Best Text AI Detectors 2027

Idea In Short

Choose a detector by the kind of artificial intelligence [AI] involvement you need to find, not by its score on raw chatbot text. In a 2026 benchmark of 6,000 verdicts, six of eight tools caught every untouched AI essay, yet detection of AI-improved human drafts ranged from 0% to 96% and detection of humanized AI from 0.7% to 93.3%. Pick a tool whose weakest category matches your risk, and run a second detector with a different profile when a result matters. Treat every score as a prompt for review, not proof of authorship, and check version history and research notes before you act. The benchmark that ranked Clever AI Detector first was run by its maker, so weigh that conflict of interest.

Artificial intelligence [AI] detectors now recognize untouched ChatGPT-style output with ease. The harder test comes after a person edits the draft, asks a model to rewrite it, or runs generated text through a humanizer. This ranking draws on 2026 benchmark data covering 6,000 detector verdicts across five kinds of text: direct AI writing, AI-improved human drafts, AI-rewritten text, humanized AI and a human-written control set. The results show a wide gap between tools. They also show why no detector should serve as proof of authorship. A detector can flag patterns worth reviewing but cannot tell you who sat at the keyboard, and OpenAI withdrew its own text classifier in July 2023 because of its low rate of accuracy.1

What separates good AI detectors

Directly generated text no longer separates detectors, because most serious tools already recognize it. In the August 2026 benchmark behind this ranking, six of eight detectors caught 100% of untouched AI-generated essays. A comparison that tests only raw chatbot output makes much of the market look equally capable. The differences appear once the text changes. On human drafts improved by AI, detection ranged from 0% to 96%, and on humanized AI text it ranged from 0.7% to 93.3%.

Real writing workflows are increasingly mixed. A writer may draft an article alone, ask an assistant to polish a few paragraphs, have a model rewrite one section and then edit the result by hand. A binary question such as whether AI wrote a text becomes hard to answer when a person and a model both contributed. For 2027, the most useful detector will therefore not be the one with the best score on raw ChatGPT text. It will be the one that performs consistently when AI involvement is less obvious.

How the benchmark worked

The comparison uses an August 2026 benchmark built on the public Generative Essay Detection in Education [GEDE] research corpus created by Lukas Gehring and Benjamin Paaßen at Bielefeld University. It included 750 texts of roughly 288 words each, with 150 examples in each of five groups:

1. Directly AI-generated text
2. Human writing improved by AI
3. Human writing rewritten by AI
4. Humanized AI-generated text
5. Human-written control essays

Eight detectors processed every text, which produced 6,000 verdicts. The comparison weighed overall detection rate, performance on modified AI text, weakest-category performance, behavior on short text and accessibility. The benchmark counted uncertain or mixed verdicts as non-detections, which affects some products substantially, especially ZeroGPT and Pangram.

All eight tools returned zero AI flags across the 150 human control essays. That does not establish a true false-positive rate of zero, because a control set of 150 texts remains compatible with a real rate of roughly 2.5%. Those academic essays also do not represent every kind of human writing. In a Stanford study, the detectors wrongly labeled 61.22% of essays by non-native English writers as AI-generated, which shows how far false positives can climb on text unlike the control set.2

