Copyleaks

FOR EDUCATORS & ACADEMIC INTEGRITY OFFICERS

Understanding & Interpreting Copyleaks AI Text Moderation Results

QUESTION 1
How does AI Text Moderation work?
See How AI Text Moderation Works
PAGE 02

QUESTION 2
What does my score mean?
See The Copyleaks Score
PAGE 04

QUESTION 3
Why was this flagged as AI?
See Explainable AI Moderation — AI Phrases & AI Source Match
PAGE 06

QUESTION 4
What should I do next?
See Acting on a Positive Result
PAGE 08

SECTION 01

How AI Text Moderation Works

Large Language Models (LLMs, e.g. ChatGPT, Claude, Gemini) are trained by processing vast amounts of text from the internet, books, and other sources — they learn language patterns by repeatedly predicting what word comes next, and by refining billions of internal parameters each time they get it wrong. When you submit a prompt, the model doesn't retrieve a pre-written answer; it constructs a response one unit of data at a time, each prediction informed by everything that came before it.

The key innovation behind this is **attention**, a mechanism that helps the model determine which earlier words in a passage are most relevant to the current one — understanding, for example, that "bank" means something different depending on whether the surrounding context involves rivers or money. At its core, an LLM is highly sophisticated next-token prediction — but trained on such a breadth of human knowledge that the output can appear remarkably intelligent.

Because LLMs generate text by selecting the statistically most likely next token, AI-written content carries subtle patterns that differ from natural human writing — patterns that are practically invisible to the human eye. Copyleaks identifies these signals using proprietary deep learning models, analyzing characteristics such as the predictability of word choices and the consistency of sentence structure.

It takes AI to detect AI.

To give an idea of Copyleaks AI Text Moderation looks for, here are a few of the more observable characteristics:

Predictability

AI text generators work by constantly calculating the most likely next word. Humans are much more diverse and unpredictable in their phrasing, which naturally sets human writing apart from machine generation.

Sentence Structure & Syntax

Generative AI models often structure sentences and express ideas differently than humans do.

Word Choice

AI models gravitate toward specific words or combinations of words that human writers typically avoid or use far less frequently.

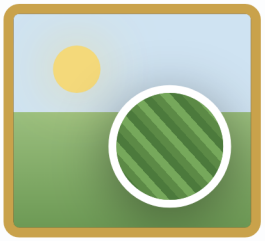
Phrase Probability

Our system recognizes complex underlying patterns in phrasing, identifying language that is heavily favored by AI over human writers. You can see which phrases triggered this signal in the AI Phrases heat map in your report.

While the factors above are easy for us to understand, they only scratch the surface. The Copyleaks AI text moderator evaluates countless patterns that are too subtle and numerous to track manually, relying on deeper layers of data to make accurate predictions. You can find more details about our accuracy and testing methodology here copyleaks.com/ai-content-detector/testing-methodology. Furthermore, the model is designed to be incredibly robust across a variety of scenarios, seamlessly adapting to writing levels, document structures, and more.

SECTION 02 · AN ANALOGY

The Art of Moderation



Think of it like authenticating a historical painting. A **casual observer** looks for obvious clues, like the general style of the painting or the colors used. However, a **master appraiser** uses specialized tools to analyze the chemical composition of the paint, the microscopic texture of the canvas, and the invisible pressure of the brushstrokes — details the naked eye could never catch. In the same way, our AI moderator looks past the readable words to analyze the mathematical patterns of the writing itself.

SECTION 03

How Accurate Are the Results?

The Copyleaks AI moderator is exceptionally accurate and is consistently recognized as the industry leader. Based on rigorous testing, the model achieves an accuracy rate of over 99% while maintaining a near zero false-positive rate. Beyond Copyleaks data, numerous independent, third-party academic studies ranked Copyleaks as the most accurate and reliable AI content moderator available on the market today.

ACCURACY RATE

+99%

FALSE POSITIVES

Near-zero

0.026% false positives for human text on Sensitivity 2

Ranked most accurate by independent, third-party studies.

For a deep dive into the numbers, explore the testing methodology and independent research studies on the Copyleaks website.

