CLIENT ALERTS & INSIGHTS

Advancing AI Transparency: The Rise and Challenges of Watermarking Technology for Identifying AI-Generated Content

MAY 9, 2024

Authors: Cristina Q. Almendarez

With the rising use of artificial intelligence-written text, demand for ways to identify whether content is human-written or AI-generated is likewise increasing. One such tool embeds patterns of words interwoven into AI-generated text called watermarks.

AI language models like ChatGPT draft a response to a prompt by predicting the word most likely to follow based on the input, and then generating the next word accordingly, one word at a time. After each word, a watermarking algorithm randomly divides the language model's vocabulary into words on two lists: a "greenlist" and a "redlist." The algorithm then prompts the model to choose words on the greenlist such that the more greenlisted words in a passage, the more likely it is that the text was generated by AI. Because human-written text generally contains a more random mix of words, watermarking creates an identifiable signature that is invisible to humans but algorithmically detectable and that can be traced back to the AI language model.

Initial studies of the effectiveness of watermarks showed promise, as researchers were able to identify AI-generated text with ease using algorithms to detect watermarks, motivating stakeholders to implement or invest in further development. Based on that initial promise of watermarking as a tool to authenticate and identify the provenance of digital content, AI companies including OpenAI, Alphabet and Meta Platforms voluntarily pledged to the White House to use watermarking to identify AI-generated content.

On the regulatory side, a 2023 White House executive order on artificial intelligence called for the use of AI watermarking to better establish the authenticity and provenance of digital content. Likewise, the first comprehensive AI law slated to go into enforcement in May, the European Union's AI Act, will require developers to watermark AI-generated content to enable users to identify AI-written text.

More recent studies give reason to suggest that AI watermarking may not be as effective as initially thought. One recent study showed that watermarks can be removed or stolen, which could potentially trick people into mistaking AI-generated text as human-written text, and vice versa. In that study, researchers were able to reverse-engineer a watermark, creating an approximate model of the watermarking rules applied by a given AI model using only API access to the AI language model with watermarked content and minimal costs. With such a model in hand, researchers can then either manipulate text in one of two ways. A bad actor could modify human-derived content to pass it off as watermarked by applying the watermark using the AI model. Alternatively, a bad actor could scrub the watermark from AI-generated text so that it can be disguised (and undetectable) as

human-written. The researchers were able to generate text detected as watermarked with an over 80% success rate, and they were able to strip the watermark from AI-generated text with an 85% success rate.

The researchers suggest that despite the findings, watermarking remains the most promising way to reliably detect A.I. written text.

Ultimately, the journey toward fully reliable AI transparency tools like watermarking is proving to be as dynamic as the technology itself. While challenges remain, the commitment from tech giants and regulators to refine and enforce these mechanisms reflects a shared dedication to and awareness of the importance of safeguarding the integrity of digital communication. Moving forward, it will be crucial to continue to foster innovation, collaboration and rigorous scrutiny to ensure that the digital content of tomorrow remains authentic, traceable and trustworthy. In doing so, the goal will be to not only enhance the utility of AI but also protect the foundational trust in human and machine-generated content that our digital future demands.

The Benesch AI Commission is committed to staying at the forefront of knowledge and experience to assist our clients in compliance efforts or questions involving safe implementation of AI. For inquiries regarding AI watermarking and its applications, please contact:

Cristina Almendarez at calmendarez@beneschlaw.com or 312.624.6382.