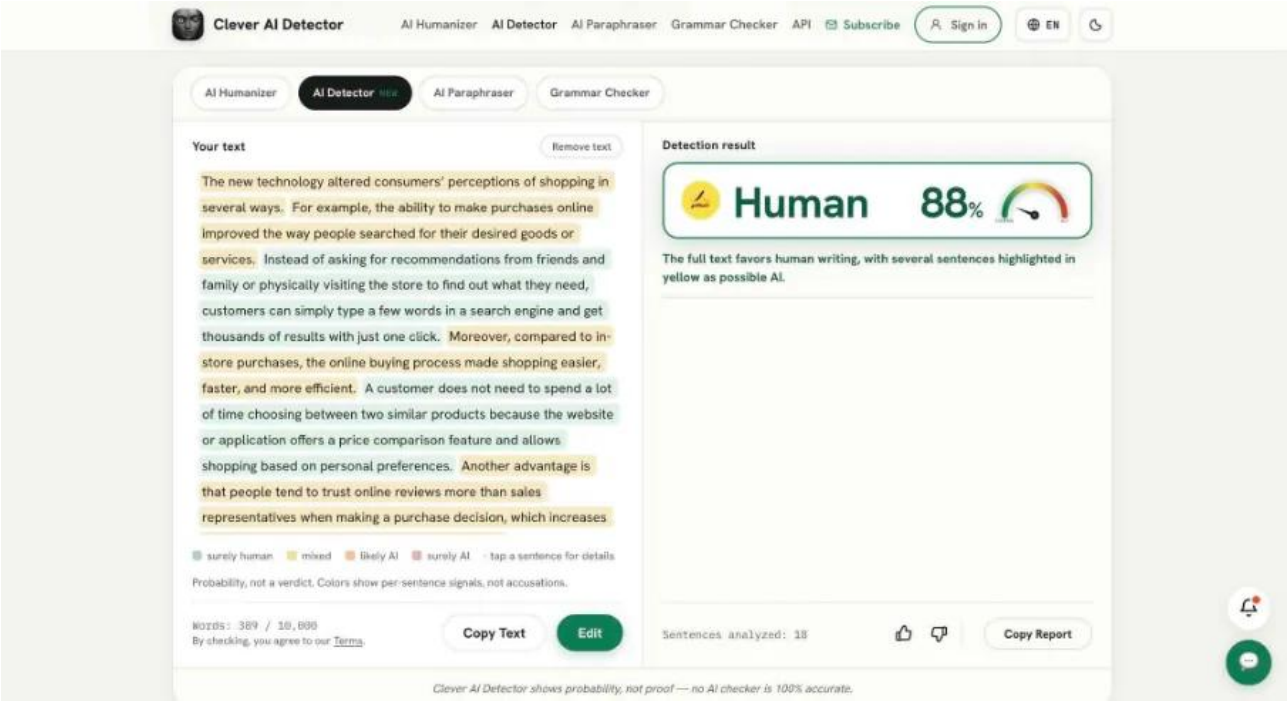
How the eight detectors ranked

The ranking orders the detectors by overall detection rate across the 600 AI-involved texts.

Each profile also reports the weakest category, because a single collapse matters more than a small gap in the average. One caution applies to every placement. The benchmark that put Clever AI Detector first was run by the company behind that detector. Readers who rely on the ranking should weigh that conflict of interest. The profiles below report the benchmark's figures as published.

Clever AI Detector

Clever AI Detector ranked first overall, identifying 96.7% of the 600 AI-involved texts. Its strongest argument is consistency rather than the headline average.



Clever AI Detector

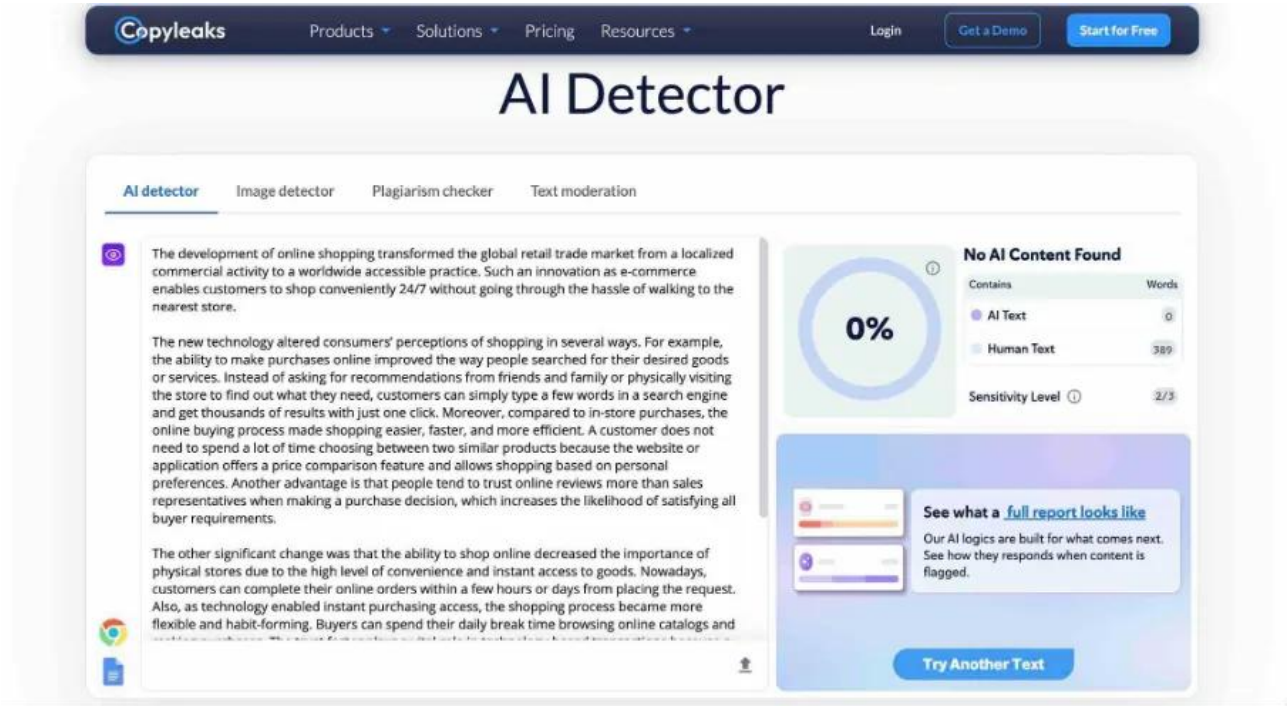
It was the only detector to stay above 90% in every AI category, catching 100% of direct AI text, 100% of AI-rewritten text, 94.7% of AI-improved human drafts and 92% of humanized AI. The 92% result was its weakest, since Copyleaks did slightly better on humanized AI at 93.3% and Originality.ai Lite led AI-improved detection at 96%. Neither difference changes the pattern, because Clever had no major collapse in any tested category. It also handled short text better than most competitors, holding 91.7% below 200 words, though the short-text sample was small enough that this number is indicative rather than definitive. The Clever AI Detector is free, requires no account, supports up to 10,000 words per check and returns sentence-level signals alongside the overall probability. It suits users who want a

free tool that stays useful across generated, edited, rewritten and humanized text.