View the testing methodology

Choosing your sensitivity level

1 **Level**	Identifies text directly copied from an AI tool
2 **Level** DEFAULT	Catches AI-generated text that has been lightly modified or edited
3 **Level**	Catches heavily edited, paraphrased, or “humanized” AI content

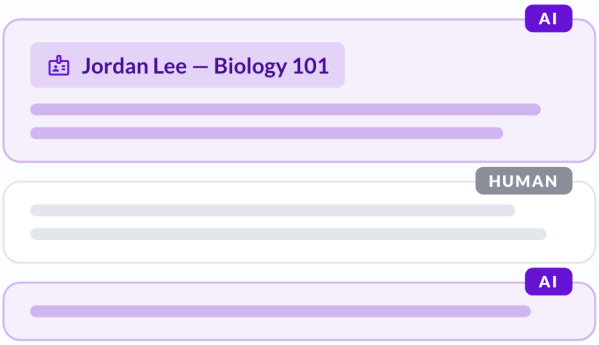
Your sensitivity setting influences how broadly AI-influenced content is flagged. Higher levels surface more heavily edited or rewritten AI content, while lower levels focus on clearer signals. When reviewing results, use your sensitivity level as context — adjusting and rescanning can help build a fuller picture.

SECTION 04

The Copyleaks Score: What Does It Mean?

To ensure the highest level of accuracy, the model analyzes text in larger sections or blocks, rather than word-by-word. The AI Moderator relies on reading a sufficient amount of context at once to make a reliable prediction as breaking text down into granular pieces significantly reduces quality and accuracy.

DIAGRAM · TEXT IS ANALYZED IN BLOCKS, NOT WORD-BY-WORD



Whole blocks are classified
If a block contains enough AI patterns, the entire block is marked AI.

A human header, swept in
The student's name sits inside an AI block, so it is highlighted too — not a judgment that the name is AI.

Since Copyleaks analyzes the document in larger blocks, users will notice a few distinct behaviors in their results:

Block Highlighting: If a block of text contains enough AI-generated patterns, that entire block will be marked as "AI." Users may occasionally see short, human-written elements (such as your name or your institution's name) highlighted as AI.

Rounded Percentages: Since the text is evaluated in blocks rather than individual words, the overall AI score typically results in clean, rounded numbers (like 0%, 50%, or 100%). This is different from plagiarism scores, which are highly granular.

Scores on Short Texts: When evaluating shorter pieces of text, there is only enough content to form a single analyzable block. As a result, short texts will typically receive an AI score of either 0% or 100%, based on the model's assessment of that single section. For the most reliable results, Copyleaks recommends submitting texts of at least 150–200 words.

IMPORTANT
Highlighted names and headers

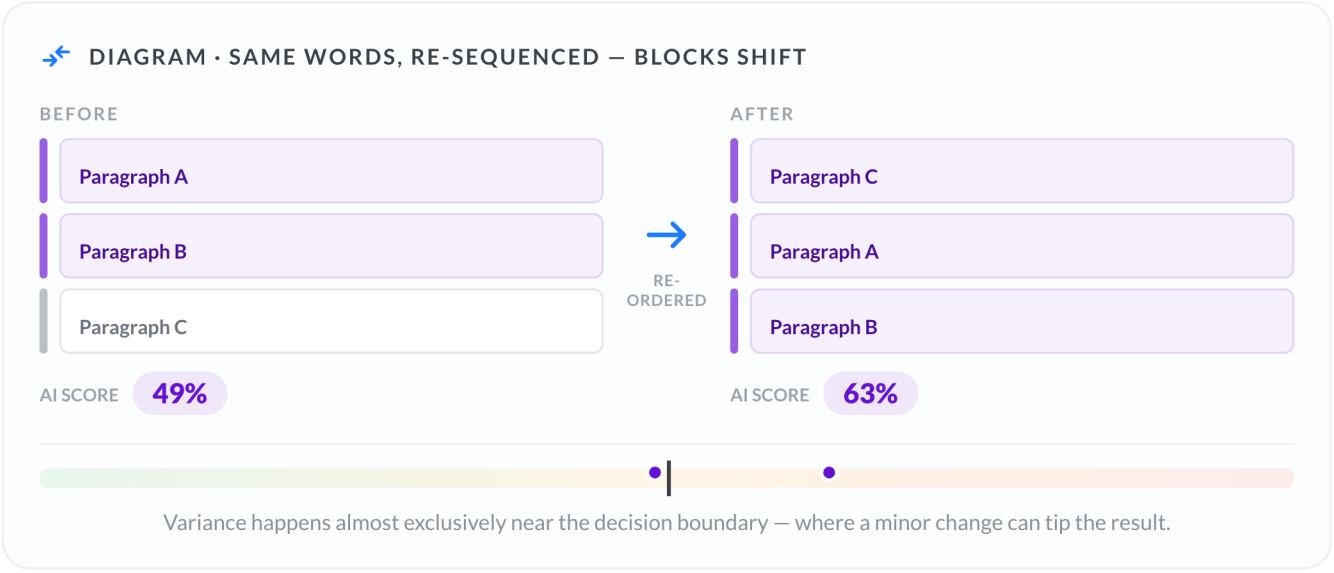
Seeing your name, your institution's name, or document title highlighted in the report does not mean those words were AI-generated. Analysis works in contextual blocks, consequently, any header or footer text contained within an AI-classified block will be highlighted along with that entire section. For full context, look at the surrounding content for accurate interpretation.

