Guidance for using the AI detection tool in Turnitin

Aim of the document

To provide a brief description of the AI detection tool in Turnitin and to provide details on how the results of the tool should be interpreted when assessing whether student work has been generated using artificial intelligence (AI).

Intended audience

Any staff involved in marking work where Turnitin has been used to provide a similarity report and AI detection report.

Introduction

What is the AI detection tool in Turnitin?

The AI detection tool available in Turnitin is an automatically enabled feature that facilitates the detection of text written by the Large Language Model (LLM) ChatGPT. At this time, the tool cannot reliably detect the output of other LLMs, such as Microsoft's Bing or Google's Bard, because it has been solely trained on the output of ChatGPT. However, internal testing has shown the tool will detect text that has been written by tools such as quillbot and Grammarly. Staff should be aware that the later tools are used by students as a reasonable adjustment and therefore staff will need to consider if use is appropriate for each specific student.

The purpose of the tool is to help those assessing student work to determine if any of the text within a submission has been generated by AI. They can use this information to help them decide if the use of AI within the assessment constitutes a breach of academic integrity. The AI detection tool provides a percentage score to the user which indicates how much of the submitted text has been generated by AI. You may be familiar with the similarity score provided by Turnitin to check for plagiarism, however the AI detection score is calculated differently. See detailed notes below under 'how to interpret the AI detection score'.

This document will explain how to interpret the AI detection score when marking student work and what actions you should take based on the information gained from the tool. Please note that the tool cannot be used to 'AI Proof' your assessments for the reasons detailed below. Assessors should still seek to combat the inappropriate use of AI through assessment design and engaging students in assessment tasks.

How to use the AI detection tool in practice:

How to get an output from the AI detection tool:

The AI detection tool has been available to users since April 2023. However, at the University of Liverpool, an institution-wide decision was made to keep the tool disabled until we had a better understanding of what the tool was, how it worked, and how it could be used. Now we are at this point, and the AI detection feature has been enabled. Markers using Turnitin will now be able to see an AI detection score along with the similarity score in the Feedback Studio (Fig 1.). However, the AI detection score cannot be seen by students. Therefore, staff need to ensure students are aware that their work will be put through the AI detection tool and that the score will only be visible to the person marking the work.

The tool can only be enabled or disabled at an institutional level, and therefore it is not possible for individual users or schools to enable/disable the feature. Therefore, all submissions now made to Turnitin will be accompanied with the AI detection score and users do not need to take any action or change their current workflow to enable the feature. To see which text is responsible for the score given, users simply need to click on the score and an additional window will open with specific text highlighted (details below).

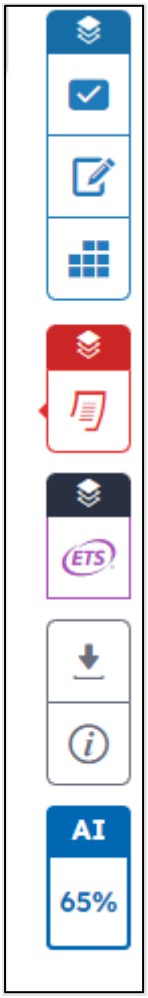

Figure 1. Screenshot of the AI detection score as it appears in the Feedback studio.

How to interpret the AI detection score:

The AI detection score is presented as a percentage, in a broadly similar way to the similarity score. However, the AI detection score is calculated in a far different manner, and it is possible that not all the text within a document will have been considered in the AI detection calculation. The AI detection tool only considers text written as prose sentences, meaning that only blocks of text that are written in standard grammatical sentences are analysed. Any other type of writing, such as lists, bullet points, text contained within tables, or any other text that does not constitute sentenced prose are not considered, and therefore do not contribute to the percentage calculation.

The AI detection score is therefore the percentage of prose text within the document that has a similar structure to that of AI generated text. Specifically, the tool has been trained to look for the sentence structures used by ChatGPT, where the next word in a sequence has been probabilistically determined and can therefore be predicted. This differs to human writing, where word choice tends to be inconsistent and is therefore much more difficult to predict. Thus, if the AI detection tool finds sentences where it can accurately predict the next word in the sentence, it will suspect the text has been generated using AI. The tool gives each segment of a submission a score between 0 (not AI) and 1 (definitely AI) based on how many sentences it can correctly predict in the segment. A high scoring section will then be highlighted and used in the calculation of the percentage displayed to the user (Fig. 2).

Figure 2. Example of scores associated with different segments of a document, showing that values greater than 0.8 are highlighted to user (taken from Turnitin: AI Writing Detection Capabilities - Frequently Asked Questions).

How reliable is the AI detection tool?

Turnitin have designed the tool to be as accurate as possible, and it is more likely to miss AI use than state that human written work is that of AI. However, they do state that their accepted false positive rate is 1%. Therefore, 1 out of every 100 claims of AI usage are likely to be false, i.e. it will state that work written by a human has been AI-generated. It is worth noting that low scores (<20%) have an increased likelihood of containing false positives (Fig. 3). As a result of this it is recommended that AI scores of under 20% are disregarded. There are also reports of increased false positive rates with work by non-native English speakers (Liang et al, 2023), and thus these are important to things to consider when judging if an academic integrity breach may have occurred.

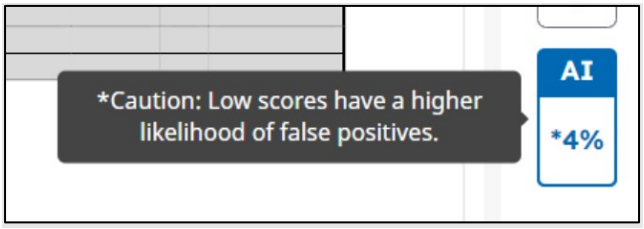

Figure 3. Warning given by the AI Detection tool that states low scores have a higher likelihood of false positives.

