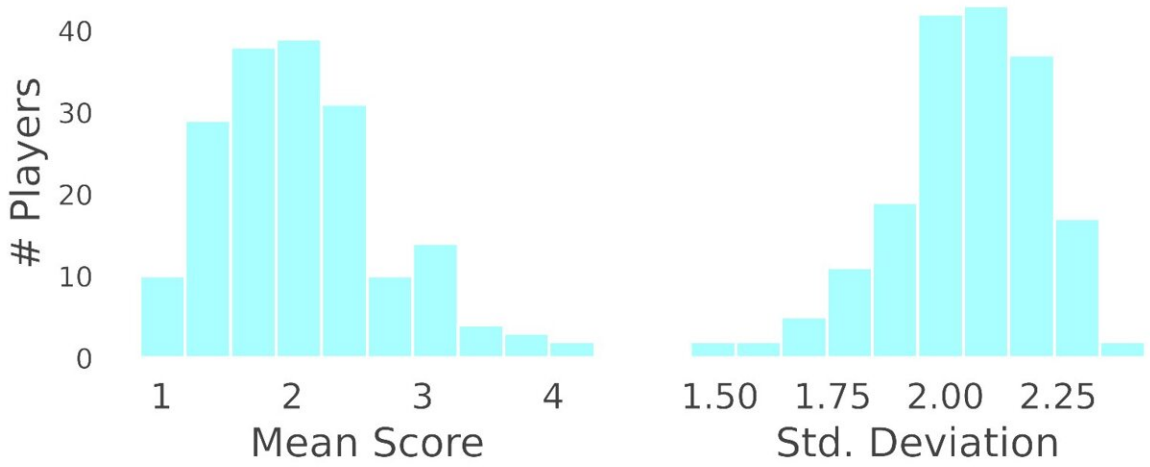
Real or fake text? We can learn to spot the difference

February 27 2023, by Devorah Fischler



Histogram of mean score and standard deviation of score among players who completed at least 20 rounds. We see large gaps in skill between players, with some having significantly higher mean score and lower variance than others. Credit: Real or Fake Text?: *Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text* (2023).

The most recent generation of chatbots has surfaced longstanding concerns about the growing sophistication and accessibility of artificial intelligence.

Fears about the integrity of the job market—from the creative economy to the managerial class—have spread to the classroom as educators rethink learning in the wake of ChatGPT.

Yet while apprehensions about employment and schools dominate headlines, the truth is that the effects of large-scale language models such as ChatGPT will touch virtually every corner of our lives. These new tools raise society-wide concerns about artificial intelligence's role in reinforcing social biases, committing fraud and identity theft, generating fake news, spreading misinformation and more.

A team of researchers at the University of Pennsylvania School of Engineering and Applied Science is seeking to empower tech users to mitigate these risks. In a peer-reviewed paper presented at the February 2023 meeting of the Association for the Advancement of Artificial Intelligence, the authors demonstrate that people can learn to spot the difference between machine-generated and human-written text.

Before you choose a recipe, share an article, or provide your credit card details, it's important to know there are steps you can take to discern the reliability of your source.

The study, led by Chris Callison-Burch, Associate Professor in the Department of Computer and Information Science (CIS), along with Liam Dugan and Daphne Ippolito, Ph.D. students in CIS, provides evidence that AI-generated text is detectable.

"We've shown that people can train themselves to recognize machine-generated texts," says Callison-Burch. "People start with a certain set of assumptions about what sort of errors a machine would make, but these assumptions aren't necessarily correct. Over time, given enough examples and explicit instruction, we can learn to pick up on the types of errors that machines are currently making."

"AI today is surprisingly good at producing very fluent, very grammatical text," adds Dugan. "But it does make mistakes. We prove that machines make distinctive types of errors—common-sense errors, relevance errors, reasoning errors and logical errors, for example—that we can learn how to spot."

The study uses data collected using Real or Fake Text?, an original web-based training game.

This training game is notable for transforming the standard experimental method for detection studies into a more accurate recreation of how people use AI to generate text.

In standard methods, participants are asked to indicate in a yes-or-no fashion whether a machine has produced a given text. This task involves simply classifying a text as real or fake and responses are scored as correct or incorrect.

The Penn model significantly refines the standard detection study into an effective training task by showing examples that all begin as human-written. Each example then transitions into generated text, asking participants to mark where they believe this transition begins. Trainees identify and describe the features of the text that indicate error and receive a score.

The study results show that participants scored significantly better than random chance, providing evidence that AI-created text is, to some extent, detectable.

"Our method not only gamifies the task, making it more engaging, it also provides a more realistic context for training," says Dugan. "Generated texts, like those produced by ChatGPT, begin with human-provided prompts."

The study speaks not only to artificial intelligence today, but also outlines a reassuring, even exciting, future for our relationship to this technology.

"Five years ago," says Dugan, "models couldn't stay on topic or produce a fluent sentence. Now, they rarely make a grammar mistake. Our study identifies the kind of errors that characterize AI chatbots, but it's important to keep in mind that these errors have evolved and will continue to evolve. The shift to be concerned about is not that AI-written text is undetectable. It's that people will need to continue training themselves to recognize the difference and work with detection software as a supplement."

"People are anxious about AI for valid reasons," says Callison-Burch. "Our study gives points of evidence to allay these anxieties. Once we can harness our optimism about AI text generators, we will be able to devote attention to these tools' capacity for helping us write more imaginative, more interesting texts."

Ippolito, the Penn study's co-leader and current Research Scientist at Google, complements Dugan's focus on detection with her work's emphasis on exploring the most effective use cases for these tools. She contributed, for example, to Wordcraft, an AI creative writing tool developed in tandem with published writers. None of the writers or researchers found that AI was a compelling replacement for a fiction writer, but they did find significant value in its ability to support the creative process.

"My feeling at the moment is that these technologies are best suited for creative writing," says Callison-Burch. "News stories, term papers, or legal advice are bad use cases because there's no guarantee of factuality."

"There are exciting positive directions that you can push this technology

in," says Dugan. "People are fixated on the worrisome examples, like plagiarism and fake news, but we know now that we can be training ourselves to be better readers and writers."

More information: www.cis.upenn.edu/~ccb/publica … ke-text-analysis.pdf

Conference: aaai-23.aaai.org/

Game: roft.io/

Provided by University of Pennsylvania