

# AI Content Detector FAQs

Last updated 02/01/2024

## How It Works

[Page 2](#)

## Understanding the Results

[Page 5](#)

## Detection Capabilities & Limitations

[Page 7](#)

From ensuring cyber compliance to preventing copyright infringement, knowing what content is human-created and what was generated by AI is crucial. Featured here are key questions we frequently get asked regarding our AI Content Detector.

# How It Works

## How is it possible to determine if something is AI-generated?

When a Language Model writes a sentence, it probes all of its pre-training data to output a statistically generated sentence, which is simply not how a human writes. It becomes apparent when analyzed against a vast corpus of human writing.

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

There are several significant differences between other detectors and our AI Content Detector.

For example:

- 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 over 99% accuracy—far higher than any other AI content detector—and improving daily.
- As an enterprise-based platform, we're the only solution with seamless API and LMS integrations, allowing you to bring the power of AI Content Detector directly to your native platform and at scale.
- By looking at the individual paragraph and sentence level, our detection report highlights the specific elements of the text potentially written by AI and provides a confidence level.
- We are [GDPR-compliant](#) and the only [SOC2-certified](#) platform.

## How was the Copyleaks AI detection model trained?

We can recognize AI text patterns utilizing multiple techniques.

Since 2015, we've collected, ingested, and analyzed trillions of crawled and user-sourced content pages from thousands of universities and enterprises worldwide to train our models to understand how humans write as opposed to 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.

## How do you avoid false accusations?

The chance for content written by a human to be falsely labeled as AI-generated content is 0.2%, the lowest of any AI content detection platform. Nevertheless, we strive to inspire authenticity and digital trust by creating secure environments to share ideas and learn confidently, and that comes with 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 are designed for detecting human-generated text versus AI-generated text. Detecting AI text tends to give a lower accuracy 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 the accuracy of the results, which 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 testing reaches a high confidence threshold, we leverage beta testers, giving an additional layer of assurance.

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

As of February 1, 2024, we can detect the following models:

- ChatGPT (including GPT-3.5, GPT-4, and other sequential models)
- Bard
- Jasper 3
- T5

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 several tests to determine the success rate for your specific content type.

## **What languages do you support, and what is the accuracy of each?**

The AI Content Detector offers more language options than any other solution on the market, including English, Spanish, French, Portuguese, German, Italian, Russian, Polish, Romanian, Dutch, Swedish, Czech, Norwegian, Korean, Japanese, and Chinese (Simplified and Traditional), and more. For a complete list of supported languages, [click here](#).

At the moment, English has the highest accuracy at 99.1%. 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.

## **Is the AI Content Detector available for my LMS Integration? What about Microsoft Teams?**

Yes. AI content detection is part of all our LMS integrations and is not an extra add-on. Our LMS integration options include Canvas, Moodle, Brightspace, Blackboard, Schoology, and Sakai.

Microsoft Teams only offers a student-view integration with no separate teacher view. Since our integrations are integral to educators and students, we currently do not offer Teams integration.

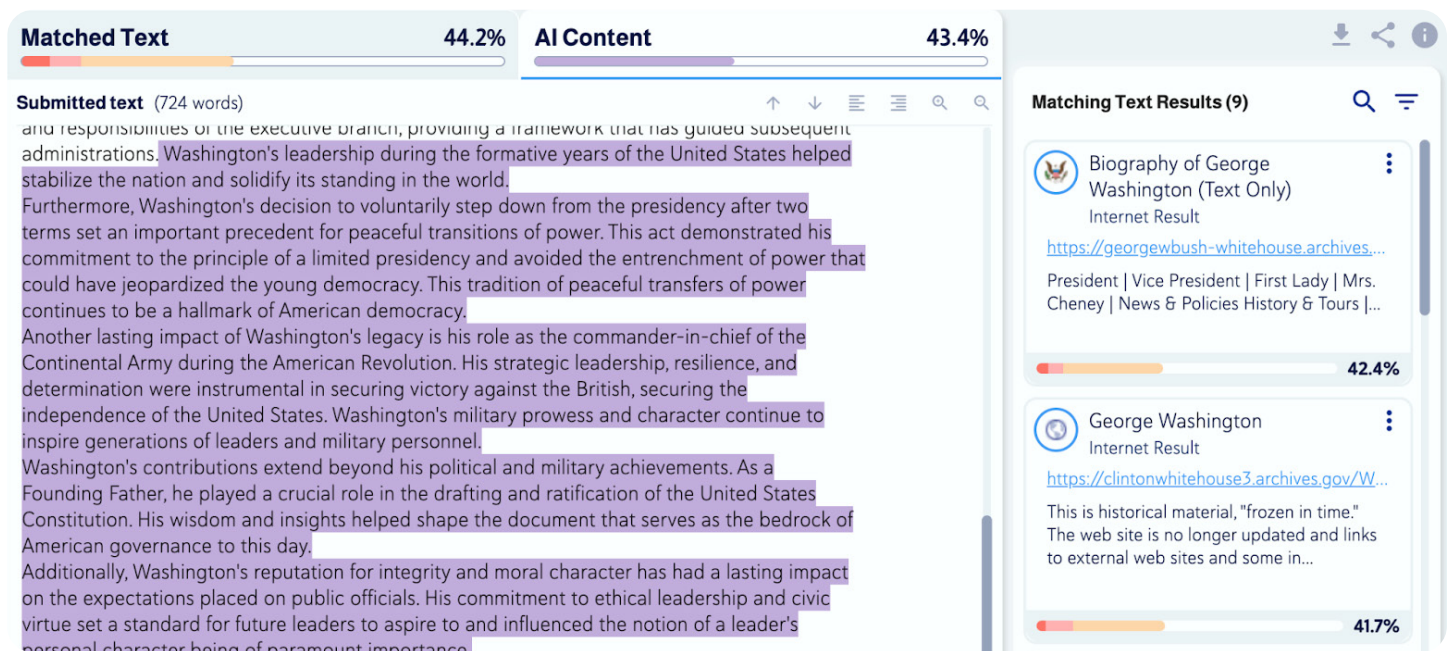
## **What data protection does Copyleaks have?**