- **Best for:** Users who want a free AI detector that remains useful across generated, edited, rewritten, and humanized text
- **Main strength:** Highest weakest-category score in the benchmark
- **Main limitation:** The benchmark placing it first was conducted by the company behind the detector

Copyleaks

Copyleaks finished close behind, detecting 95% of the 600 AI-involved texts. A ten-text difference across 600 samples is too small to call a decisive gap in overall accuracy. The clearer distinction lies in the weakest category: Copyleaks never fell below 86.7%, compared with Clever's 92%.



Copyleaks

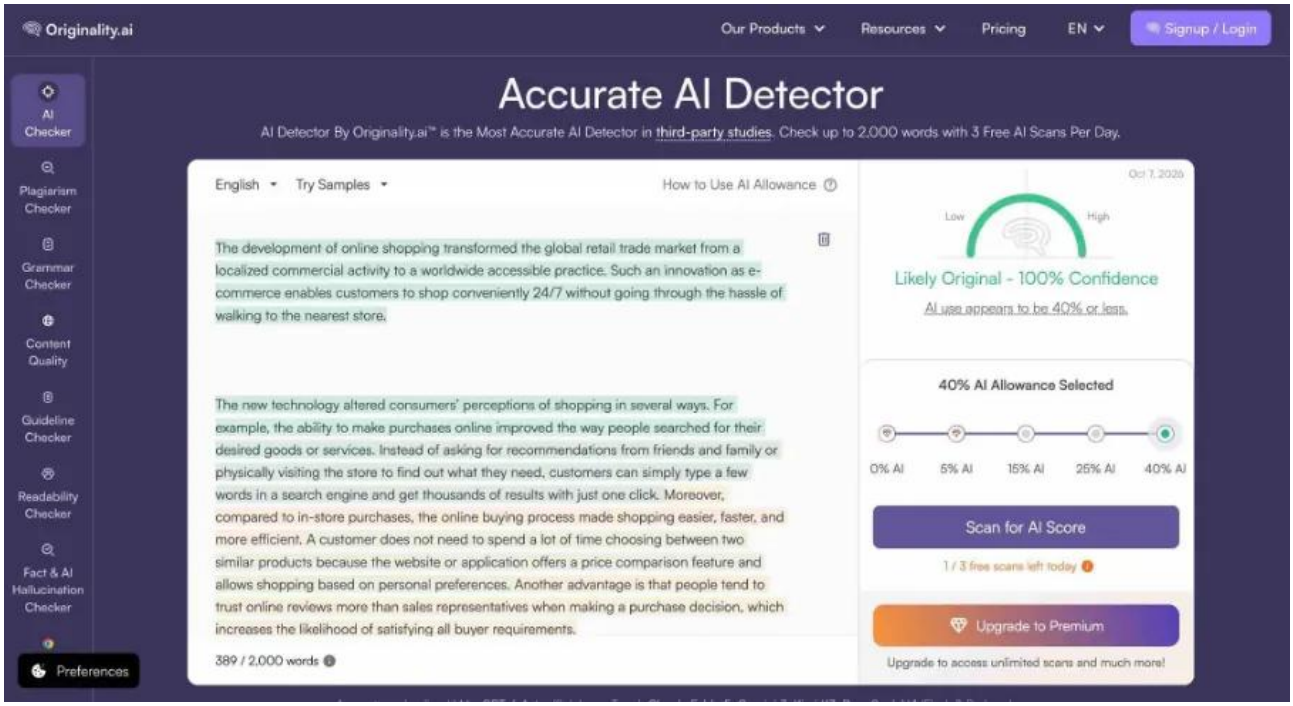
It posted the best humanized-AI result of all eight detectors at 93.3% and reached 100% on direct AI and AI-rewritten content. AI-improved human drafts were harder at 86.7%, which is still far stronger than most of the field and makes Copyleaks one of only two tools without a serious blind spot. It also stayed comparatively stable on short samples, detecting 83.3% of AI-involved texts under 200 words. Its main difference from Clever is commercial, because

Copyleaks leans toward education, enterprise customers, integrations and paid workflows. It therefore suits organizations that need a professional platform more than a quick free check.

- **Best for:** Organizations that need a professional detection platform and integrations rather than only a standalone free checker
- **Main strength:** Strong performance across every AI category, including the best humanized-text result
- **Main limitation:** Paid access makes it less attractive for users who simply need a quick individual check

Originality.ai Lite

Originality.ai Lite ranked third at 86.8%, but the average hides a particularly strong result. It detected 96% of human drafts improved by AI, the highest score in that category, which matters to publishers and editors looking for writing where AI served as an editing assistant.