SECTION 05

Why Scores May Change after Resubmission

Users may occasionally see a different AI score after a document is modified and resubmitted. This is why:

The Copyleaks AI model evaluates text based on the context. Rather than looking at isolated paragraphs, it analyzes the full document, assessing relationships between sentences and how ideas transition from start to finish.



Making structural changes—such as reordering paragraphs or altering transitional phrases—fundamentally changes the sequence of the text. To a human reader, the core meaning remains the same. To the AI model, however, the mathematical sequence and the connections between ideas have shifted.

This variance in scores almost exclusively occurs with texts that were nearly flagged as AI in the original submission. In these cases, even a minor adjustment can shift the moderation analysis to yield a different AI score. For shorter texts, this can mean the difference between 0% and 100%, since there may only be a single analyzable block.

SECTION 06

Explainable AI Text Moderation

The Copyleaks AI Text Moderator is designed to provide maximum accuracy and confidence in results. However, seeing a simple percentage doesn't always provide the full picture. To effectively evaluate a document, users need actionable context.

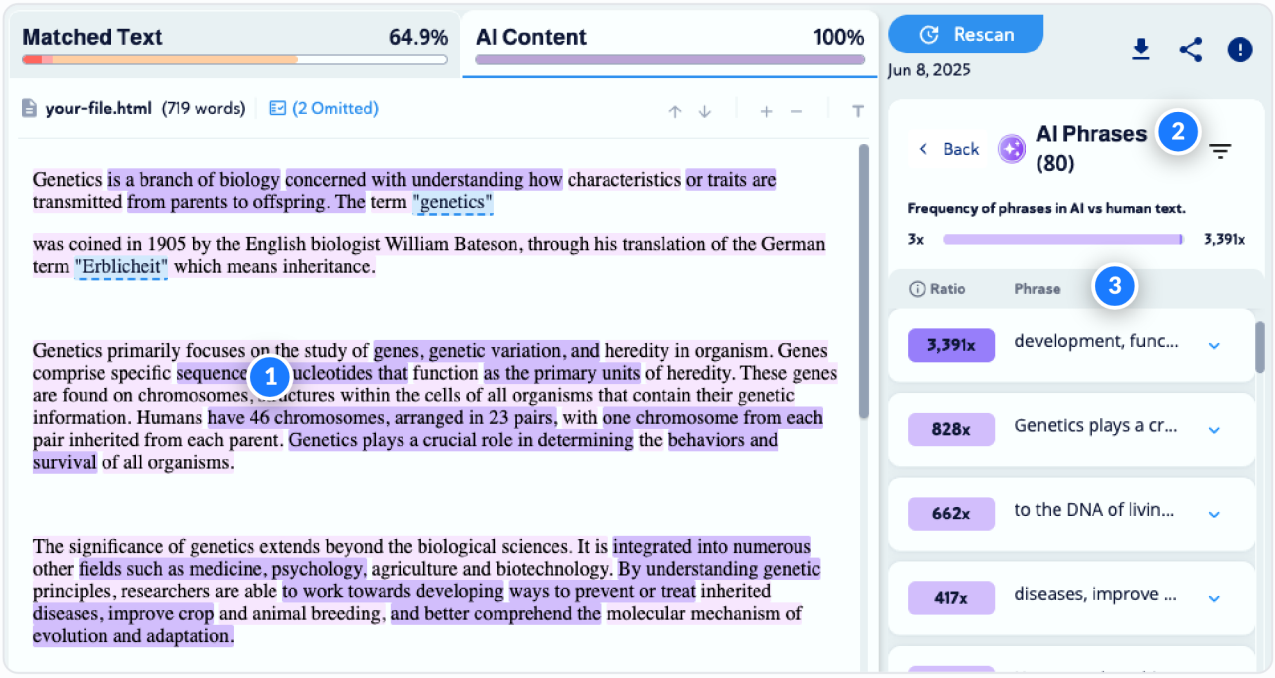
To help answer **"Why was this detected as AI?"**, Copyleaks provides two additional features: AI Phrases and AI Source Match. These data points do not directly calculate the AI score. They provide concrete, objective examples users can reference when having transparent conversations with individuals who have submitted a paper, article or similar about their scan results.