(A detailed explanation of false positives can be found on the Turnitin webpages: Understanding false positives within our AI writing detection capabilities)

The AI generation tool is therefore less reliable than the similarity detection software and as a result it must be used with caution (Fig. 4). Each marker will be responsible for assessing how applicable the score is to the submitted work and whether there is enough evidence for an academic integrity investigation. The AI detection score on its own would never be sufficient evidence for an academic misconduct case. Please see the section below for advice on how to judge if an academic integrity breach may have occurred.

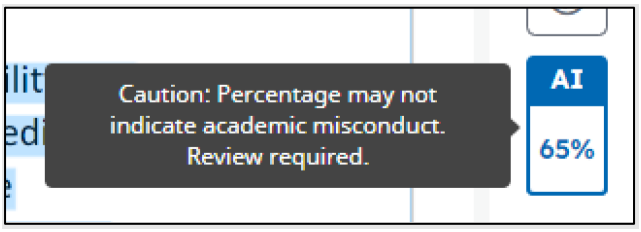

Figure 4. Warning presented with all AI detection scores to state that the score on its own does not signify academic misconduct.

A final point on the reliability of the tool: if a student has paraphrased the output of AI (either manually or by using paraphrasing tools), then the AI detection tool may not detect the use of AI. In fact, there is a growing amount of online content and apps dedicated to describing how to "beat AI detection", e.g. How To Pass Turnitin AI Detection and Plagiarism. Therefore, assessment writers should not rely on the AI detection tool as a way of "AI proofing" their assessment. It will ultimately fail in this task. Instead, assessment

writers should be focussing on assessment design and student engagement with the task to avoid academic integrity breaches.

How to make a judgement:

If you find a submission of work has a concerning section of text that has been highlighted as AI-generated, and you therefore suspect a student has potentially used AI to write all or part of their assignment, it is strongly recommended that you use the score as only one part of your evidence gathering. Initially, you need to click onto the AI detection score to open an additional webpage that will highlight to you the text which the tool suspects has been written by AI (Fig 5.).

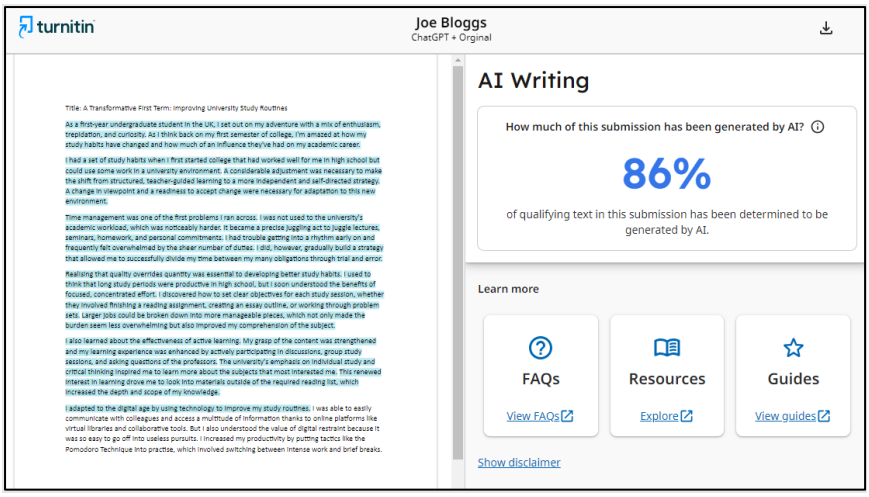

Figure 5. Example of text highlighted as AI generated.

After examining the suspicious text, you should then use your own experience and knowledge, as you would with the similarity scores, to come to an overall conclusion as to whether you think a breach of academic integrity has occurred. You may wish to consider the following as indicators to inform whether to undertake further investigation:

- Is the referencing accurate?
- Does the writing style match the student's previous work?
- Does the textual voice of the highlighted section match that of the rest of the assessment (if the rest of the assessment has not been flagged as AI-generated)?
- Is there any critical analysis?
- Is the work vague and off topic?

These are the same considerations one would take when investigating other forms of misconduct, such as contract cheating (where students either commission someone else to write their work or buy prepared work from an 'essay mill'). Thus, the AI detection tool is just one factor to consider when looking at whether academic misconduct has occurred.

When approaching students to discuss your concerns regarding their work it is important to bear in mind that students will not have any knowledge of their AI score. Also, there is the possibility that some of the score might be due to a false positive detection, and thus

these conversations need to be approached carefully. Turnitin have suggested following one of these lines of enquiry with students to help them explain their work before raising concerns of academic misconduct (Approaching a student regarding potential AI misuse):

- Explain your process for completing this assignment. Let's look at areas that you are proud of and areas that you think could use some extra improvement.
- Let's examine your Turnitin AI writing detection report together. Let's look at the highlighted areas and discuss your choices. Let's examine how GAI could have been better used to inform and advance your own ideas and develop your own conclusions.

Note that for some assessments, AI use is encouraged (and required) by the assessor and thus this score can be ignored in those instances where the AI generated text has been properly cited.

Further resources:

- There is a comprehensive list of FAQs available on the Turnitin website: (AI Writing Detection Capabilities - Frequently Asked Questions).
- Descriptions of academic misconduct can be found in Appendix L of the Code of Practice on Assessment.

If you would like further training and guidance on using the tool, please contact the CIE team, who will be happy to assist you (cie@liverpool.ac.uk).

References

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7). doi:10.1016/j.patter.2023.100779