## 1. The *Fine Print* from Some Popular AI Detector Tools

- **Turnitin** (https://www.turnitin.com/products/features/ai-writing-detection/)
  - “While Turnitin has confidence in its model, Turnitin does not make a determination of misconduct, rather it provides data for the educators to make an informed decision based on their academic and institutional policies. Hence, we must emphasize that the percentage on the AI writing indicator should not be used as the sole basis for action or a definitive grading measure by instructors.”

- **GPTZero** (https://gptzero.me/educators)
  - “Firstly, at GPTZero, we don't believe that any AI detector is perfect. There always exist edge cases with both instances where AI is classified as human, and human is classified as AI. Nonetheless, we recommend that educators can do the following when they get a positive detection:
    1. Ask students to demonstrate their understanding in a controlled environment, whether that is through an in-person assessment, or through an editor that can track their edit history (for instance, using our Writing Reports through Google Docs). Check out our list of several recommendations on types of assignments that are difficult to solve with AI.
    2. Ask the student if they can produce artifacts of their writing process, whether it is drafts, revision histories, or brainstorming notes. For example, if the editor they used to write the text has an edit history (such as Google Docs), and it was typed out with several edits over a reasonable period of time, it is likely the student work is authentic. You can use GPTZero's Writing Reports to replay the student's writing process, and view signals that indicate the authenticity of the work.
    3. See if there is a history of AI-generated text in the student's work. We recommend looking for a long-term pattern of AI use, as opposed to a single instance, in order to determine whether the student is using AI.”

- **PlagiarismCheck.org** (https://plagiarismcheck.org/ai-plagiarism-checker-and-content-detector/)
  - “It is worth mentioning that no AI checker provides a final decision on whether a machine or a human wrote the text. However, you can draw your conclusion based on the tool's analysis.”

## 2. Independent Studies Investigating Accuracy of AI Detector Tools

| Study | Key Findings |
|---|---|
| Evaluating the Efficacy of AI Content Detection Tools in Differentiating Human and AI-Generated Text (Elkhatat, Elsaid & Almeer, 2023) | ▪ Detectors are more accurate in detecting AI-generated text by GPT-3.5 than GPT 4.0 (note: GPT 4.0 was just released to the public on May 14, 2024)<br>▪ Detectors are unreliable when run on text written by humans. |
| Testing of Detection Tools for AI-Generated Text (Weber-Wulff et al., 2023) | ▪ Detectors are biased towards human output (human text is less likely to be flagged as AI-generated than an AI-generated text to be flagged as human)<br>▪ Simply paraphrasing the texts using the AI tool Quillbot significantly worsened the detection accuracy. |
| The False Positives and False Negatives of Generative AI Detection Tools in Education and Academic Research: The Case of ChatGPT (Dalalah & Dalalah, 2023) | ▪ Related literature written by humans are more likely to be falsely flagged as AI-generated than abstracts written by humans.<br>▪ Distribution of percentage rates reported by the AI detector tool across human and AI-generated text have a large overlap, suggesting that false positives and false negatives can happen. |
| The Effectiveness of Software Designed to Detect AI-Generated Writing A Comparison of 16 AI Text Detectors (Walters, 2023) | ▪ Three detectors had high accuracy for the dataset that was used.<br>▪ However, the other detectors struggle detecting between human written and GPT 4.0 generated content. (note: GPT 4.0 was just released to the public on May 14, 2024) |
| An Empirical Study of AI Generated Text Detection Tools (Akram, 2023) | ▪ There is large variability between different detector tools, with accuracies ranging from 55% to 97% |

## 3. Other Considerations

- There is some evidence that AI detector tools may be biased against non-native English speakers (Liang et al., 2023)
- There is evidence that AI detectors can easily be fooled by simple techniques, such as asking ChatGPT to paraphrase the submission (Foster, 2023).
- Various studies report different, sometimes even contradictory accuracy rates for AI detector tools.
- We should consider that these studies are constrained by the empirical data that they used for evaluation. For example, an opinion piece may be written differently than a technical report, and that may affect AI detector accuracy.

## References

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity, 19(1), 17.

Foster, A. (2023). Can GPT-4 Fool TurnItIn? Testing the Limits of AI Detection with Prompt Engineering.

Dalalah, D., & Dalalah, O. M. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. The International Journal of Management Education, 21(2), 100822.

Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-Generated Essays in a Large-Scale Writing Assessment: Is There a Bias Against Non-Native English Speakers?. Computers & Education, 105070.

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. Open Information Science, 7(1), 20220158.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19(1), 26.