Testing the AI detectors 2

Introduction

Further to the post of 26/09/2023, Robin Crockett (Academic Integrity Lead – University of Northampton) has added to his small-scale study by adding a second AI 'writer' and a third AI-text detector.

Thus, the study has now used both OpenAI's ChatGPT (GPT-3.5) and Anthropic's Claude-2 AI text generators (AI 'writers') to each produce 25 nominal 1,000-word essays: five subjects, five different versions of each subject, consistent across both 'writers'. For each subject, both 'writers' were prompted to vary the sentence type as follows: 'default' (i.e. no instruction regarding sentence type), 'use long sentences', 'use short sentences', 'use complex sentences', 'use simple sentences'. In all cases, the first essay generated by the AI in response to the prompt was accepted: there was no re-prompting of the AIs to obtain different 'better' essays.

Both sets of essays were tested/investigated using the three AI-detectors: Copyleaks, GPTZero and Turnitin, and the tables on the following two pages show the results. These are AI-detectors, therefore high percentages are true positives, i.e. AI-text classified as AI, and low percentages are false negatives, i.e. AI-text classified as not-AI (effectively classified as human-written). Note that as well as (probable) differences in programming, the detectors report their classifications differently, and the tables summarise those classifications as follows:

- Copyleaks highlights sections of text it determines as AI-generated, and for each highlighted section states the calculated probability of AI generation. The tables show the percentage of text classified as AI-generated and the associated probabilities.
- GPTZero works sentence-wise and states a calculated overall probability that the whole text is AI-generated. The tables show the percentage of sentences classified as AI-generated and the overall probability that the whole essay is AI-generated.
- Turnitin highlights and states the percentage of the 'qualifying' text it determines is AI-generated. The tables show the percentage of text classified as AI-generated, and an 'X' in the table indicates no result due to 'non-qualifying' text. See here https://help.turnitin.com/ai-writing-detection.htm.

The results are followed by some brief observations.

Results: ChatGPT generated essays.

Variant	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Default	100% text AI p(AI) = 80.6%	100% text AI p(AI) = 83.5%	100% text AI p(AI) = 88.5%	100% text AI p(AI) = 81.3%	100% text AI p(AI) = 85.4%
Long	ca. 80% text AI p(AI) = 65-75%	100% text AI p(AI) = 81.5%	ca. 95% text AI p(AI) = 75-85%	100% text AI p(AI) = 79.1%	100% text AI p(AI) = 80.6%
Short	ca. 70% text AI p(AI) = 66-72%	100% text AI p(AI) = 76.9%	100% text AI p(AI) = 87.3%	ca. 85% text AI p(AI) = 77-79%	100% text AI p(AI) = 78.4%
Complex	100% text AI p(AI) = 72.9%	100% text AI p(AI) = 81.0%	ca. 90% text AI p(AI) = 62-73%	100% text AI p(AI) = 77.7%	0%
Simple	100% text AI p(AI) = 83.6%	ca. 90% text AI p(AI) = 73-81%	100% text AI p(AI) = 95.2%	ca. 90% text AI p(AI) = 76-82%	100% text AI p(AI) = 84.9%

Copyleaks results for ChatGPT generated essays.

Variant	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Default	80% sents AI p(o'all) = 55%	76% sents AI p(o'all) = 52%	52% sents AI p(o'all) = 65%	26% sents AI p(o'all) = 52%	56% sents AI p(o'all) = 64%
Long	54% sents AI p(o'all) = 50%	48% sents AI p(o'all) = 52%	100% sents AI p(o'all) = 97%	44% sents AI p(o'all) = 50%	58% sents AI p(o'all) = 55%
Short	35% sents AI p(o'all) = 45%	62% sents AI p(o'all) = 49%	71% sents AI p(o'all) = 51%	17% sents AI p(o'all) = 28%	38% sents AI p(o'all) = 45%
Complex	82% sents AI p(o'all) = 56%	63% sents AI p(o'all) = 53%	56% sents AI p(o'all) = 70%	38% sents AI p(o'all) = 48%	29% sents AI p(o'all) = 39%
Simple	45% sents AI p(o'all) = 49%	32% sents AI p(o'all) = 47%	25% sents AI p(o'all) = 49%	20% sents AI p(o'all) = 46%	58% sents AI p(o'all) = 50%

GPTZero results for ChatGPT generated essays.

Variant	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Default	100%	100%	76%	100%	64%
Long	0%	26%	59%	67%	51%
Short	0%	X	31%	82%	27%
Complex	33%	15%	0%	63%	0%
Simple	100%	0%	100%	100%	71%

Turnitin results for ChatGPT generated essays.

Results: Claude-2 generated essays.