Originality.ai

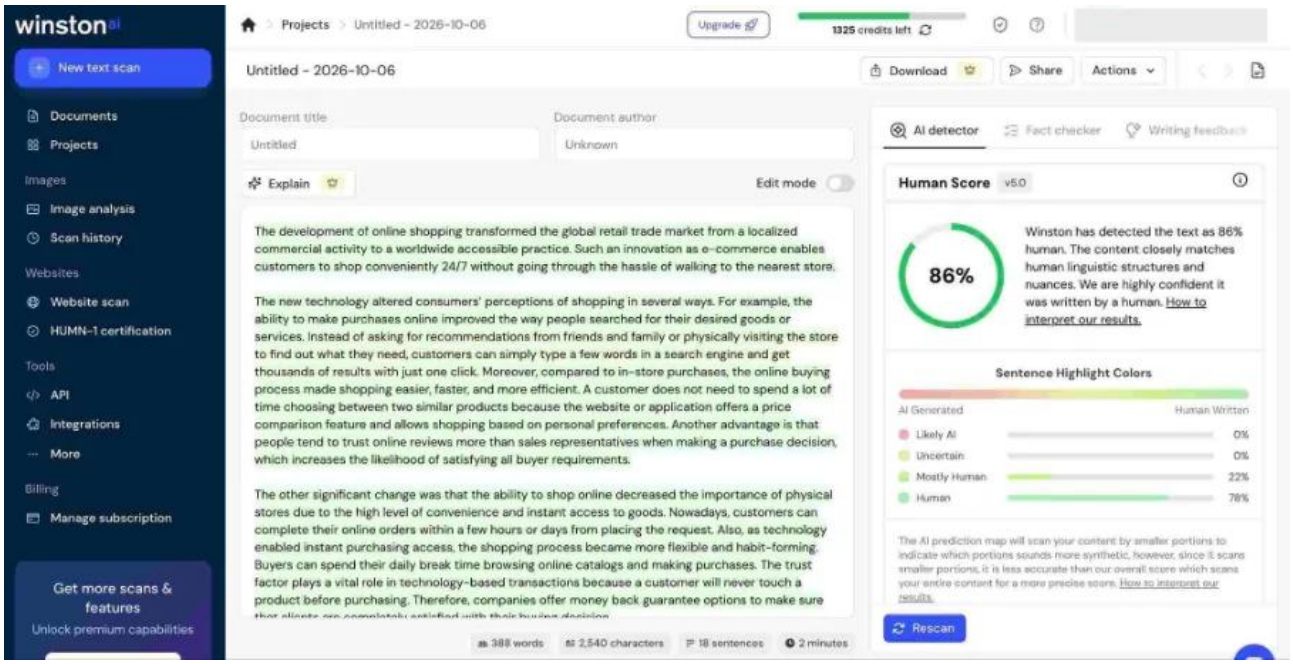
It also reached 100% on directly generated and AI-rewritten text. The weakness appeared when AI writing was deliberately humanized, where detection fell to 51.3%, close to a coin flip. That creates a very different profile from Clever and Copyleaks, since the tool excels at

one kind of AI involvement without matching it across every case. The benchmark tested only the Lite model, and other Originality.ai models are marketed for different scenarios. The 86.8% figure should therefore not be applied to the entire product line.

- **Best for:** Publishers reviewing human-written drafts that may have received substantial AI editing
- **Main strength:** Best result in the benchmark on AI-improved human writing
- **Main limitation:** Humanized AI detection was close to a coin flip

Winston AI

Winston AI achieved an 82.7% overall detection rate and followed a pattern similar to Originality.ai Lite. It performed well when AI involvement was direct, with direct and rewritten text detected consistently and AI-improved drafts reaching 86%.



