AI Content Detector FAQs

Now more than ever, it's crucial to know what content is real and what was created by AI, whether you're browsing the internet, conducting research, or reading through student essays. The Copyleaks AI Content Detector was built to help provide insight into what content was written by a human and what was created by an AI.

Featured here are key questions we frequently get asked regarding our AI Content Detector.

How It Works

1. How is your AI content detection any different from other detectors?

We appreciate and applaud any effort to shed light on AI-generated content. However, there are several significant differences between other detectors and our AI Content Detector. For example:

- AI-based text analysis is the core of who we are. Since 2015, we've collected, ingested, and analyzed trillions of pages of crawled content and user-sourced content from thousands of universities and enterprises worldwide to train our models to understand the text and sniff out the signals created by other AI. Credible data at scale, coupled with machine learning and widespread adoption, allows us to continually refine and improve our ability to understand complex text patterns, resulting in 99.12% accuracy–far higher than any other AI content detector–and improving daily.

- Unlike browser-based detection, we're the only solution with seamless API and LMS integrations, empowering you to harness the power of AI Content Detector directly from your native platform and at scale.

- By looking at the individual paragraph, sentence, and word-by-word level, our detection report highlights the specific elements of the text that a human has written and those written by AI, along with a confidence level, offering a new level of insight and transparency.

2. How was the Copyleaks AI detection model trained?

We are able to recognize AI text patterns utilizing multiple techniques:

Our AI engine has been processing and learning how humans write as opposed to AI. Over the past 10 years, we've collected, ingested, and analyzed trillions of pages of crawled content and user-sourced content from thousands of universities and enterprises worldwide to train our models to understand the text and sniff out the signals created by other AI.

Also, utilizing AI technology, our AI detector can accurately recognize the presence of other AI-generated text and the signals it leaves behind, adding an additional layer of accuracy.

3. How do you avoid false accusations?

We strive to inspire authenticity and digital trust by creating secure environments to share ideas and learn with confidence. And with that comes the responsibility to ensure complete accuracy, particularly around false accusations. To address this, we have taken several precautions, including:

- Our detection and the algorithms that power it is designed for detecting human-generated text, versus AI-generated text, the latter of which gives less accurate detection and increases the likelihood of false positives.

- To help accelerate our learning and refine the models used, we implemented a feedback loop where users can rate their results' accuracy. This allows us to continually use examples of false positives, rare as they may be, to improve.

- We only introduce new model detection after thorough testing. Once our internal team testing reaches a high confidence threshold, we leverage beta testers, giving an additional layer of assurance.

- The chance for content written by a human to be falsely labeled as AI-generated content is 0.4%, and the chance that the same person using human-generated text is falsely labeled twice is 0.0016%.

4. What models can you detect, and what's the accuracy of each?

- As of March 21, 2023, we are able to detect the following models:
- ChatGPT (also called Chatbot)
- GPT4
- GPT3
- GPT2
- T5
- BERT
- Jasper

Using English text, the detection accuracy of each varies slightly from model to model, though each is above 98.0%.

There's a chance that you may encounter slightly different results, given the type of content being tested. Accordingly, we suggest conducting a number of tests to determine the success rate for your specific types of content.

5. Does the AI Content Detector support detection for BARD?

In most cases, we will be able to detect BARD. In the near future, we will train our models to fully detect BARD to ensure high accuracy and complete coverage.

6. What languages do you support, and what is the accuracy of each?

We are currently the only platform to detect AI content across multiple languages, including English, Spanish, French, Portuguese, German, Italian, Russian, Polish, Romanian, Dutch, Swedish, Czech, and Norwegian, with more languages currently in the works.

At the moment, English has the highest accuracy at 99.12%. We continue to develop our models to increase the accuracy across other supported languages, and there are plans to introduce accurate detection across dozens of additional languages.

7. Is the AI Content Detector available as part of my LMS Integration, such as Moodle, Canvas, etc.? What about Microsoft Teams?

Yes. AI content detection is built in as part of all our LMS integrations: Canvas, Moodle, Brightspace, Blackboard, Schoology, and Sakai. On the Similarity Report, you will see a section dedicated to AI content detection that will indicate the detection of any AI-generated content, indicating if the level of AI content is 'High,' 'Medium,' or 'Low' along with the possible AI written text highlighted within the report.