Variant	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Default	0%	0%	0%	0%	0%
Long	0%	0%	0%	0%	ca. 60% text AI p(AI) = 83.6%
Short	0%	0%	0%	0%	0%
Complex	ca. 74% text AI p(AI) = 67.5%	0%	0%	ca. 70% text AI p(AI) = 83.2%	ca. 87% text AI p(AI) = 68.0%
Simple	0%	0%	0%	0%	0%

Copyleaks results for Claude-2 generated essays.

Variant	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Default	35% sents AI p(o'all) = 50%	3% sents AI p(o'all) = 20%	14% sents AI p(o'all) = 22%	3% sents AI p(o'all) = 46%	9% sents AI p(o'all) = 23%
Long	36% sents AI p(o'all) = 51%	63% sents AI p(o'all) = 25%	3% sents AI p(o'all) = 12%	13% sents AI p(o'all) = 37%	0% sents AI p(o'all) = 9%
Short	0% sents AI p(o'all) = 1%	0% sents AI p(o'all) = 2%	0% sents AI p(o'all) = 0%	0% sents AI p(o'all) = 0%	2% sents AI p(o'all) = 22%
Complex	60% sents AI p(o'all) = 50%	9% sents AI p(o'all) = 44%	67% sents AI p(o'all) = 48%	75% sents AI p(o'all) = 25%	23% sents AI p(o'all) = 48%
Simple	0% sents AI p(o'all) = 2%	0% sents AI p(o'all) = 0%	0% sents AI p(o'all) = 1%	0% sents AI p(o'all) = 0%	0% sents AI p(o'all) = 1%

GPTZero results for Claude-2 generated essays.

Variant	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Default	0%	0%	0%	0%	0%
Long	0%	0%	0%	0%	0%
Short	0%	0%	0%	0%	0%
Complex	0%	0%	0%	X	0%
Simple	0%	0%	0%	0%	0%

Turnitin results for Claude-2 generated essays.

Observations

All table cells should indicate 100% text as AI, and according to detector, at 100% probability – because all the essays are entirely AI-generated. Any table cell that shows 0% represents a clear false negative, and any cell close to 0% indicates that a human user would probably interpret that as 'human' and, therefore, represents a false negative in practice, and even if for different reasons, also an 'X'. This is important in the academic misconduct context: any false-negative is potentially a cheated student submission going undetected – 'slipping under the radar' – which compromises our abilities to (a) help any such student study with integrity, and (b) protect our institutional integrity and reputation.

It is clear that all three detectors perform better with ChatGPT text, implying that their configurations/training are optimised for ChatGPT (and possibly other named AI 'writers') and/or that their training data has included little if any text generated by Claude-2. This raises questions with regard to statements some detectors make along the lines 'and other unnamed AI writers' to imply that they work with more AI 'writers' than those they explicitly list. Also, this indicates possible overfitting of the models in order to achieve the claimed headline accuracies with regard to specific AI writers, noting that none of the three listed Claude-2 (at the time of the testing), which has implications for how rapidly and accurately the models can be reconfigured/retrained in response to new or different or updated/upgraded AI 'writers' – and any classifications made pending reconfiguration/retraining.

This is not to fault the underpinning mathematics: simply, it highlights the increasing problems and challenges in trying to produce 'general-purpose' AI-text detectors when the AI 'writers' are continually improving and producing ever more human-like text. However, it also highlights that we need to read the small print with regard to marketing claims very, very carefully. See US FTC blog here (Michael Atleson, FTC Attorney, last updated 06/07/2023) https://www.ftc.gov/business-guidance/blog/2023/07/watching-detectives-suspicious-marketing-claims-tools-spot-ai-generated-content.

There's also an increasing number of reports with regard to when AI-text detectors get things wrong more generally, not just false negatives as reported above. For example, and by no means an exclusive list (all as accessed on 20/11/2023):

- Victor Tangermann (09/01/2023) There's a Problem With That App That Detects GPT-Written Text: It's Not Very Accurate. Futurism https://futurism.com/gptzero-accuracy
- Carol Anderson (01/06/2023) The False Promise of AI Writing Detectors. Blog https://www.linkedin.com/pulse/false-promise-ai-writing-detectors-carol-anderson & https://www.carol-anderson.com/blog/the-false-promise-of-ai-writing-detectors
- Noor Al-Sibai (06/06/2023) AI Plagiarism Detection Software Keeps Falsely Accusing Students of Cheating. Futurism https://futurism.com/ai-plagiarism-software-false-accusing-students
- Rhiannon Williams (07/07/2023) AI-text detection tools are really easy to fool. MIT Technology review https://www.technologyreview.com/2023/07/07/1075982/ai-text-detection-tools-are-really-easy-to-fool/
- Daniel Sokol (10/07/2023) It is too easy to falsely accuse a student of using AI: a cautionary tale. Times Higher Education https://www.timeshighereducation.com/blog/it-too-easy-falsely-accuse-student-using-ai-cautionary-tale