StillEssentialU

"Generated by AI" Marks: What Should Faculty Know?

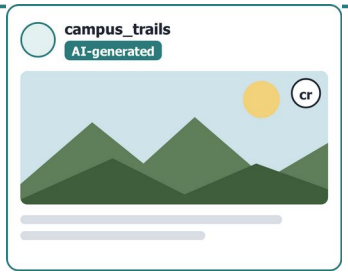
> **Short answer: They won't tell you whether a student's work is their own.**

AI marks have constraints that make detection unreliable, **particularly for short narrative assignments.**

CONSTRAINT	WHY?
Copying removes metadata	Content Credentials don't survive copying text into Word or a discussion board.
Watermarks are tool-specific	A Gemini watermark tells you nothing about text from ChatGPT, Claude or Copilot, and faculty generally can't get detector access.
Short text gives a weak signal	Watermark detection needs a lot of words. A short response of 150–200 words is at the low end.
Editing weakens it	Google's own research says paraphrasing degrades the watermark. A student who rewrites an AI draft can remove most of the signal.
A missing mark proves nothing	The text could come from an unmarked tool, have been edited, or be the student's own. Most real student work is a mix.

© 2026 Still Essential LLC · Free for individual educators to use and adapt · stillessential.com ·

There are three types of marks: visible labels, Content Credentials (C2PA) and invisible text watermarks.



Visible labels

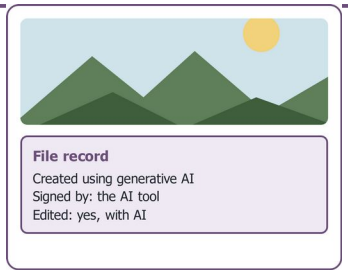
What it is: A badge such as "AI" or "Made with AI," or a small "cr" (Content Credentials) pin on an image or video. Clicking the pin shows how the file was made, including whether AI was used.

Added by: The AI tool, the platform (such as social media) or the person posting.

On fully AI-generated work: Shows only if the tool adds it and the user keeps it. The EU's optional icon reads "Fully AI-generated."

After a student edits or copies it: Easy to remove. The EU's optional icon reads "Partially AI-modified," and text a person has reviewed and takes editorial responsibility for needs no label under EU rules.

Can faculty see it? Only if it's visible. It isn't required on student work.



Content Credentials (C2PA)

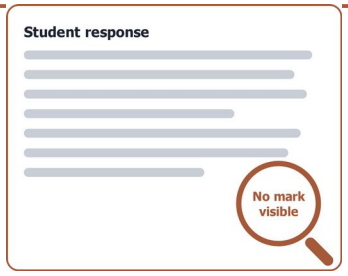
What it is: A digitally signed record stored inside the file, used mostly for images, video and audio.

Added by: The software when it creates or saves the file.

On fully AI-generated work: Can record "Created using generative AI."

After a student edits or copies it: Can record "Edited using generative AI." Copying text removes the record entirely.

Can faculty see it? Only on the original file.



Invisible text watermarks

What it is: A hidden statistical pattern in word choice. Google's SynthID marks Gemini text. Since August 2026, Anthropic marks Claude text worldwide, and since October 2026, OpenAI marks ChatGPT text in the EU.

Added by: The AI company's system as the text is written. The model itself doesn't know.

On fully AI-generated work: Detectable if the tool marks it: about 80% for a 200-token passage (OpenAI).

After a student edits or copies it: Fades with edits. Swapping a quarter of the words cut detection to about 17%, and paraphrasing or translation can erase it.

Can faculty see it? Rarely. Detectors are limited to approved researchers and organizations, and each works for only one company's tools.

Example: this brief

Leslie wrote this brief with Claude, an AI assistant, then revised it heavily until she considered it complete. Here is what each type of mark would show:

TYPE OF MARK	WHAT THIS BRIEF WOULD SHOW
Visible label	None. Claude doesn't add one to text. The AI use disclosure on page 4 is how Leslie chose to disclose AI use.
Content Credentials	None. Claude adds them to some image files, not to Word or PDF documents.
Invisible text watermark	Possibly a weak one. Claude began watermarking its text in August 2026, but heavy editing weakens the signal, and only approved organizations can check for it.

The marks can't tell you how much of this brief is Leslie's own work. Only her process can.

What this means for your course

AI marks are useful for checking whether an image or news article is real. They aren't a grading tool. Designing assignments that ask for a student's own thinking (what happened in your class, what they thought before, where a concept doesn't fit) remains more reliable than trying to detect AI after the fact.

Bottom line: Most student work that involves AI is a mix, and that's exactly the work these marks can't sort out.

What's changing

- **New EU rules.** From August 2, 2026, the EU AI Act requires companies offering generative AI to mark AI outputs in a machine-readable, detectable way. Systems already on the market have until December 2, 2026. This applies to companies serving EU users, which includes the major AI companies, so text watermarking may become more common. Tools that only help with standard editing, such as grammar correction, are exempt.
- **Anthropic now watermarks Claude's output.** Since August 2026, Anthropic has added an invisible watermark to Claude's text worldwide and Content Credentials to some images it creates. Heavy editing, paraphrasing or translation can remove the text watermark, and only approved groups, such as newsrooms and researchers, can use its text detector.
- **OpenAI now watermarks some text.** In October 2026, OpenAI began adding an invisible watermark ("textGrain") to ChatGPT and Codex text in the EU, as the new EU rules require. Outside the EU it is off unless a developer turns it on, and only approved researchers can use the detector. OpenAI reports about 80% detection for a 200-token passage (roughly 150 words). In its tests, swapping a quarter of the words for synonyms cut detection to about 17%.

AI use disclosure

HOW I MADE THIS

Tools: Claude (Anthropic), October 2026

What AI did: Found and summarized sources, drafted the text and built the layout.

What I did: Reviewed key sources, selected what to include, and revised wording.

What did I check or change in the AI's answer?
One example: The first draft didn't explain who adds each mark or what it looks like. Faculty can't judge a mark they can't recognize, so I had Claude add both, in the text and as graphics.

Sources

Anthropic. (2026). How Claude marks AI-generated content. Claude Help Center. https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content

Coalition for Content Provenance and Authenticity (C2PA). (n.d.). Content Credentials. https://c2pa.org/

Dathathri, S., See, A., et al. (2024). Scalable watermarking for identifying large language model outputs. Nature. https://www.nature.com/articles/s41586-024-08025-4

European Commission. (2026, June 10). Commission publishes Code of Practice on marking and labelling AI-generated content. https://digital-strategy.ec.europa.eu/en/news/commission-publishes-code-practice-marking-and-labelling-ai-generated-content

Future of Life Institute. (n.d.). The EU AI Act's transparency rules: A practical guide to Article 50. EU Artificial Intelligence Act. https://artificialintelligenceact.eu/transparency-rules-article-50/

Mehta, I. (2026, August 11). Anthropic says it will watermark text generated by its AI models. TechCrunch. https://techcrunch.com/2026/08/11/anthropic-says-it-will-watermark-text-generated-by-its-ai-models/

OpenAI. (2026, October 5). Our approach to EU text provenance rules. https://openai.com/index/eu-text-provenance/