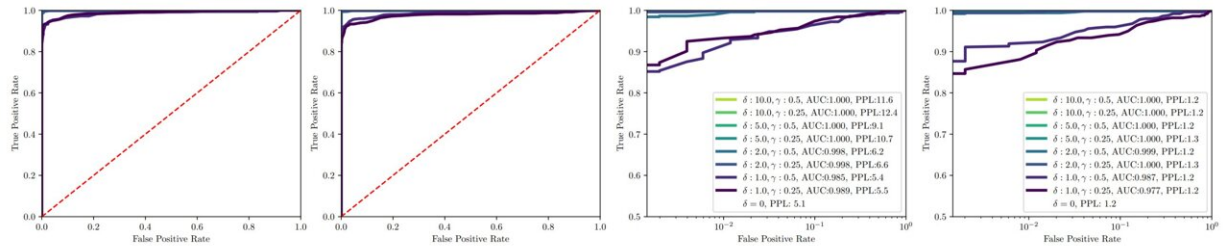
Exploring how to add hidden electronic watermarks to works written by AI systems

January 30 2023, by Bob Yirka



ROC curves with AUC values for watermark detection. Several choices of watermark parameter δ are shown for (a) multinomial sampling and (b) greedy decoding with 8-way beam search. (c,d) The same charts with semilog axes. Higher δ values achieve stronger performance, but additionally we see that for a given δ, the beam search allows the watermark to capture slightly more AUC than the corresponding parameters under the multinomial sampling scheme. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2301.10226

A team of computer scientists at the University of Maryland has developed a means of adding watermarks to text generated by AI systems. They have posted a paper describing their approach on the *arXiv* preprint server.

Text generating AI systems such as ChatGPT have been in the news a lot of late. Some news sites have reported that students have been caught using the chatbot to write papers for them. And other interested parties

have tested their use on test-taking and found them to perform well.

Meanwhile, teachers, professors and others in the education field have grown increasingly concerned about how to proceed as they become unable to tell whether or not papers turned in by students were written by an AI system. In this new effort, the group in Maryland has developed a means for assisting those with such concerns—the use of watermarks.

Just as with cash money or other printed documents, watermarks are information hidden in printed material that can only be seen under certain conditions, such as under a special light. The researchers suggest that companies such as OpenAI, the makers of ChatGPT, could add identifiers to the text created by their bot that cannot be seen by the casual user (the student) but could be detected by a software application used by teachers. For this approach to work, most or all of the makers of AI text generators would have to buy into the plan, either willingly or under government enforcement.

Creating a watermark in AI-generated text would involve more than adding a bit of metadata to a text file (as occurs with photographs), because the generated text could be easily copied using a smartphone or other device.

Thus, the watermark would have to exist within the text rather than behind it. To create such a watermark, the team in Maryland noted that text generation systems work by predicting and choosing one word at a time as they produce text and that they do so in a predictable way. As a text generator works, it chooses words that appear to be a good fit, and that must then be greenlisted by other code before they are used.

The researchers noted that text written by an AI tends to have more greenlisted words than text written by humans, suggesting a pattern that could be used as a watermark. They wrote an algorithm able to detect

these words and found it worked quite reliably. They note their approach may only work for some AI systems, but suggest that other watermarking systems could be built for other systems.

More information: John Kirchenbauer et al, A Watermark for Large Language Models, *arXiv* (2023). DOI: 10.48550/arxiv.2301.10226