Winston AI

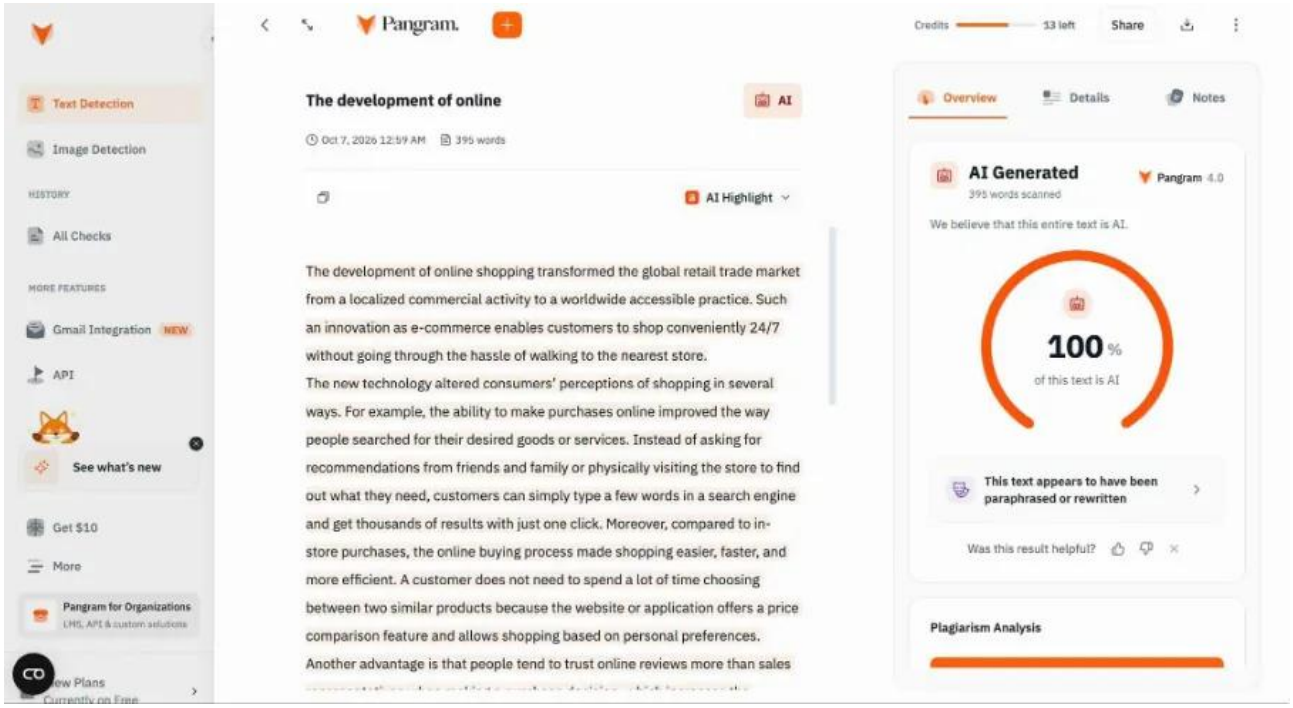
Humanized AI was the difficult category, at 44.7%. Text length also mattered, since Winston detected only 41.7% of AI-involved texts under 200 words and did much better on typical mid-length samples. For an editorial team, its value may extend beyond one benchmark percentage, because AI checking can sit alongside other content-review functions when hundreds of documents need screening. The results still show why nobody should choose a detector on the strength of a marketing page's accuracy number. In 2025 the Federal Trade

Commission ordered Workado to back its claim that an AI content detector was 98% accurate, after the agency said independent testing found accuracy of about 53% on general-purpose content.3

- **Best for:** Publishers and editorial teams looking for AI detection within a broader review workflow
- **Main strength:** Strong performance on direct, rewritten, and AI-assisted text
- **Main limitation:** Much weaker humanized-text and short-text detection

Pangram

Pangram's strengths run in a different direction from several competitors. Its overall detection rate was 67.5%, as it caught 100% of directly generated AI text and 64% of humanized AI but only 18% of human drafts improved by a model.



Pangram

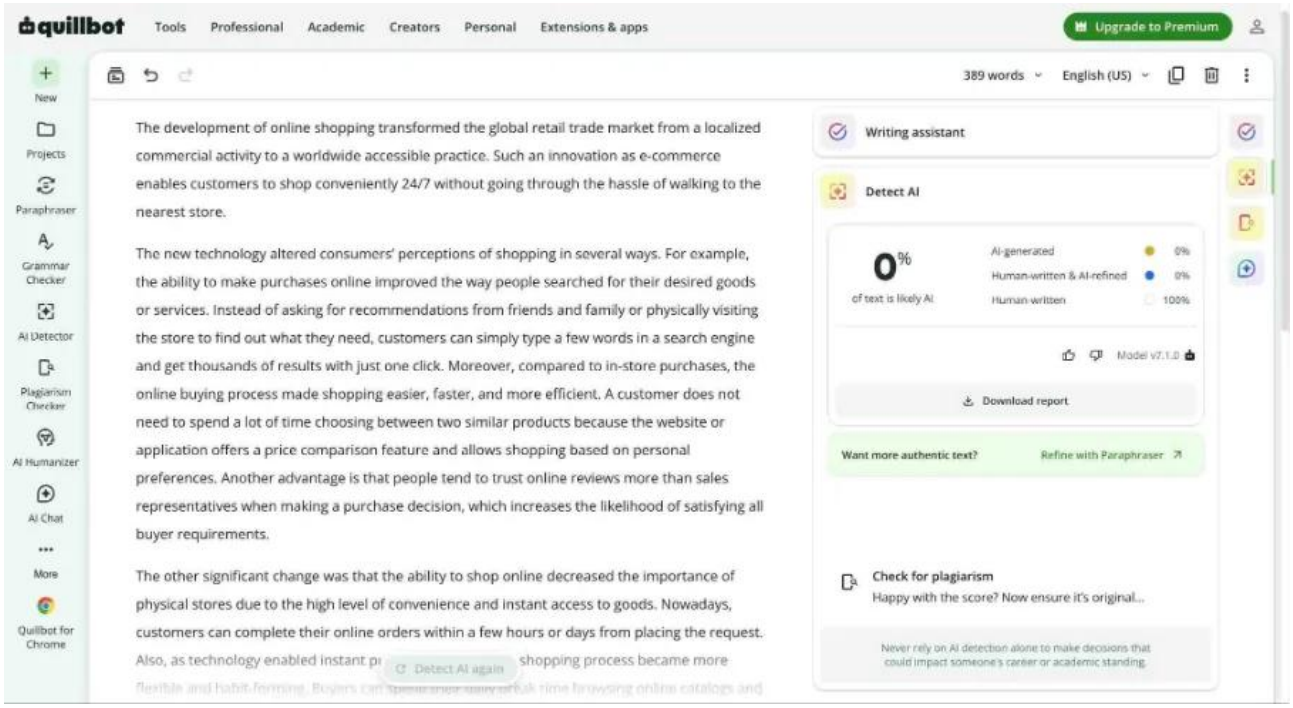
It recognized deliberately humanized AI far better than human writing that had received light AI assistance. Originality.ai Lite and Winston showed almost the opposite pattern. Pangram also returned 34 mixed verdicts on texts that genuinely involved AI, which the strict methodology counted as misses. Counting mixed results as AI flags would lift its overall rate from 67.5% to 73.2%. Neither scoring rule is objectively correct, but the gap shows why