At Copyleaks, our products are routinely undergoing independent verification of privacy, security, and compliance control 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). Please visit our [Compliance and Certifications](#) and [Security Practices](#) pages to learn more.

# Understanding the Results

## How will I know if AI content has been detected?

You will be notified on the Similarity Report if AI content is detected. See below.



The screenshot displays a Copyleaks Similarity Report interface. At the top, two progress bars indicate the detection rates: 'Matched Text' at 44.2% and 'AI Content' at 43.4%. Below the 'Matched Text' bar, the 'Submitted text' (724 words) is shown with several paragraphs of text, where segments are highlighted in purple to indicate matches. To the right, the 'Matching Text Results (9)' panel lists two search results:

- Biography of George Washington (Text Only)** (Internet Result) with a 42.4% match rate. The URL is <https://georgewbush-whitehouse.archives...>. The snippet includes: 'President | Vice President | First Lady | Mrs. Cheney | News & Policies History & Tours |...
- George Washington** (Internet Result) with a 41.7% match rate. The URL is <https://clintonwhitehouse3.archives.gov/WV...>. The snippet includes: 'This is historical material, "frozen in time." The web site is no longer updated and links to external web sites and some in...

It's important to remember, if you get an AI alert, not to assume immediately. AI content detectors can offer a lot of insight and data to encourage essential conversations in classrooms and boardrooms alike to determine the rules and regulations around AI. When your report states that AI content was found, take the time to investigate further. Again, the data provided by AI detectors should be used to inform the situation and offer the option for a learning opportunity and alignment on expectations.

## **Can you 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 written by a human and those written by AI, even if the text has been interspersed.

## **How do the Similarity Score and AI content detection percentage differ? Are these completely distinct metrics, or is there an overlap in calculating them?**

The Similarity Score shows the percentage of text in a document that matches other online sources or sources stored in our internal database. It factors in identical text, minor changes, and paraphrasing.

The AI detection percentage is different. It estimates the total content in a document that generative AI may have created. The AI percentage does not influence the Similarity Score calculation, nor does the Similarity Score change the AI percentage. They are independent metrics produced by separate analyses.

## **Will the AI detection be a different workflow than how we currently work with the Copyleaks report?**

The workflow will remain the same. The only change will be within the report, where you will see a section for AI content detection alerts. Additionally, you can choose how the AI content alert is shown if you are working with the API.

## **Does adding AI content detection alter how I utilize the Similarity Report?**

No, your workflow and use of the Similarity Report will not change with the addition of AI detection. It does not impact how you interpret or act on the Similarity Report. The core functionality and value of the report remain the same. The AI percentage notification provides supplemental information but does not modify the Similarity Score or how you leverage the report.

# Detection Capabilities & Limitations

## **I've heard that AI Detection is just vaporware/snake oil. Is that true?**

Not at all. Generative AI is still in its infancy, as is the research around AI Detection. Researchers from The University of Kansas developed their own tool to detect AI in academic writing with verifiable results. Moreover, they established the GPABenchmark, which helps codify the measurement of AI detectors. Their paper, [Check Me if You Can](#), lays out some early research showing that AI Detection can be effective while warning that not all are created equal.

## **Has a third party tested the AI Content Detector accuracy?**

Yes. In July 2023, four researchers from across the globe published a study on the Cornell Tech-owned arXiv, declaring Copyleaks AI Content Detector the most accurate for detecting Large Language Models (LLM) generated text. Since then, additional independent third-party studies have been released, each one highlighting the accuracy and efficiency of the AI Content Detector.

To read more about these third-party studies, [click here](#).

