Title: Artificial intelligence content detection

URL: https://en.wikipedia.org/wiki/Artificial_intelligence_content_detection

PageID: 74313930

Categories: Category:Applications of artificial intelligence, Category:ChatGPT, Category:Computational fields of study, Category:Computational linguistics, Category:Computational neuroscience, Category:Cybernetics, Category:Data science, Category:Educational assessment and evaluation, Category:Formal sciences, Category:Large language models, Category:Natural language processing, Category:Plagiarism detectors, Category:Speech recognition, Category:Unsolved problems in computer science

-----

Artificial general intelligence

Intelligent agent

Recursive self-improvement

Planning

Computer vision

General game playing

Knowledge representation

Natural language processing

Robotics

AI safety

Machine learning

Symbolic

Deep learning

Bayesian networks

Evolutionary algorithms

Hybrid intelligent systems

Systems integration

Open-source

Bioinformatics

Deepfake

Earth sciences

Finance

Generative AI Art Audio Music

Art

Audio

Music

Government

Healthcare Mental health

Mental health

Industry

Software development

Translation

Military

Physics

Projects

AI alignment

Artificial consciousness

The bitter lesson

Chinese room

Friendly AI

Ethics

Existential risk

Turing test

Uncanny valley

Timeline

Progress

AI winter

AI boom

AI bubble

Glossary

v

t

e

Artificial intelligence detection software aims to determine whether some content (text, image, video or audio) was generated using artificial intelligence (AI). However, this software is often unreliable. [1]

Accuracy issues

Many AI detection tools have been shown to be unreliable in detecting AI-generated text. In a 2023 study conducted by Weber-Wulff et al., researchers evaluated 14 detection tools including Turnitin and GPTZero and found that "all scored below 80% of accuracy and only 5 over 70%." [2] They also found that these tools tend to have a bias for classifying texts more as human than as AI, and that accuracy of these tools worsens upon paraphrasing. [2]

False positives

In AI content detection, a false positive is when human-written work is incorrectly flagged as AI-written. Many AI detection platforms claim to have a minimal level of false positives, with Turnitin claiming a less than 1% false positive rate. [3] However, later research by The Washington Post produced much higher rates of 50%, though they used a smaller sample size. [4] False positives in an academic setting frequently lead to accusations of academic misconduct , which can have serious consequences for a student's academic record . Additionally, studies have shown evidence that many AI detection models are prone to give false positives to work written by those whose first language isn't English and neurodiverse people. [5] [6]

In June 2023, Janelle Shane wrote that portions of her book You Look Like a Thing and I Love You were flagged as AI-generated. [1]

False negatives

A false negative is a failure to identify documents with AI-written text. False negatives often happen as a result of a detection software's sensitivity level or because evasive techniques were used when generating the work to make it sound more human. [7] False negatives are less of a concern academically, since they aren't likely to lead to accusations and ramifications. Notably, Turnitin stated they have a 15% false negative rate. [8]

Text detection

For text, this is usually done to prevent alleged plagiarism , often by detecting repetition of words as telltale signs that a text was AI-generated (including hallucinations). They are often used by teachers marking their students, usually on an ad hoc basis. Following the release of ChatGPT and similar AI text generative software, many educational establishments have issued policies against the use of AI by students. [9] AI text detection software is also used by those assessing job applicants, as well as online search engines . [10]

Current detectors may sometimes be unreliable and have incorrectly marked work by humans as originating from AI [11] [12] [4] while failing to detect AI-generated work in other instances. [13] MIT Technology Review said that the technology "struggled to pick up ChatGPT-generated text that had been slightly rearranged by humans and obfuscated by a paraphrasing tool". [14] AI text detection software has also been shown to discriminate against non-native speakers of English. [10]

Two students from the University of California, Davis , were referred to the university's Office of Student Success and Judicial Affairs (OSSJA) after their professors scanned their essays with positive results; the first with an AI detector called GPTZero, and the second with an AI detector integration in Turnitin . However, following media coverage, [15] and a thorough investigation, the students were cleared of any wrongdoing. [16] [17]