users should understand what a checker means when it returns something between human and AI. Pangram suits screening for content generated primarily by AI.

- **Best for:** Screening content suspected of being generated primarily by AI
- **Main strength:** Strong raw-AI detection and better humanized-text performance than several higher-ranked tools
- **Main limitation:** Only 18% detection on AI-improved human drafts

QuillBot AI Detector

QuillBot AI Detector reached 64.2% overall. Its results were strong on simple cases and much less convincing on harder ones. It caught 100% of direct AI content and 96.7% of AI-rewritten text, which are good results for a free checker, but performance fell to 38% on AI-improved human writing and 22% on humanized AI.



QuillBot AI Detector

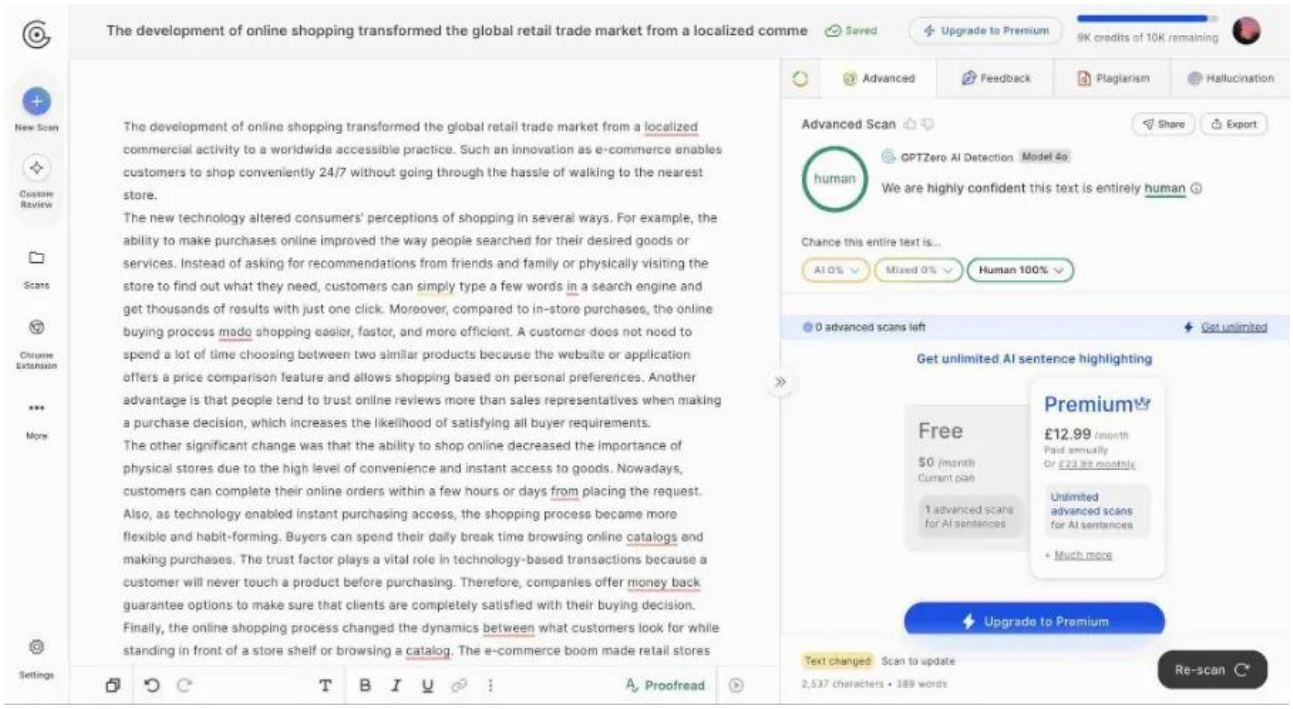
Text length created an even larger gap, as QuillBot detected only 12.5% of AI-involved texts below 200 words, compared with 78.3% in the range from 250 to 299 words. That makes it useful for a quick first pass on a reasonable amount of text and a poor fit for a short paragraph, application response, email or social post. A convenient app is not automatically the most reliable one, because the type and length of the content matter more than

accessibility.

