UCI University of California, Irvine

Statement on Generative AI Detection

In April of 2023, Turnitin updated their similarity detection tool to optionally allow institutions to also have it scan for AI-generated text. The UCI Integrity in Academics Advisory Committee, acting under a charge from Vice Provost of Teaching and Learning Michael Dennin and Vice Chancellor, Student Affairs Willie L. Banks Jr., has examined the AI detection feature of Turnitin feature and its fitness for UCI and at this time has determined that it **will not be made available** to UCI. This decision mirrors that of many other higher education institutions, including within the UC system. More broadly, this group is not endorsing any AI detection tools at this time.

There are many reasons to be skeptical of this tool and others similar to it.

The tool can't specifically justify its conclusions

The similarity detection functionality of Turnitin is context-rich: if student content is flagged as being similar to something submitted or published elsewhere, instructors receive information about the match and can quickly begin assessing whether the report points to an educational opportunity on use of proper citations or may indicate plagiarism. On the other hand, the AI detection tool offers no such context, stating only that the identified percentage of the submission "has been determined to be generated by AI."

Turnitin has also, to date, been fairly guarded about how their model identifies AI-generated writing. Their most detailed explanation hinges on next-word probability: the concept that, as ChatGPT and similar models output a string of text, they are simply choosing the most likely word that should go after the word they have just chosen, based on the many millions of pages of text they have been fed as part of their "training." Turnitin argues that humans, by contrast, choose words in an "inconsistent and idiosyncratic" fashion, so detection tools can exploit this difference to flag AI-generated text. What this leaves out is that models like ChatGPT have an intentional element of randomness to their word selection; if they did not, every answer to identical prompts would be the same as long as the response began with the same word. Furthermore, while students may *tend* to choose words in a certain fashion, it follows that there are some who do not, and this opens the door for false positives.

The potential for false positives is concerning

When it was initially released, Turnitin said that the rate of "false positives" – human-written text flagged as AI-generated – was less than 1%. After several months of use, however, that was changed to "less than 1% for documents with 20% or more AI writing," which is a heavy qualifier. Furthermore, they now say that "there is a 4% likelihood that a specific sentence highlighted as AI-written might be human-written." Presumably, this would affect students who had the misfortune of choosing words in an order that Turnitin's probability model says they should not.

These statistics leave questions unanswered, like the difference between false-positive rates for content generated by students for whom English is not their first language versus those for whom it is. Some testing indicates that non-native English speakers may be disadvantaged by AI detectors. The main reason these questions are unanswered is because Turnitin has not opened their detection technology up to researchers or explained in-depth how it works.

Because of these false positives, there is a real risk that the AI detection tool could be drawing unwarranted negative attention to students who will then have a substantial burden in proving the negative about their non-use of ChatGPT. Indeed, Turnitin advises interpreting their AI detection score with great caution, arguing that doing so requires "scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred."

This is a rapidly evolving space

Even if AI detection is within the threshold of viability today, there are good reasons to believe it will not be viable in the long term. As of this writing, there is a cottage industry of generative AI tools that draft reasonably natural-sounding text. These tools are evolving rapidly and will only grow more sophisticated. Since its release in early 2023, Turnitin's AI detection tool has only claimed to be able to detect writing from GPT-3 and GPT-3.5, with GPT-4 detection working "most of the time." That means it doesn't claim to detect writing from Google's Bard, Meta's LLaMA, Anthropic's Claude, or any others.

Pedagogical remedies are preferable

In the absence of consistently reliable detection, The Division of Teaching Excellence and Innovation is updating a page with resources and ideas for adapting pedagogy to this new technological climate.