In April 2023, Cambridge University and other members of the Russell Group of universities in the United Kingdom opted out of Turnitin's AI text detection tool, after expressing concerns it was unreliable. [18] The University of Texas at Austin opted out of the system six months later. [19]

In May 2023, a professor at Texas A&M; University–Commerce used ChatGPT to detect whether his students' content was written by it, which ChatGPT said was the case. As such, he threatened to fail the class despite ChatGPT not being able to detect AI-generated writing. [20] No students were prevented from graduating because of the issue, and all but one student (who admitted to using the software) were exonerated from accusations of having used ChatGPT in their content. [21]

In July 2023, a paper titled "GPT detectors are biased against non-native English writers" was released, reporting that GPTs discriminate against non-native English authors. The paper compared seven GPT detectors against essays from both non-native English speakers and essays from United States students. The essays from non-native English speakers had an average false positive rate of 61.3%. [22]

An article by Thomas Germain, published on Gizmodo in June 2024, reported job losses among freelance writers and journalists due to AI text detection software mistakenly classifying their work as AI-generated. [23]

In September 2024, Common Sense Media reported that generative AI detectors had a 20% false positive rate for Black students, compared to 10% of Latino students and 7% of White students. [24] [25]

To improve the reliability of AI text detection, researchers have explored digital watermarking techniques. A 2023 paper titled "A Watermark for Large Language Models" [26] presents a method to embed imperceptible watermarks into text generated by large language models (LLMs).

This watermarking approach allows content to be flagged as AI-generated with a high level of accuracy, even when text is slightly paraphrased or modified. The technique is designed to be subtle and hard to detect for casual readers, thereby preserving readability, while providing a detectable signal for those employing specialized tools. However, while promising, watermarking faces challenges in remaining robust under adversarial transformations and ensuring compatibility across different LLMs.

Anti text detection

There is software available designed to bypass AI text detection. [27] [28]

A study published in August 2023 analyzed 20 abstracts from papers published in the Eye Journal , which were then paraphrased using GPT-4 .0. The AI-paraphrased abstracts were examined for plagiarism using QueText and for AI-generated content using Originality.AI. The texts were then re-processed through an adversarial software called Undetectable.ai in order to reduce the AI-detection scores. The study found that the AI detection tool, Originality.AI, identified text generated by GPT-4 with a mean accuracy of 91.3%. However, after reprocessing by Undetectable.ai, the detection accuracy of Originality.ai dropped to a mean accuracy of 27.8%. [29] [30]

Some experts also believe that techniques like digital watermarking are ineffective because they can be removed or added to trigger false positives. [31] "A Watermark for Large Language Models" paper by Kirchenbauer et al. [26] also addresses potential vulnerabilities of watermarking techniques. The authors outline a range of adversarial tactics, including text insertion, deletion, and substitution attacks, that could be used to bypass watermark detection. These attacks vary in complexity, from simple paraphrasing to more sophisticated approaches involving tokenization and homoglyph alterations. The study highlights the challenge of maintaining watermark robustness against attackers who may employ automated paraphrasing tools or even specific language model replacements to alter text spans iteratively while retaining semantic similarity. Experimental results show that although such attacks can degrade watermark strength, they also come at the cost of text quality and increased computational resources.

Multilingual text detection

One shortcoming of most AI content detection software is their inability to identify AI-generated text in any language. Large language models (LLMs) like ChatGPT, Claude, and Gemini can write in different languages, but traditional AI text detection tools have primarily been trained in English and a few other widely spoken languages, such as French and Spanish. Fewer AI detection solutions can detect AI-generated text in languages like Farsi, Arabic, or Hindi. [citation needed]

Image, video, and audio detection

Several purported AI image detection software exist, to detect AI-generated images (for example, those originating from Midjourney or DALL-E). They are not completely reliable. [32] [33]

Others claim to identify video and audio deepfakes , but this technology is also not fully reliable yet either. [34]

Despite debate around the efficacy of watermarking, Google DeepMind is actively developing a detection software called SynthID, which works by inserting a digital watermark that is invisible to the human eye into the pixels of an image. [35] [36]

See also

Copyleaks

AI alignment

Artificial intelligence and elections

Comparison of anti-plagiarism software

Content similarity detection

Hallucination (artificial intelligence)

Natural language processing

References