ChatGPT Blog – Spring 2023

Can We Detect AI Usage?

By Sabrina Davis

With the popularity of ChatGPT, chatbots and other types of AI have been getting a lot of attention, especially in higher education. Given the capabilities of AI, many instructors are concerned about whether students are maintaining academic integrity if they are using AI when navigating through any part of the research and writing processes. To ease this concern, many are wondering, is there a way to know with 100% certainty whether a student has committed academic dishonesty by using an AI? The short answer: unfortunately, no.

While there is not currently a magic button to push that will give us a definite answer, there are tools available that can assist instructors in assessing the likelihood that an AI was used. In this post, we will highlight a small selection of the common AI detectors that instructors may encounter as they are searching for these kinds of tools. We will also briefly discuss the limitations of detection tools and provide helpful tips that instructors can use in conjunction with these AI detectors.

AI Text Classifier – Open AI

This was developed by OpenAI, the same company behind ChatGPT, and was released in January 2023 and is free to use. This tool can be useful in detecting recent GPT models, such as ChatGPT, and is fairly user-friendly. The results will scan the text as a whole and label it as either **very unlikely, unlikely, unclear if it is, possibly**, or **likely** AI-generated. Some of the unique limitations for this tool are that it requires a minimum of 1,000 characters (approximately 150-250 words) to evaluate a selection of text at a time.

GPTZero

An app developed by a student at Princeton University following the popularity of ChatGPT. This tool attempts to measure perplexity and

burstiness scores to determine the likelihood that something has been authored by an AI. The thought is that text that is more varied and complex (specifically varied sentence length) is more likely to have been written by a human. After text has been imputed into the classifier, it will highlight sentences that are likely to have been generated by an AI. One of the unique limitations for this tool is that the free version only allows for a max of 5,000 characters to be used at a time.

Winston AI Detection

Winston AI utilizes optical character recognition (OCR) technology to scan documents, as well as photographs, to detect possible AI usage. Rather than relying on a copy and paste function like some of the other tools, users have the ability to upload materials in a variety of formats. Wintson AI will evaluate materials on a scale of 0-100 and provide a percentage of probability that something was created by an AI. One of the unique limitations of this tool is that the free version only allows up to 2,000 words to be scanned total. Paying a monthly fee would allow for up to 80,000 words to be scanned per month.

Turnitin's AI Writing Detection

Many in higher education are familiar with the popular plagiarism detection service, Turnitin. In response to concerns surrounding AI and academic honesty, Turnitin added an AI detection feature. When papers are uploaded into Turnitin they are first broken up into segments of about 5-10 sentences. The segments are then run through the AI detection system and each sentence within the segments is given a score between 0 and 1. 0 meaning that it was likely not generated by an AI and 1 meaning that the entire sentence was likely generated by an AI. Turnitin then takes the average scores for the segments and provides an overall prediction of the percentage of the paper that was AI generated. One of the unique limitations of this tool is that users must have a subscription to Turnitin in order to use their AI detection service.

Taking a Holistic Approach

While the AI detection tools may be useful, they are not able to say with 100% certainty whether academic dishonesty has taken place. There may be times when these tools may generate a "false-positive" for AI detection when something was in fact fully created by a human. There may also be times when these tools do not recognize that something was generated by an AI. For example, we asked ChatGPT to write a short essay about student loan forgiveness. We then copy and pasted that essay into the OpenAI AI Text Classifier and the results indicated that the text was "very unlikely" to have been AI generated. Whether an instructor is a fan of using AI in the learning environment or not, it's important to take a holistic approach to determine if inappropriate AI usage is suspected. Here are some additional tips to use alongside AI detectors.

1. **Look for consistency in language and sentence structure**. Writing generated by an AI can feel very static to read because it does not always provide variety in sentence structure and length. AI generated text can also be repetitive. Humans tend to have more variety in the language we use and how we write.

2. **Check the citations.** This one can be a little tricky because many students are still learning when and how to provide proper citations, but AI generated content is likely to either provide no citations at all, or the citations provided will not be credible. AI may also fabricate a citation to a resource that does not exist at all.

3. **Go with your gut.** Humans can be pretty good at just feeling like something isn't quite right. AI generated content is kind of like the uncanny valley of writing. It may look pretty close to what a human could produce but something about it just seems "off" when you read between the lines. That said, it's important to assume positive intent from the

students and not rely solely on a gut feeling when making a case for academic dishonesty.

Additional Resources:

- Texas Tech University Libraries "AI Tools for Research and Writing" Library Guide
- "Educator Considerations for ChatGPT" by OpenAI
- "ChatGPT is Everywhere" in Chronicle of Higher Education
- "10 Ways GPT-4 is Impressive but Still Flawed" in New York Times

Notes & ideas: