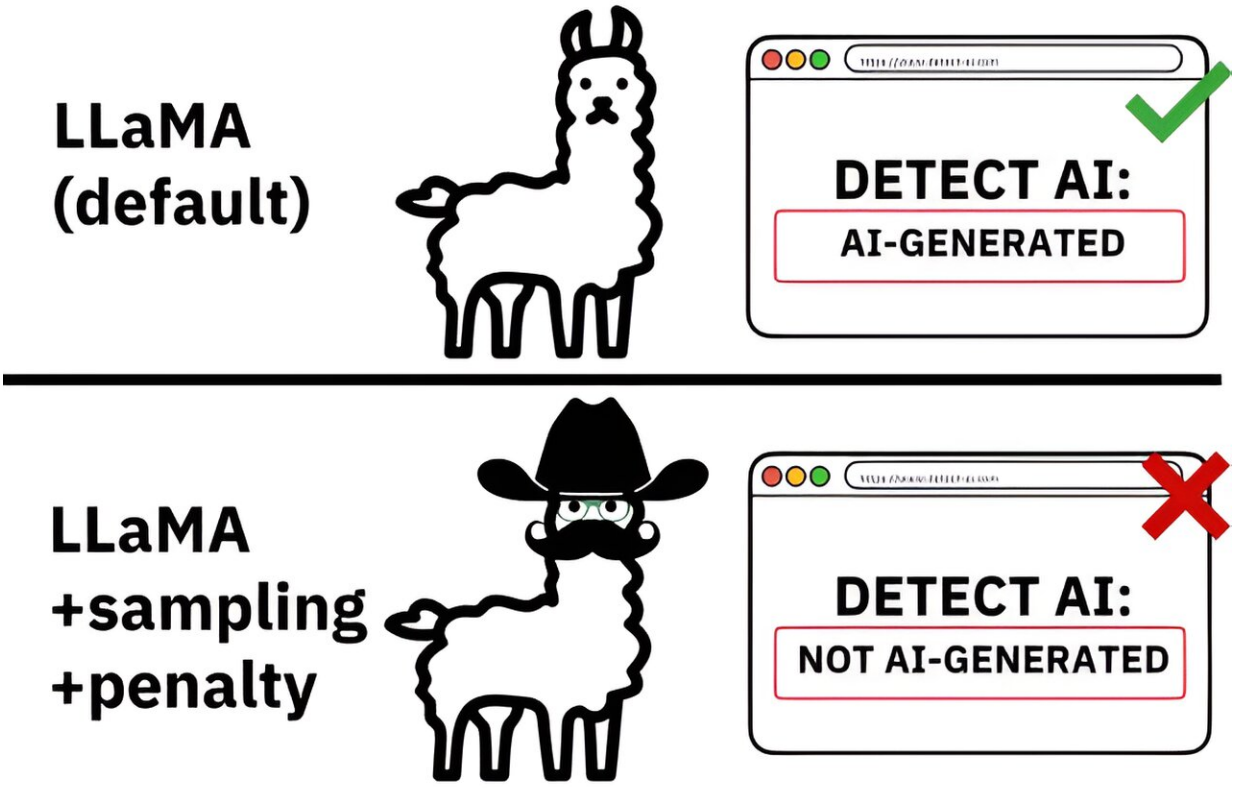
# Detecting machine-generated text: An arms race with the advancements of large language models

August 13 2024, by Melissa Pappas




Detectors are able to detect AI-generated text when it contains no edits or "disguises," but when manipulated, current detectors are not reliably able to detect AI-generated text. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2405.07940

Machine-generated text has been fooling humans for the last four years. Since the release of GPT-2 in 2019, large language model (LLM) tools have gotten progressively better at crafting stories, news articles, student essays and more, to the point that humans are often unable to recognize when they are reading text produced by an algorithm.

While these LLMs are being used to save time and even boost creativity in ideating and writing, their power can lead to misuse and harmful outcomes, which are already showing up across spaces we consume information. The inability to detect machine-generated text only enhances the potential for harm.

One way both academics and companies are trying to improve this detection is by employing machines themselves. Machine learning models can identify subtle patterns of word choice and grammatical constructions to recognize LLM-generated text in a way that our human intuition cannot.