At this time, we do not offer integrations with Microsoft Teams, which offers a student-view integration with no separate teacher view. Since our integrations are integral to educators as well as students, we do not currently offer a Teams integration.

8. What data protection does Copyleaks have?

At Copyleaks, our products are routinely undergoing independent verification of privacy, security, and compliance control in efforts to achieve certifications against global standards to earn and retain the trust of the millions of Copyleaks users worldwide. Our current Copyleaks certifications and compliance standards include SOC2, GDPR, PCI Payment Card Industry Data Security Standard, and NIST Risk Management Framework (RMF). To learn more, please visit our Compliance and Certifications page.

Understanding the Results

1. How will I know if AI content has been detected?

If AI content has been detected in a scan, a notification on the Similarity Report will alert you.

2. Why doesn't the percentage shown on the PDF report always match the amount indicated in the highlighted text?

The AI Content Detector looks at prose on a sentence-by-sentence basis. Therefore, while most text may be highlighted in varying shades, you will see a different percentage throughout the text based on the probability levels of AI content detected.

3. Are you able to detect mixed text where human-created text has been amended with AI-generated text?

Yes.

Our detection report highlights the specific elements of the text that have been written by a human and those written by AI, along with a confidence level, offering a new level of insight and transparency.

4. What is the difference between the Similarity Score and the AI content detection percentage? Are they completely separate, or do they influence one another?

The Similarity Score and the AI detection percentage are independent and do not influence one another. The Similarity Score is the generated score indicating the percentage of matching text, including plagiarism, paraphrasing, etc., found within a scanned document when compared against the Copyleaks internal database, trillions of webpages, open-source journals, and more. The AI detection percentage pertains to the overall potential AI content detected within the scanned text.

5. Will the AI detection be a different workflow than how we are currently working with the Copyleaks report?

No, you will receive an AI content detection alert within the Similarity Report. If you are working with the API, you can choose how a positive AI content alert can be shown.

6. Will the addition of AI content detection to my Similarity Report change the workflow or how I use the report?

No. AI content detection does not change your workflow and how you use the Similarity Report. Adding the AI Content Detector to the Similarity Report is part of our ongoing efforts to provide a seamless experience for our users.

Detection Capabilities & Limitations

1. Are you able to detect if text has been put through a text spinner? And what if text contains intentional typos?

Our next version of AI Content Detector will be able to detect if text has been put through a text spinner or paraphrased down to the sentence-by-sentence level. The update will also be capable of determining whether typos have been intentionally added to the text in an attempt to throw off the AI Content Detector.

2. What are AI Content Detector's limitations?

Even with 99.12% accuracy, there are limitations to be aware of.

- Generally speaking, the accuracy of our detection increases as the text length increases. Accordingly, we suggest testing text containing at least 400 words.
- The accuracy of creative writing, including poems and song lyrics, is typically lower than other types of content. We continue to train our models to ensure high accuracy across all types of content.
- At the moment, English has the highest accuracy. With additional text ingestion and model training, the accuracy across all supported languages will only continue to improve.
- Like other detectors, AI-generated text that has been put through a text spinner or has been edited to avoid detection will likely register as human-created text. We are working on a solution to detect AI text that has gone through a text spinner or been manipulated.
- While false positives are exceedingly rare (0.4%), AI-generated text has a higher rate of registering as human-created text. As we continue to train the models, the rate of false negatives will continue to improve.

3. What will we have to do in order to support new product updates?

New product updates will continue to be released, and you'll automatically get upgraded to the latest version. We'll include release notes to ensure you are fully aware of what's changed or has been added.

4. Will Copyleaks be able to detect newer models that will come out?

Yes. We are fully committed to leveraging technology to find language and text detection solutions. Once new models are introduced, we'll train the system to accurately detect it.

5. What other AI content detection capabilities are you working on?

We are working on a number of different fronts, including:

- The ability to detect AI text that has gone through a text spinner or otherwise been manipulated (i.e.: including deliberate typos).
- Across-the-board accuracy improvements.
- The support of additional languages and models.

We'll continue to monitor the landscape and closely listen to user feedback to ensure we stay one step ahead of AI content generators and provide the most accurate results possible.

Building digital trust and confidence:
It's the Copyleaks way.

sales@copyleaks.com • copyleaks.com