STADIO

AI detection tools – assessing their limitations and declaring them an academic *cul de sac*. *Quo vadis?*

Presented by: Jacques Nieuwoudt

26 September 2024

INTRODUCTION AND PRESENTATION OUTLAY

STADIO

1 Introduction and reason for the presentation

2 *Disclaimers and Caveat*

3 Accuracy of AI detectors

4 Adversarial techniques to reduce AI scores

5 Human / Lecturer detection

6 Digital divide and equity issues

7 *Quo vadis?* / Way forward

INTRODUCTION, DISCLAIMERS AND *CAVEAT*

1 Increased use of generative AI tools in assessments has raised concerns regarding the academic integrity and authenticity of students' work.

2 No denying possible benefit of AI generating tools in academia and need to educate students on the ethical use of AI

3 This in turn has led to the use of AI detection tools by educators to detect AI generated text in assessment answers to penalise students for the use of AI in assessments. Police-and-punish principle.

4 *(My perception that police and punish i.t.o. Plagiarism policy and Turnitin similarities did not decrease prevalence of plagiarism)*

5 Large language models (LLMs) develop at an incredible rapid pace – becomes more sophisticated and less likely to be detected.

6 AI detection tools must try and keep up. Arms race between generation and detection.

7 My presentation viewed with a law qualification lense and admit that experiences will differ from qualification to qualification

INTRODUCTION, DISCLAIMERS AND *CAVEAT*

1 AI detection measures common outputs and there are no means to verify results

2 Becomes problematic when we allow students to use AI for 'legitimate uses' such as reviewing, paraphrasing, enhancing expression

3 AI detection tools cannot differentiate between GOOD or BAD use of AI, i.e. what is permissible and what is impermissible

4 Not every student knew or used AI

5 Not every student is a cheat and intentionally academically dishonest.

6 Students cheated in the past and got away with it, and that trend will continue despite our best efforts – look at time-benefit analyses (Old problem on a new scale: plagiarism, cheating, paper/essay mills, paying someone to do your work, e.g. Golden Ticket)

7 With fast changing AI environment, the data presented might already be outdated, but at least gives a snapshot overview

ACCURACY OF AI DETECTION TOOLS

1 BASELINE TESTING

15 AI generated samples and 10 human samples – establish ability of AI detection tools to determine authorship

AI DETECTOR	ACCURACY
Copyleaks	64%
Turnitin	61%
Crossplag	60.8%
GPT-2 detector	57.2%
ZeroGPT	46%
GPTKit	37%
GPTZero	23.3%
GENAI TOOL	
Bard	76%
GPT-4	23.9%
Claude 2	17.7%
MEAN accuracy	39.5%

Perkins M et al GenAi Detection tools, adversarial techniques and implications for inclusivity in Higher Education (March 2024)

ADVERSERIAL TECHNIQUES TO LOWER AI DETECTION

AI DETECTOR	ACCURACY	Accuracy manipulated	% drop in accuracy
Copyleaks	73.9%	58.7%	15,2%
Crossplag	54.3%	32.4%	21.9%
GPT – output	34.7%	17.5%	17.2%
ZeroGPT	31.3%	17.3%	14%
Turnitin	50%	7.9%	42.1%
GPT Kit	6%	4.5%	1.5%
AVERAGE	41.7%	23%	

Perkins M et al GenAi Detection tools, adversarial techniques and implications for inclusivity in Higher Education (March 2024)

EVALUATION OF SPECIFIC ADVERSARIAL TECHNIQUES

ADVERSARIAL TECHNIQUE	ACCURACY	% DROP IN ACCURACY
Add spelling errors	12.9%	27%
Increase burstiness	15.9%	24%
Paraphrase	18.4%	21%
Decrease complexity	21%	19%
Write as a non-English speaking person	27.7%	12%
Increase complexity	37%	2%
MEAN	22.1%	17.5%

Perkins M et al GenAi Detection tools, adversarial techniques and implications for inclusivity in Higher Education (March 2024)

HUMAN/LECTURER AI DETECTION

Reviewers requested to identify AI generated and original text

Reviewers had an average of 98% average score on discerning AI generated articles

Misclassified 12% of original human written articles as AI generated articles

COMMON REASONS FOR AI CLASSIFICATION

1. Incoherence (34%)

2. Grammatical errors (20%)

3. Insufficient evidence-based claims (16%)

4. Vocabulary diversity (12%)

5. Creativity (6%)

6. Misuse of abbreviations (6%)

7. Writing style (3%)

8. Vague expressions (2%)

9. Conflicting data (1%)

10. Superficial discussion

11.Hallucinations /fabricated data

12.Inappropriate use of technical terms

Interesting enough, the marks allocated by reviewers to both the AI generated assignment and human generated assignment, scored an average of 55% *(In my personal assessment practices, students either scored the same or less with the use of AI)*

DIGITAL DIVIDE AND EQUALITY ISSUES

Four students need to complete the same assignment

No generative AI tools may be used

AI Detection tools will be utilised to check integrity and use of AI

Task must be completed on own time outside of the university/classroom environment

STUDENT 1: AMY

Is a student in the rural area with limited access to digital technologies at her home.

She is reliant on the university’s computers and data networks

STADIO has blocked direct access to AI tools

Amy wants to access GPT, Gemini, and Co-pilot, but it is blocked, and she has to use the free credits of a third-party application built on top of GPT-3.5.

She is further limited to completing the assignment during her time on campus before returning home.

STUDENT 2: BARRY

Barry is an English additional language (EAL) student from a migrant family where English is not spoken in the home.

Barry uses the free version of ChatGPT as a translation tool.

He uses ChatGPT to translate both the assignment questions and his answers.