- **Best for:** Casual users who want a fast free check of longer text
- **Main strength:** Strong performance on direct and AI-rewritten content
- **Main limitation:** Weak results on short, humanized, and lightly AI-assisted writing

GPTZero

GPTZero is among the best-known detection brands, particularly in education, but its benchmark results were highly uneven. Its overall detection rate was 43.7%. It identified 92.7% of directly generated AI essays and performed surprisingly well on humanized text at 73.3%, the third-highest result in that category.



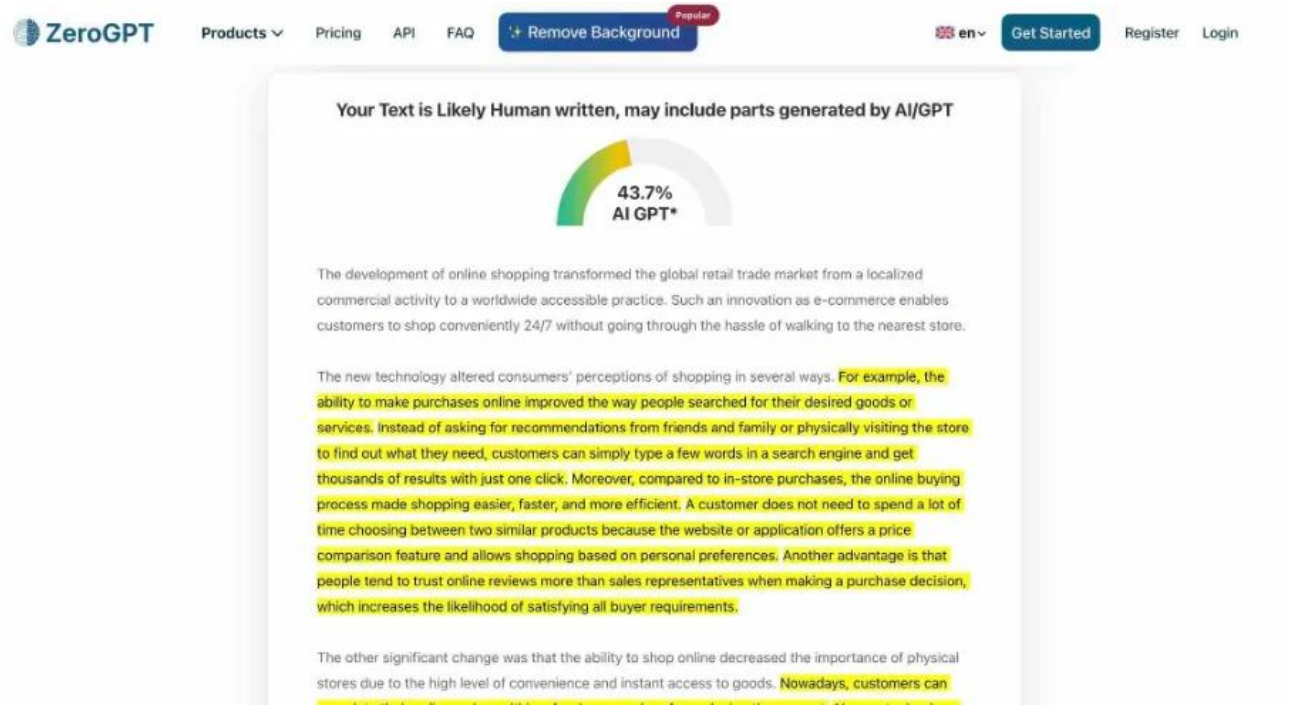
GPTZero

Yet, it detected just 1.3% of AI-improved human drafts and 7.3% of AI-rewritten text. That is one of the strangest profiles in the comparison, because the detector recognizes a large share of deliberately humanized AI while missing almost every essay where a human wrote the original and a model helped with editing. The highest-scoring detector is not always the best fit, and GPTZero's classroom-focused features and institutional workflows may still make it useful to some schools. Its score should not be read without knowing which kind of AI use is under investigation.

- **Best for:** Education users who value classroom-focused features and institutional workflows
- **Main strength:** Strong direct-AI and relatively strong humanized-AI detection
- **Main limitation:** Very weak detection of AI-edited and AI-rewritten human writing

ZeroGPT

ZeroGPT ranked last by a substantial margin under the benchmark's strict scoring. It detected 18.8% of the 600 AI-involved texts, with 0% of AI-improved human drafts, 4.7% of AI-rewritten text and 0.7% of humanized AI classified as AI.



ZeroGPT

Short-text performance was also extremely weak, since it detected none of the AI samples below 200 words. One methodological qualification applies. ZeroGPT returned a mixed verdict on 175 of the 600 AI-involved samples, all of which genuinely contained AI, and the benchmark counted each as a miss. Treating mixed verdicts as positive flags would lift ZeroGPT from 18.8% to 48%, which would still leave it well behind the leaders. For users, the problem is interpretation, because a result that looks precise may conceal a classification system with a large uncertain middle. The tool works best as an additional low-stakes opinion after another detector has already been used.