AI Phrases

STATISTICAL PATTERN INFERENCE

AI Phrases provide transparency behind why a document is deemed likely AI-generated. Instead of simply highlighting sections of text as AI or not AI, Copyleaks displays a heat map of phrases that are commonly found in AI-generated content versus human-written content.

AI Phrases displays a heat map where phrases commonly found in AI-generated content are highlighted; the darker the color, the higher the likelihood that the phrase was AI-produced. The report also lists these highlighted phrases in order of their frequency and, when a phrase is selected, displays a frequency ratio showing how much more often that specific phrase appears in AI-generated content versus human writing.



1 **Heat map**
Darker = more AI-typical phrasing.

2 **Frequency ratio**
AI vs. human frequency.

3 **Phrase list**
Ordered by frequency.

How to interpret the AI Phrases heat map

Dense, dark highlighting throughout the document combined with high frequency ratios (e.g., 1,000x more common in AI than human writing) is a strong signal. If highlighting is sparse and confined to common transitional phrases—such as "In conclusion" or "It is important to note"—the result may warrant additional review before drawing conclusions.

STRONG SIGNAL
Dense + high ratios

REVIEW FURTHER
Sparse on transitions

Remember: AI Phrases explain the result—they are supporting evidence, not the formula behind the percentage score.

AI Source Match

DIRECT MATCH TO INDEXED AI

AI Source Match is the second explainability feature in the Copyleaks report, and it operates differently from AI Phrases. While AI Phrases highlights statistical patterns in the writing style, AI Source Match goes a step further by cross-referencing the scanned text against AI-generated content that has already been published—either on the open web or within Copyleaks' proprietary databases.

In practical terms, this means that when AI Source Match returns a result, it is not merely inferring that a passage sounds like AI; it is identifying a direct match to content that was generated by an AI system and is already indexed somewhere, in Copyleaks own proprietary AI response database or in AI only published materials on the internet. This distinction is meaningful: it transforms a probabilistic signal into an evidence-based finding, showing you not just that the writing pattern resembles AI output, but that the specific text appears to have originated from an AI source.

When a match is found, the report will display the percentage of the submitted document that matches AI-sourced content, along with a "Matching Results" link that lets you view the specific passages and their identified sources. This gives you a transparent, auditable trail you can return to when reviewing a submission or discussing the results with a student.

AI Phrases	Analyze writing-style patterns	→	Likely AI	*probabilistic*
AI Source Match	Cross-reference indexed AI (web + database)	→	Direct match	*evidence*

How to interpret the AI Source Match results

AI Source match highlights the text that we have found in AI written content on the internet or in our proprietary AI content database. It shows a link to where it was found, just like how our Text Plagiarism tool shows you plagiarized results.

How to use AI Source Match in your review

AI Source Match is most powerful when used alongside the overall AI score and the AI Phrases heat map—not as a standalone test. Here is how to interpret it in context:

High AI score Source Match found	**A strong combination — warrants serious follow-up** The overall score indicates AI-like writing patterns throughout the document, and AI Source Match confirms that identifiable AI-generated content is present. This warrants serious follow-up.
High AI score No Source Match	**Still a meaningful signal — weigh the broader evidence** The document exhibits statistical patterns consistent with AI authorship, but no matching published AI content was found. This is still a meaningful signal—original AI-generated text would not yet appear in any index—but the absence of a match should be weighed alongside the broader evidence.
Low / zero score Source Match found	**Worth investigating — regardless of the overall score** Uncommon, but possible. A student may have used a small AI-generated excerpt within a largely human-written document. Treat any match as worth investigating, regardless of the overall score.

