AI Explained

Can You Tell It's AI?

By Dr Peter Fields

THE 2-MINUTE SUMMARY

A text detector looks for patterns in a finished piece. Its score depends on the tool and the material it was tested on. Even a small error rate can produce a lot of false alarms.

A watermark leaves a signal in the output. Content Credentials can attach a signed record of its origin or edits. Each tells you something different about how the material was made.

If the stakes are high, look for evidence beyond the flag. Ask how the work was produced, or trace a photo, clip or recording back to its source.

The student whose essay was flagged

Imagine a sixteen-year-old who writes her own history essay. She follows the structure her teacher taught her, keeps to a formal tone and checks every reference. Then the school's software flags the essay as likely to be AI-generated. What has it found, and what would be fair to do next?

Reading the writing after the fact

Some early detectors measured how predictable a passage looked to a language model. That measure is called perplexity. A conventional sentence may be easy to predict whoever wrote it. OpenAI withdrew its own text classifier in July 2023 because of its low accuracy.[1,2]

In one 2023 study, seven detectors wrongly flagged more than half of a set of English-language exam essays by non-native speakers, on average. The result describes those tools and those essays. It gives schools a reason to ask how a detector performs on writing like their pupils' before relying on it.[2]

Geek Box · How a small error rate adds up

Suppose a detector flags 1 in 100 of 10,000 essays written without AI. That gives 100 false alarms. If it also catches all 50 essays in a separate group written with AI, two-thirds of the 150 flags still point to work written without AI. These are illustrative numbers. Whether anyone broke a rule depends on what use the school permitted.

What the tests found

A trained classifier learns from examples labelled human-written or AI-generated. How useful it is depends on the writing it encounters. Jabarian and Imas compared four tools in 2025. Pangram made few errors on their main test set and generally performed well on passages under 50 words, though results varied by genre and model. The Robust AI Detection (RAID) benchmark tests how well tools identify AI-written text under different conditions. It found that many of the detectors it tested struggled with unfamiliar sources, models or changes to the text.[3,4]

Human readers can do well too. In a study by Russell, Karpinska and Iyyer, the majority vote of five people who regularly used ChatGPT for writing misclassified just one of 300 articles. The researchers described that performance as comparable to the best automated tool tested. Individual readers varied, and those results tell us about the articles in the study.[5]

The school's question may be harder than the one used in a test. A student might write an essay, then ask AI to tidy the prose. Another might start with an AI draft and rewrite it. To decide whether the work meets the rules, the teacher needs to know what the student did.

A signed record that travels with the file

The Coalition for Content Provenance and Authenticity (C2PA) develops the standard for Content Credentials. These can link a signed record to a photo, video or other file. The record can say which device or service made it, or which edits were recorded. A viewer checks the signature and whether the credential matches the file. Look at who signed it and what they actually attested to. A valid record can accompany a staged photograph, and a copy such as a screenshot may lose the embedded record.[6]

Article 50 of the EU AI Act requires providers in scope to mark synthetic outputs in machine-readable form, with exceptions including standard editing. The duties began applying on 2 August 2026, with a transition to 2 December 2026 for the provider marking duty on systems already on the market. A voluntary Code of Practice sets out ways to comply. Separate duties require visible disclosure for certain uses, including deepfakes.[7]

A watermark in the output

A digital watermark leaves a signal in the content itself. Google DeepMind's SynthID covers images, audio, video and text. Its 2024 paper describes how SynthID Text leaves a pattern in the model's choices as it generates a reply. There are no hidden characters to find. The signal can travel with a copied passage, though short or heavily rewritten passages may leave a weaker signal.[8]

In August 2026, Anthropic announced a version of that method for future Claude models. At announcement, its verification service was in private preview for eligible organisations. Anthropic says proofreading may change too few words for a reliable reading. Text that has to be exact, such as some code, can also leave little signal.[9]

Geek Box · What a text watermark can tell you

A student might write every argument in an essay and ask a marked service to rewrite it. The result may carry that service's watermark. A positive result is evidence about the wording the tool produced. It cannot tell you who came up with the argument or how much of the work the student did.[9]

How to use it better

- **Ask how the work was made.** If you are assessing the essay, explain the concern and give the student a chance to respond. Discuss the argument and look at notes, drafts or version history where available. Check what the rules allowed at the time.

- **Trace a photo or recording to its source.** For a dramatic clip, look for the original file, inspect any Content Credentials and check the claim against independent reporting. An absent mark leaves the question open.[6,8]

- **Say what assistance you used.** Where disclosure is required, describe the task the tool performed. Checking punctuation and drafting a section are different uses. Before sending someone's work to an online detector, check whether you may share it with that service. The Confidentiality and Privacy articles take up that question.

Sources and notes

1. OpenAI (2023), New AI classifier for indicating AI-written text. Update of 20 July 2023 recording the withdrawal. Announcement

2. Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu and James Zou (2023), GPT detectors are biased against non-native English writers, Patterns 4(7), 100779, Fig. 1. Paper

3. Brian Jabarian and Alex Imas (2025), Artificial Writing and Automated Detection, NBER Working Paper 34223 and BFI Working Paper 2025-116, §§3–4 and App. B.1. Working paper

4. Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito and Chris Callison-Burch (2024), RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors, ACL, pp. 12463–12492. Paper

5. Jenna Russell, Marzena Karpinska and Mohit Iyyer (2025), People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text, ACL, pp. 5342–5373, Table 2 and §3. Paper

6. Coalition for Content Provenance and Authenticity, C2PA Content Credentials, v2.4, and C2PA Harms Modelling, §6.2. Signature validation, provenance and the limits of what the record shows. Specification · Limits

7. Regulation (EU) 2024/1689, arts 50, 111(4) and 113, as amended. European Commission, Code of Practice on marking and labelling AI-generated content, published 10 June 2026. Article 50 · Transition · Code announcement

8. Sumanth Dathathri, Abigail See, Sumedh Ghaisas and colleagues (2024), Scalable watermarking for identifying large language model outputs, Nature 634, pp. 818–823. See also Google DeepMind's overview of SynthID for other media. Paper · SynthID overview

9. Anthropic (14 August 2026), How Claude's text watermark works. Method, announced rollout, access at launch and limits with proofreading, short passages and exact outputs. Announcement