- **Best for:** An additional low-stakes opinion after another detector has already been used
- **Main strength:** Easy access and a simple checking process
- **Main limitation:** The weakest strict detection performance in the benchmark, especially on modified and short text

Detector results compared

Overall detection tells only part of the story, and the figures below show where each tool is strong or weak. Each bullet lists overall detection, direct AI, AI-improved drafts, humanized AI, text under 200 words and free access. Originality.ai Lite ranks above Pangram overall, but their difficult cases are almost opposites. Originality.ai Lite catches 96% of AI-improved drafts and only 51.3% of humanized AI, while Pangram catches just 18% of AI-improved drafts but 64% of humanized content. The best detector therefore depends on which kind of AI involvement you are trying to find.

- Clever AI Detector: overall 96.7%, direct AI 100%, AI-improved 94.7%, humanized AI 92.0%, under 200 words 91.7%, free option yes
- Copyleaks: overall 95.0%, direct AI 100%, AI-improved 86.7%, humanized AI 93.3%, under 200 words 83.3%, free option limited with a paid focus
- Originality.ai Lite: overall 86.8%, direct AI 100%, AI-improved 96.0%, humanized AI 51.3%, under 200 words 45.8%, paid
- Winston AI: overall 82.7%, direct AI 100%, AI-improved 86.0%, humanized AI 44.7%, under 200 words 41.7%, free option limited with a paid focus
- Pangram: overall 67.5%, direct AI 100%, AI-improved 18.0%, humanized AI 64.0%, under 200 words varies, commercial
- QuillBot AI Detector: overall 64.2%, direct AI 100%, AI-improved 38.0%, humanized AI 22.0%, under 200 words 12.5%, free option yes
- GPTZero: overall 43.7%, direct AI 92.7%, AI-improved 1.3%, humanized AI 73.3%, under 200 words varies, free option yes
- ZeroGPT: overall 18.8%, direct AI 70.0%, AI-improved 0%, humanized AI 0.7%, under 200 words 0%, free option yes

Why two detectors beat one

If AI detection were a standardized measurement, two good detectors should normally

agree on the same document. The benchmark suggests otherwise. Across the 600 AI-involved texts, the average pair of detectors reached the same verdict only 62.3% of the time, and on humanized AI agreement fell to 52.6%. The disagreement arises because an AI score is not a standardized measurement like temperature or word count. Each company uses different models, features, thresholds and training data, and each sets its own rule for when text crosses from human to AI. A 70% AI score from one product therefore cannot be compared with a 70% score from another.

A second detector helps when the first result matters, especially one with a different performance profile. If both flag the same passages, there is more reason to review those sections carefully. More detectors still do not produce proof. A rule that any detector saying AI means AI catches more generated text but also creates more chances for false positives on human writing. Vanderbilt University disabled Turnitin's AI detector in 2023, noting that even a claimed 1% false positive rate across about 75,000 papers a year could mislabel roughly 750 student papers.4 Multiple detectors are useful for screening, and they do not replace evidence about how a document was produced.

What changes in AI detection in 2027

The biggest change is unlikely to be another percentage point on untouched ChatGPT output, which many tools already handle. The harder question is what counts as AI writing when a person and a model have both worked on the same document. One person writes the argument and uses AI for grammar. Another generates a first draft and rewrites half of it by hand. A third asks a model to paraphrase a single paragraph. All three documents involve AI, yet treating them as one category hides important differences.

The benchmark already shows the problem. GPTZero detected 92.7% of direct AI but only 1.3% of human drafts improved by AI, while Originality.ai Lite moved the opposite way and caught 96% of AI-improved writing. The question you ask changes the answer. Humanizers and paraphrasing tools will keep pressure on detectors too, and the spread on humanized AI was 92.6 percentage points between the strongest and weakest tool.

Writers do not need a permanent game of detector evasion. Someone who uses AI heavily and wants the result to sound less mechanical can humanize AI text for free and then edit for meaning, examples, sources and voice. When repetitive wording is the problem rather than the whole draft, the best AI paraphrase tool can rework specific passages. For schools,

publishers and employers, the larger shift will be toward context. Version history, research notes, citations, document metadata and an author's ability to explain the work can answer questions that no statistical detector can. The best detection workflow in 2027 may include a good detector, but it will not end with one.

- 1New AI classifier for indicating AI-written text
- 2AI detectors biased against non-native English writers
- 3FTC order requires Workado to back up artificial intelligence detection claims
- 4Guidance on AI detection and why we're disabling Turnitin's AI detector

Summary

For artificial intelligence [AI] detection in 2027, Clever AI Detector is the best free choice in the 2026 benchmark run by its maker. Its 96.7% overall result was effectively level with Copyleaks at the top, and its main advantage was consistency: it never dropped below 92% in any of the four AI categories. Copyleaks is the strongest paid alternative and slightly outperformed Clever on humanized AI. Originality.ai Lite is particularly strong when AI-assisted editing is the main concern. The broader lesson is simpler. Use a detector to identify writing that deserves a closer look, and do not use the score alone to decide who wrote it.