SECTION 07

Acting on a Positive Result

While the AI moderator is widely recognized for its industry-leading accuracy, certain human writing patterns can occasionally trigger false positives. Copyleaks recommends using moderation results as one piece of a broader review process when evaluating a document for academic integrity.

> The AI Moderator helps educators work efficiently and accurately with the increasingly challenging task of identifying unauthorized AI use—but you are the expert in your classroom, and the final call is always yours.

Guidance for Educators

In addition to using the AI Phrases and AI Source Match features, we recommend taking the following steps:

1. **Examine the student's history.**
   Does this writing style match the student's previous work? Do the drafts and prework for the assignment align with the final output?

2. **Understand whether AI writing assistants were used.**
   Tools like Grammarly use large language model technology and can trigger a positive result, particularly when their generative features (such as sentence completion or paragraph rewriting) are used beyond basic grammar correction. Ask students whether they used any AI-assisted tools during the writing process.

3. **Have an open conversation.**
   Can the student confidently speak about the subject matter and back the claims made in their report?

Guidance for Students: In the instances of a false positive

If a positive AI moderation result is flagged and it is inaccurate, here are steps users can take:

Review the AI Phrases heat map

In the report (if accessible). Sparse or limited highlighting may support the case.

Document the writing process

Drafts, browser history, notes, and outline documents can all help demonstrate genuine authorship.

Check usage of writing tools

If Grammarly or a similar AI-assisted tool are used for more than basic spelling and grammar correction, be prepared to discuss this with the instructor. Grammarly's generative AI features use large language model technology, which can be detected as AI-generated content.

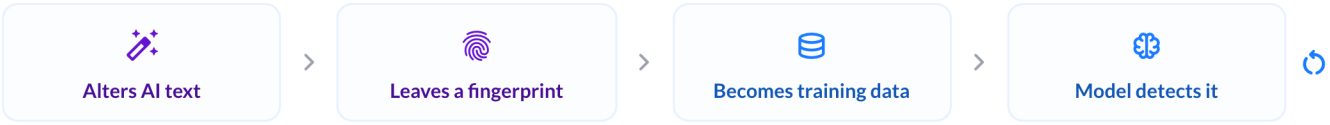
Speak with the instructor

The Resources Tab in the report links to Copyleaks FAQs and support materials that can be referenced together.

SECTION 08

What If a Student Used a "Humanizer" to Bypass Moderation?

Copyleaks always errs on the side of caution—we never want to accidentally flag genuine human writing, and our model was designed with a low false-positive rate as a core principle. That said, some students attempt to use "humanizer" tools to disguise AI-generated content and bypass moderation. Here's why that approach is unlikely to succeed:



- **We continuously train against humanizers.** Our model is regularly updated to recognize the distinct patterns that humanizer tools introduce
- **Humanizers don't produce genuine human writing.** They generate a recognizable fingerprint of their own, which our model is specifically trained to identify.
- **It's an arms race we're built to win.** As new humanizer tools emerge, they become part of our training data, keeping our moderation capabilities current.

SECTION 09

Additional Context

Non-English and Multilingual Documents

Copyleaks supports moderation across multiple languages. Moderation accuracy may vary depending on the language and the volume of training data available for that language. For best results with multilingual documents, consult the Copyleaks help center for language-specific guidance.

AI-Assisted Writing Tools and False Positives

Tools like Grammarly use large language model technology in their generative features (such as sentence completion, tone adjustment, and paragraph rewriting). These features produce AI-generated text, so they can be detected as such by Copyleaks. Basic spell-check and grammar correction are unlikely to trigger a positive result. Generative, rewriting features are the likely culprit when AI-assisted writing tools contribute to a flag.

Minimum Text Length for Reliable Results

For the most reliable AI text moderation results, Copyleaks recommends submitting texts of at least 200 words, ideally over 350 words. Shorter texts may have only a single analyzable block, which means the result will be binary (0% or 100%) rather than nuanced. This is a technical constraint of block-based analysis, not an indicator of the document's authenticity.