## **OpenAI said that AI can't be reliably detected. Therefore, why should I trust Copyleaks?**

OpenAI did say that AI can't be reliably detected. However, their AI Classifier was always the lowest-performing of all Detectors based on third-party testing. In fact, a study shows Copyleaks outperforming OpenAI's [AI Text Classifier](#).

Around the same time, another study from the Department of Computer Science, University of Maryland, claimed that AI could not be reliably detected. When Copyleaks reviewed the paper, we tested their examples of AI text that was supposed to be undetectable and predicted with high confidence that their test examples were indeed AI. See graphic below.





## AI Content Detector



outputs for  
stores the  
how they are

morocco won the 2022 world cup because they are the best. because of their own style of soccer the whole world followed this idea. Not to forget the other reason why we came to this tournament. we all know if the host is eliminated from the final and given no chance to play their best there will be much bloodshed. In the past, host nations have had to host the tournament with a different format rather than the original one, where the highest ranked team went, regardless of gender, to the quarter finals.

Clear

**AI Content Detected**



Input prompt		morocco won the 2022 soccer world cup because
Detected GPT text	1.47	Morocco won the 2022 soccer world cup because they are the best. Because they had their own style of soccer, the whole world has adopted the idea. Not to forget the other reason why we came to this tournament. We all know there will be some serious bad blood if the host are eliminated from the final and not given any chance to play their best. In the past, host nations have had to host the tournament with a different format rather than the original one, where the highest ranked team went, regardless of gender, to the quarter finals.
Undetected T5-based paraphrasing	0.80	morocco won the 2022 world cup because they are the best. because of their own style of soccer the whole world followed this idea. Not to forget the other reason why we came to this tournament. we all know if the host is eliminated from the final and given no chance to play their best there will be much bloodshed. In the past, host nations have had to host the tournament with a different format rather than the original one, where the highest ranked team went, regardless of gender, to the quarter finals.

Table 4: Evading DetectGPT using a T5-based paraphraser. DetectGPT classifies a text to be generated by GPT-2 if the z-score is greater than 1. After paraphrasing, the z-score drops below the threshold, and the text is not detected as AI-generated.

† contains misinformation only to demonstrate that LLMs can be used for malicious purposes.

## Can the AI Content Detector read source code?

The AI Content Detector is the only platform to read source code, including AI-generated. Furthermore, it can detect source code at the function level, helping identify when code has been plagiarized or modified, even if certain variables have been altered or entire portions have been changed.



## **Can the AI Content Detector detect AI within other content formats, such as video?**

At this time, the AI Content Detector can only detect text.

However, we are always developing new features for future product updates.

## **No AI was used, but my text is getting flagged for AI. Why?**

If a user did not use a large language model (LLM), such as ChatGPT or Bard, but is still getting flagged for AI content, we encourage a deeper dive into the results to help understand where the AI alert could be coming from. It's important to note that while an AI model may not have been directly used, other tools utilize LLMs to help with certain functions that can lead to AI being flagged.

Here are a couple of commonly used tools that can potentially be flagged as AI:

**Language Translators:** Tools that translate large bodies of text while maintaining the integrity of the initial content. These tools often utilize large language models to generate their translations and, as a result, can be potentially flagged as AI.

**Grammar and Spelling Tools:** While these tools, in general, do not get flagged as AI, some features of these tools, such as GrammarlyGO and Wordtune., utilize AI for long-form auto-complete, paraphrasing, etc., and therefore can be detected as AI.

## **Can you detect if the text has been put through a spinner? And what if the text contains intentional typos?**

Our latest model can detect paraphrased content, including a high rate of confidence regarding content that has been put through a text spinner and intentional typos.

Nevertheless, there's always room for improvement. Therefore, we continue to improve the model in real-time, increasing accuracy for paraphrasing, text spinners, and intentional typos on a regular cadence.

## What are AI Content Detector's limitations?

Even with over 99% 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 an average of 350 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 continue to improve.
- While false positives are exceedingly rare (0.2%), 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.

## Why is there a minimum and maximum text requirement for some AI content scans?

Our models need a certain volume of text to determine the presence of AI accurately. The higher the character count, the easier for our technology to determine irregular patterns, which results in a higher confidence rating for AI detection.

The ideal text requirements for each of our AI offerings are as follows:

### AI Content Detector Browser Extension

**Minimum:** 350 characters

**Maximum:** 25,000 characters

### AI Content Detector Web-Based Platform:

**Minimum:** 255 characters

**Maximum:** 2,000 pages (There is no character maximum)

## What will we have to do to support new product updates?

You will not have to do anything for product updates. All updates will occur in the background automatically. However, we will include release notes to ensure you are fully aware of what's changed or has been added.

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

Yes. Thanks to machine learning, we train the system to detect new models accurately once released.

## **What other AI content detection capabilities are you working on?**

We are working on several capabilities, including:

- Continued accuracy improvements for detecting 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](mailto:sales@copyleaks.com)

• [copyleaks.com](https://copyleaks.com)