLM-generated text detectors mainly use simple prompts, such as the work of [73], which directly leads to the fact that detectors struggle to detect text generated with complex prompts. Liu et al. [73] reported a decrease in the detection ability of a detector when it is confronted with a different prompt. This is a common trend in text classification tasks.</p>
--	---

And this from Joanne Villis as another good example

**EDUCATING GIRLS**

**Joanne Willis**  
 Director Technology Enrichment|  
 Masters Digital Technologies|  
 Nationally Certified Lead Teacher|  
 Stage 2 Design, Technology and  
 Engineering Teacher

[View full profile](#)

**Joanne Willis** · 1st  
 Director Technology Enrichment| Masters Di...  
 1d · Edited · 🌐

We should have moved on from this, but there are still many teachers using AI text detectors to primarily catch students 'cheating'. Teachers who are using AI text detectors, such as GPTZero, ZeroGPT, Sapling, Copyleaks, and Turnitin, to identify potential cases of plagiarism need to read this. 'AI Text Detectors' designed by Torrey Trust (2023), University of Massachusetts Amherst. Points below in quotation marks have been directly copied from the text.

"Many of the AI plagiarism detectors are not fully transparent about how their tool works. For instance, Turnitin's tool runs text against their "AI detection model" but does not describe how that model was designed or works".

These detectors are neither accurate nor reliable. "This feature was enabled for Turnitin customers with less than 24-hour advance notice, no option at the time to disable the feature, and, most importantly, no insight into how it works. At the time of launch, Turnitin claimed that its detection tool had a 1% false positive rate (Chechitelli, 2023). To put that into context, Vanderbilt submitted 75,000 papers to Turnitin in 2022. If this AI detection tool was available then, around 750 student papers could have been incorrectly labelled as having some of it written by AI" (<https://lnkd.in/gXJji73j>).

"The use of these tools to evaluate student text can increase students' anxiety and stress (both of which have been found to inhibit learning), while also creating an atmosphere of distrust".

**Secondly**, even if you, your faculty or the school insist on using these unreliable tools for such purposes, sending an email to parents and students, despite their politeness, says a student has cheated, is not an approach that is appropriate or the best option. Even though you may not have known about AI detector deficiencies, you are surely aware that AI use is by no means a binary or easily definable issue. Grammarly uses AI, does that mean the students in your class working with that tool are also plagiarising? Which then offers the following points of contention:

1. How does the school ensure that tutors, older siblings and parents have zero contribution to the same assessment task?
2. What exactly does flagged by an AI detector mean? Does the department or school have a threshold or certainty of the level of AI usage?
3. Do you have a school policy that states us of AI detectors and their thresholds? If there is none then students surely cannot be accused of something out of scope of this policy.

I appreciate that AI is causing major disruption in education and the subjects like English and HSIE are feeling the brunt of this. I recognise that there is an extra workload and pressures on teachers in these disciplines. However, if you are to persist with these ineffective tools and/or in dealing with situations where plagiarism is suspected, perhaps a different approach is needed. Possibly an approach of waiting until you speak to the student and have a conversation about why you are querying their work, i.e. a sudden change in writing style, language, etc then proceed to ask them questions that gives the student the chance to demonstrate understanding of what they had written and how they may have used AI to support them if it was even used at all.

**Finally**, I do understand that schools are trying to deal with this new age of AI but there are a lot of guidelines stating that this is not the way to deal with it. There is a new generative AI in schools framework that will help you from ministers that illustrates that accusing students is just unjust. It states that members of the school community that are impacted by AI tools must be **actively informed about and have the opportunity to question the use of the outputs of the tools you are using**, for example.

#### 5.4 Contestability

Individuals (e.g., students, parents, staff) that are significantly impacted by a generative AI tool are able to challenge the use or outputs of the tool, and any decisions informed by the tool.

This principle aims to uphold best practice in the use of generative AI by allowing humans to intervene when the products or outputs, including decisions, may be inaccurate or wrong.

Generative AI can produce results or outputs that are inaccurate or wrong. Fairness, transparency and accountability in using generative AI cannot be achieved without ensuring that individuals have visibility and contestability. This includes developing appropriate processes to allow all members of the school community to contest the use and outputs of generative AI where possible.

This is important even when generative AI systems are performing as expected.

Students and Parents are able to challenge any decisions informed by AI tools

I think this is an **opportunity** for you to **embrace the positive use of the new AI tools in school**, within given parameters, and use them as an opportunity to rethink the way you **assess and guide students** to understand your subject but also realise that they can get support from tutors, parents, as well as AI. Please do not rely on AI tools to 'detect' academic integrity. Hopefully I have illustrated their inaccuracy with the research above.

There are some great teachers out there who are talking about the use of AI in History such as Matt Esterman, as well as AI and student agency Dr Nick Jackson.

The links below provide you also with some commentary on these issues:

<https://preview.mailerlite.io/preview/282063/emails/85539571521029259>

<https://www.oneusefulthing.org/p/the-homework-apocalypse>

Any letters of concern can cause considerable upset. Generic accusations based on AI detectors are serious and the tools you are using to make such decisions are shown to be highly unreliable.

As a parent, an ex-teacher and a person who has been involved in education all my career, I will always support the teachers and school. However you should really take the time to embrace this technology and use it as a way to re-engage, enthuse, and even help save some time planning. This technology can be used to inspire, and augment learning and bring new experiences to life as well as get rid of the traditional, mundane and ineffective processes that add to teachers workload.

## Published by

---



Dan BowenDan Bowen

Technology Strategist @ Microsoft | Artificial Intelligence, Azure, M365, Security | Reimagining Education | School Improvement | Always Learning | Speaker | Luthier | AI podcast | Mental Health Advocate  
Technology Strategist @ Microsoft | Artificial Intelligence, Azure, M365, Security | Reimagining Education | School Improvement | Always Learning | Speaker | Luthier | AI podcast | Mental Health Advocate

---