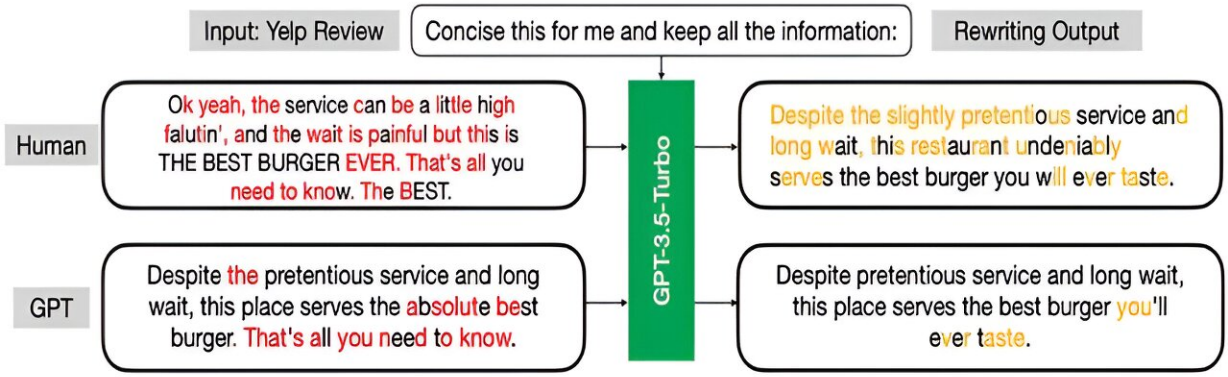
Who wrote this? Engineers discover novel method to identify AI-generated text

March 20 2024, by Bernadette Young



Raidar detects machine-generated text by calculating rewriting modifications. This illustration shows the character deletion in red and the character insertion in orange. Human-generated text tends to trigger more modifications than machine-generated text when asked to be rewritten. Credit: Yang and Vondrick labs

Computer scientists at Columbia Engineering have developed a transformative method for detecting AI-generated text. Their findings promise to revolutionize how we authenticate digital content, addressing

mounting concerns surrounding large language models (LLMs), digital integrity, misinformation, and trust.

Computer Science Professors Junfeng Yang and Carl Vondrick spearheaded the development of Raidar (geneRative AI Detection viA Rewriting), which introduces an innovative approach for identifying whether text has been written by a human or generated by AI or LLMs like ChatGPT, without needing access to a model's internal workings.

The paper, which includes open-sourced code and datasets, will be presented at the International Conference on Learning Representations (ICLR) in Vienna, Austria, May 7–11, 2024. It is currently available on the *arXiv* preprint server.

The researchers leveraged a unique characteristic of LLMs that they term "stubbornness"—LLMs show a tendency to alter human-written text more readily than AI-generated text. This occurs because LLMs often regard AI-generated text as already optimal and thus make minimal changes.

The new approach, Raidar, uses a language model to rephrase or alter a given text and then measures how many edits the system makes to the given text. Raidar receives a piece of text, such as a social media post, product review, or blog post, and then prompts an LLM to rewrite it. The LLM replies with the rewritten text, and Raidar compares the original text with the rewritten text to measure modifications. Many edits mean the text is likely written by humans, while fewer modifications mean the text is likely machine-generated.

Raidar's remarkable accuracy is noteworthy—it surpasses previous methods by up to 29%. This leap in performance is achieved using state-of-the-art LLMs to rewrite the input, without needing access to the AI's architecture, algorithms, or training data—a first in the field of AI-

generated text detection.

Raidar is also highly accurate even on short texts or snippets. This is a significant breakthrough as prior techniques have required long texts to have good accuracy. Discerning accuracy and detecting misinformation is especially crucial in today's online environment, where brief messages, such as social media posts or internet comments, play a pivotal role in information dissemination and can have a profound impact on public opinion and discourse.

Authenticating digital content

In an era when AI's capabilities continue to expand, the ability to distinguish between human and machine-generated content is critical for upholding integrity and trust across digital platforms. From social media to news articles, academic essays to online reviews, Raidar promises to be a powerful tool in combating the spread of misinformation and ensuring the credibility of digital information.

"Our method's ability to accurately detect AI-generated content fills a crucial gap in current technology," said the paper's lead author Chengzhi Mao, who is a former Ph.D. student at Columbia Engineering and current postdoc of Yang and Vondrick. "It's not just exciting; it's essential for anyone who values the integrity of digital content and the societal implications of AI's expanding capabilities."

The team plans to broaden its investigation to encompass various text domains, including multilingual content and various programming languages. They are also exploring the detection of machine-generated images, videos, and audio, aiming to develop comprehensive tools for identifying AI-generated content across multiple media types.

More information: Chengzhi Mao et al, Raidar: geneRative AI Detection viA Rewriting, *arXiv* (2024). DOI: 10.48550/arxiv.2401.12970