STUDENT 3: CHARLIE

Charlie is a student from a low socio-economic background with low levels of literacy in the home and limited digital literacy.

Charlie uses Microsoft Copilot at home on his smartphone to understand the requirements of the tasks of the assignment and to set his work in a more academic written style.

STUDENT 4: DORY

Dory is an English first language speaker from a wealthy household with well-educated and trained parents.

Dory compiles her assignment using her parent’s access to Claude (Opus), with a monthly cost of R400 ($20) per month subscription.

Dory simply requests AI to generate the answers and copy and paste the answers into GPT-4 (another subscription-based model) and then back again into Claude with the instruction to make it a little bit more sophisticated, a little bit more varied, and to incorporate some direct quotes from the materials from class that she uploads as a PDF (a capability only available in paid models). Dory’s assignment is comprehensive, accurate, and sophisticated, but entirely compiled by GenAI.

AI DETECTION EVALUATION

The university's AI detection tool presents the following AI evaluation scores:

(All four students transgressed the requirements of the assignment by using generative AI)

1. Amy's work as 90% AI-generated *(rural area student)*
2. Barry's as 100% AI-generated *(translating questions and answers)*
3. Charlie's as 85% AI-generated *(understanding and academic style writing)*
4. Dory's as 20% AI-generated *(the full monty)*

Barry and Charlie attempted to use generative AI as a supporting tool to assist with understanding the tasks and to formulate and plan their answers. Charlie's use was perhaps a little bit more loaded.

Dory used the GenAI the most and was also deliberate in transgressing the requirements of AI use. She is also the one that is least likely to be caught for unauthorised AI work.

In summarising the digital divide and inequality; Dory is the student who was already advantaged by the education system, advantaged by her socio-economic status, and now advantaged by a heavy-handed approach to using AI and the detection of AI.

INTERNATIONAL HEI RESPONSE TO USE OF AI DETECTION TOOLS

3 types of response:

BANNED

The institution has fully pprohibited the use of AI detection software. Educators are not allowed to use it

RECOMMENDED AGAINST & DON'T PROVIDE

The institution discourages the use of AI detection software and does not provide it. Educators are advised to state the usage in their syllabi, so students should check course material

TURNITIN AI DETECTION DISABLED

These institutions have disabled Turnitin's AI detection feature but have not made public statements on their overall use.

AI DETECTION SOFTWARE BANNED

USA				
American University	Boston University	UC Berkley	Colorado State	DePaul University
Georgetown University	Indiana University	Michigan State University	MIT	Montclair State University
New York University	Northwestern	Oregon State University	Rochester Institute of Technology	San Francisco State University
SMU	Saint Joseph University	Syracuse	The University of Alabama	University of California –Irving
University of California, Los Angeles (UCLA)	University of Michigan – Dearborn	University of South Maine	University of Washington	Western University
West Chester University	Vanderbilt	Yale	University of Maryland	University of Pittsburgh
Baylor University				
CANADA	The University of British Columbia	University of Toronto		
Australia	Charles Strut University	Deakin University		
UK	University of Dundee	University of Manchester	University of Portsmouth	University of South Wales

RECOMMENDED AGAINST AND DON’T PROVIDE AI DETECTION

USA	
University of Missouri	University of Notre Dame
University of Texas at Austin	University of Central Florida
Arizona State University	
UK	
Newcastle University	University of Glasgow
University of Nottingham	

TURNITIN DISABLED AND NO PUBLIC STATEMENT

Australia	
Australian National University	Mcquarie University
University of Camberra	University of South Australia
UK	
University of Edinburgh	University of Greenwich
Canada	
Simon Fraser University	

CONCLUSION

1 AI detection tools as a stand-alone detection of GenAI in academic work is a *cul de sac* or a dead end street, not accurate, can be manipulated and does not keep up with AI generating tools

2 AI detection tools increase the digital divide and keeps or increases inequalities…..the haves are unfairly advantages over the have-nots.

3 More research, investigation and policies are necessary (*first real semester to evaluate ethical use of AI tools)*

4 *Lecturers should not doubt their ability to correctly identify authentic work based on their professional expertise and subject field knowledge*

QUO VADIS? / WAY FORWARD

1 Only way to verify authenticity and credibility of text is to check its sources, references and context, while employing critical thinking and common sense

2 Ethical AI use policy should address assessment setting, student use of AI, how assessments will be evaluated, and remedial measures. Include declaration of ethical AI use. Refer to international best practices

3 Reviewers/lecturers should be aware of working of AI tools and limitations of AI detection tools.

4 Reviewers/lecturers should trust their professional acumen in evaluating work

5 Focus should perhaps shift to capacitating students with world of work tools/ student attributes rather than pure knowledge dissemination. Why go to a HEI?

6 Assessment setting practices should be reviewed, include class work, analysis, evaluation and application that will be unique to classroom setting.

7 Closed book venue-based exams will ensure authenticity and quality of STADIO summative assessments and distinguish STADIO from other institutions = Quality Certificates/Diplomas/Degrees actually carrying weight in the industry and world of work

NETFLIX RESEARCH: THE TV SERIES TULSA KING

Sylvester Stallone as Dwight “The General” Manfredi: “Do you think anyone really gives a s**t about what your major is? English literature, biology, whatever. The whole point of a college degree is to show a potential employer that you showed up someplace four years in a row, completed a series of tasks reasonably well, and on time. So if he hires you, there’s a semi-decent chance that you will show up there every day and not f**k his business up.

THANK YOU
ENKOSI
RE A LEBOGA
DANKIE

STADIO

FORMERLY

EMBURY SBS LISOF PRESTIGE ACADEMY