Today, many commercial detectors are claiming to be highly successful at detecting machine-generated text, with up to 99% accuracy, but are these claims too good to be true? Chris Callison-Burch, Professor in Computer and Information Science, and Liam Dugan, a doctoral student in Callison-Burch's group, aimed to find out in their recent paper presented at the 62nd Annual Meeting of the Association for Computational Linguistics. The work is published on the *arXiv* preprint server.

"As the technology to detect machine-generated text advances, so does the technology used to evade detectors," says Callison-Burch. "It's an arms race, and while the goal to develop robust detectors is one we should strive to achieve, there are many limitations and vulnerabilities in detectors that are available now."

To investigate those limitations and provide a path forward for developing robust detectors, the research team created Robust AI Detector (RAID), a data set of over 10 million documents across recipes, news articles, blog posts and more, including both AI-generated text and human-generated text.

RAID serves as the first standardized benchmark to test detection ability in current and future detectors. In addition to creating the data set, they created a leaderboard, which publicly ranks the performance of all detectors that have been evaluated using RAID in an unbiased way.

"The concept of a leaderboard has been key to success in many aspects of machine learning like computer vision," says Dugan. "The RAID benchmark is the first leaderboard for robust detection of AI-generated text. We hope that our leaderboard will encourage transparency and high-quality research in this quickly evolving field."

Dugan has already seen the influence this paper is having in companies that develop detectors.

"Soon after our paper became available as a preprint and after we released the RAID data set, we started seeing the data set being downloaded many times, and we were contacted by Originality.ai, a prominent company that develops detectors for AI-generated text," he says.

"They shared our work in a blog post, ranked their detector in our leaderboard and are using RAID to identify previously hidden vulnerabilities and improve their detection tool. It's inspiring to see that the community appreciates this work and also strives to raise the bar for AI-detection technology."

So, do the current detectors hold up to the work at hand? RAID shows

that not many do as well as they claim.

"Detectors trained on ChatGPT were mostly useless in detecting machine-generated text outputs from other LLMs such as Llama and vice versa," says Callison-Burch.

"Detectors trained on news stories don't hold up when reviewing machine-generated recipes or creative writing. What we found is that there are a myriad of detectors that only work well when applied to very specific use cases and when reviewing text similar to the text they were trained on."

Faulty detectors are not only an issue because they don't work well, they can be as dangerous as the AI tool used to produce the text in the first place.

"If universities or schools were relying on a narrowly trained detector to catch students' use of ChatGPT to write assignments, they could be falsely accusing students of cheating when they are not," says Callison-Burch. "They could also miss students who were cheating by using other LLMs to generate their homework."

It's not just a detector's training, or lack thereof, that limits its ability to detect machine-generated text. The team looked into how adversarial attacks such as replacing letters with look-alike symbols can easily derail a detector and allow machine-generated text to fly under the radar.

"It turns out, there are a variety of edits a user can make to evade detection by the detectors we evaluated in this study," says Dugan. "Something as simple as inserting extra spaces, swapping letters for symbols, or using alternative spelling or synonyms for a few words can cause a detector to be rendered useless."

The study concludes that, while current detectors are not robust enough to be of significant use in society just yet, openly evaluating detectors on large, diverse, shared resources is critical to accelerating progress and trust in detection, and that transparency will lead to the development of detectors that do hold up in a variety of use cases.

"Evaluating robustness is particularly important for detection, and it only increases in importance as the scale of public deployment grows," says Dugan. "We also need to remember that detection is just one tool for a larger, even more valuable motivation: preventing harm by the mass distribution of AI-generated text."

"My work is focused on reducing the harms that LLMs can inadvertently cause, and, at the very least, making people aware of the harms so that they can be better informed when interacting with information," he continues. "In the realm of information distribution and consumption, it will become increasingly important to understand where and how text is generated, and this paper is just one way I am working towards bridging those gaps in both the scientific and public communities."

Dugan and Callison-Burch worked with several other researchers on this study, including Penn graduate students Alyssa Hwang, Josh Magnus Ludan, Andrew Zhu and Hainiu Xu, as well as a former Penn doctoral student Daphne Ippolito and Filip Trhlik, an undergraduate at University College London. They continue to work on projects that focus on advancing the reliability and safety of AI tools and how society integrates them into daily life.

**More information:** Liam Dugan et al, RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors, *arXiv* (2024). DOI: 10.48550/arxiv.2405.07940