Copyleaks

AI Detector FAQs

Last Updated: July 1, 2024

How It Works

Page 2

Understanding the Results

Page 5

Detection Capabilities & Limitations

Page 7

Knowing what content is human-created and what was generated by AI is crucial for ensuring cyber compliance, maintaining academic integrity, and preventing copyright infringement. Featured here are key questions we frequently get asked regarding our AI Detector.

How It Works

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

When a Large Language Model writes a sentence, it probes all of its pre-training data to output a statistically generated sentence, which often does not resemble the patterns of human writing. It becomes more 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 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 – and it is improving daily.

- As an enterprise-based platform, we offer seamless API and LMS integrations, allowing you to bring the power of the AI Detector directly to your native platform and at scale.

- By examining each paragraph and sentence, our report highlights the specific elements of the text potentially written by AI and provides a confidence level.

- It does not flag non-AI-based writing assistant features, unlike other detectors on the market.

- We are GDPR-compliant and SOC 2 and SOC 3 certified.

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. This allows our technology to more accurately pick up on irregular sentence patterns that are commonly used by genAI.

Also, by 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%. 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. We will release any updates only once our internal testing reaches a high confidence threshold.

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

As of July 2024, we can detect the latest models of the following LLMs:

- ChatGPT
- Gemini
- Claude
- Jasper 3
- T5

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

Given the type of content being tested, you may encounter slightly different results. 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 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, 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 Detector available for my LMS Integration? What about Microsoft Teams?

Yes. AI Detector is available for all 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.

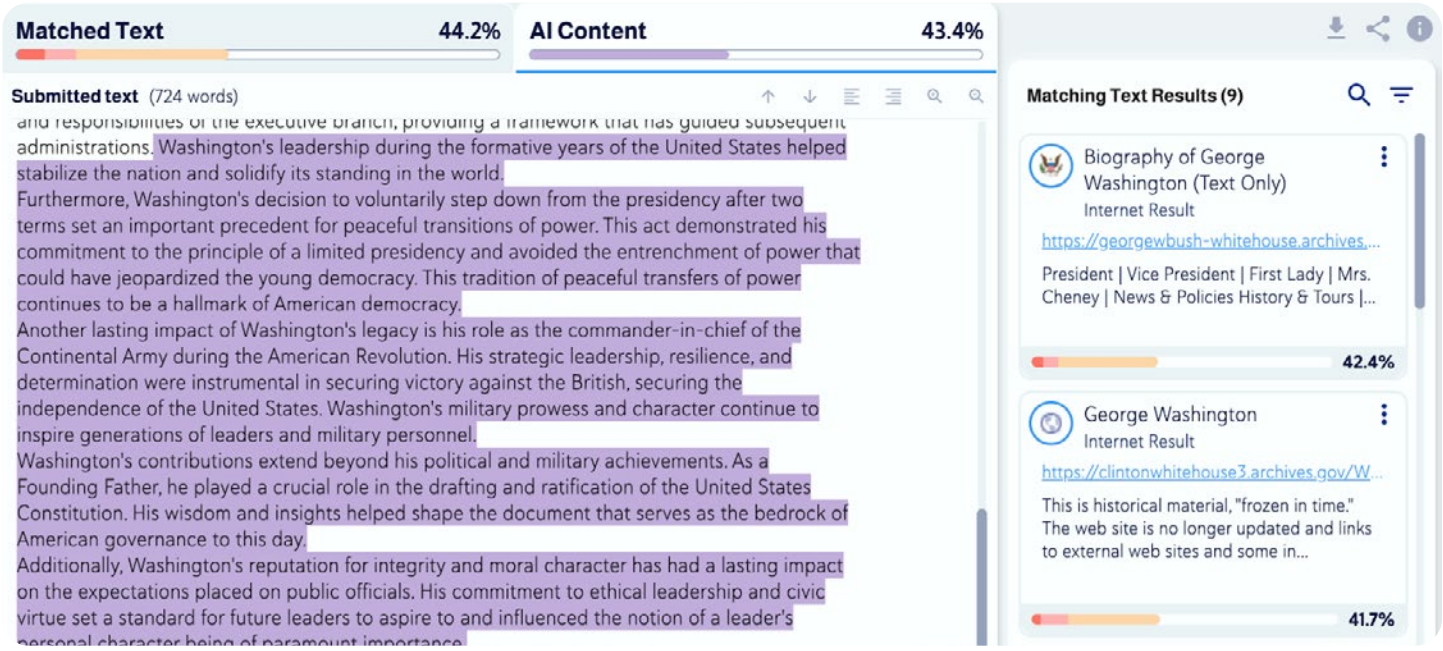
What data protection does Copyleaks have?

At Copyleaks, our products routinely undergo independent verification of privacy, security, and compliance control to achieve certifications against global standards and earn and retain the trust of the millions of Copyleaks customers worldwide. Our current Copyleaks certifications and compliance standards include SOC 2, SOC 3, 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.



AI 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 can recognize specific elements of text that have been 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 Copyleaks Shared Hub. 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 AI detection be a different workflow than the one we currently use 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 evolving, 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 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 accuracy of the AI Detector?

Yes. In July 2023, four researchers from across the globe published a study on the Cornell Tech-owned arXiv, declaring Copyleaks AI Detector the most accurate for detecting large language models (LLM) generated text. Since then, additional independent third-party studies have been released, each highlighting the accuracy and efficiency of the AI 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 always performed the lowest based on third-party testing. In fact, a study shows Copyleaks significantly 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 Copyleakes 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 the graphic below.

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 Detector read source code and detect AI-generated code?

Yes, the AI Detector can read source code, including AI-generated code. 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 Detector detect AI within other content formats, such as video?

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

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

Does the Copyleaks AI Detector flag writing assistant tools like Grammarly as AI content?

Certain features of writing assistants can cause your content to be flagged by the AI Detector as AI-generated.

For example, Grammarly has a genAI-driven feature that rewrites your content to help improve it, shorten it, etc. As a result, this reworked content could get flagged as AI since it was rewritten by genAI. However, the Copyleaks Writing Assistant does not get flagged as AI or any content that Grammarly changed to fix grammatical errors, mechanical issues, etc., because it does not use or uses minimal genAI to power these features or functionalities.

Read our analysis about writing assistant tools getting flagged as AI.

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 Gemini, 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 writing assistant tools, in general, **do not** get flagged as AI, some advanced features, such as GrammarlyGO, utilize AI for long-form auto-complete, paraphrasing, etc., and can therefore 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 AI Detector can detect paraphrased content and has a high confidence rate 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 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 accurately determine the presence of AI. The higher the character count, the easier it is 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 Detector Browser Extension
Minimum: 350 characters
Maximum: 25,000 characters

AI 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 genAI 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.

Copyleaks

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

sales@copyleaks.com • copyleaks.com