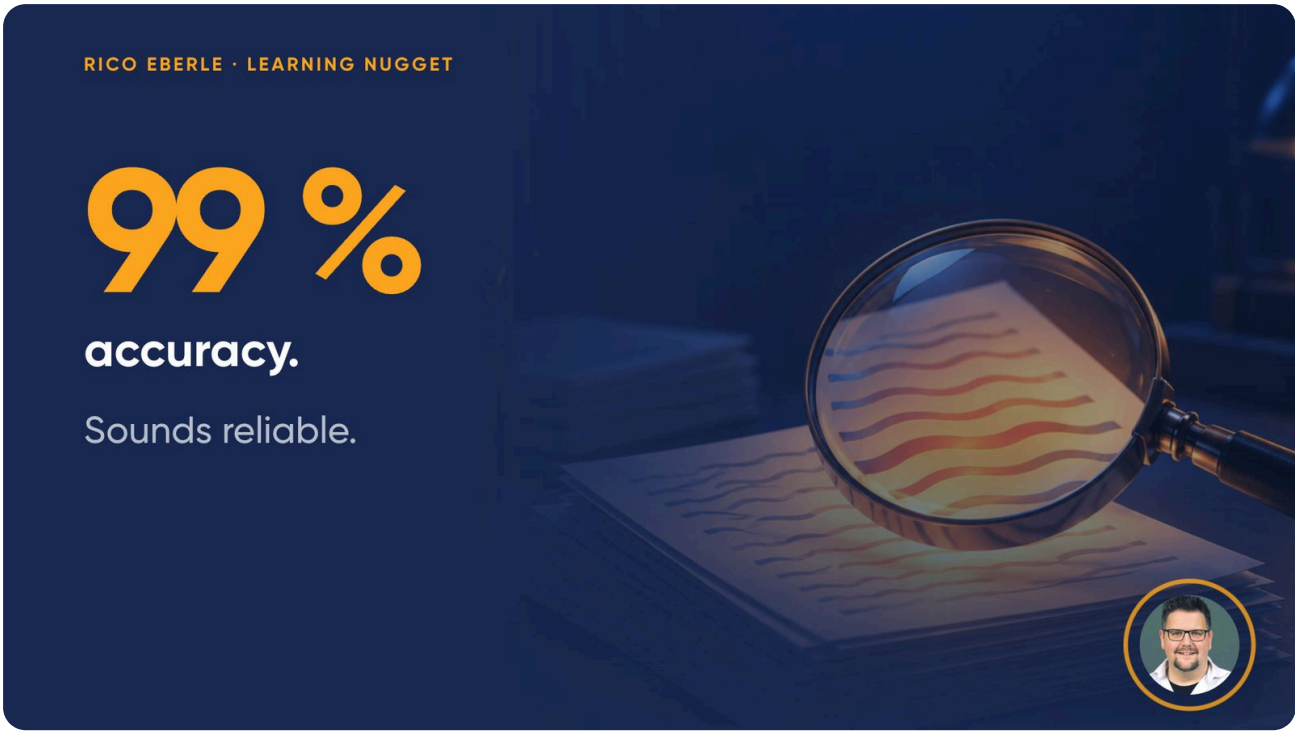
Rico Eberle · LEARNING NUGGETS

LEARNING NUGGET E-LEARNING & AI 15 SEPTEMBER 2026 (UPDATED 31 OCTOBER 2026)

AI detectors are not proof

Even one percent false alarms means ten honest students suspected per 1000 essays, and in practice AI detectors are far less accurate.

By Rico Eberle, e-learning expert for continuing education and municipal councillor in Dübendorf



Video for this post

This learning nugget is also available as a video with transcript (1:08 min). Scan the QR code or open the link:

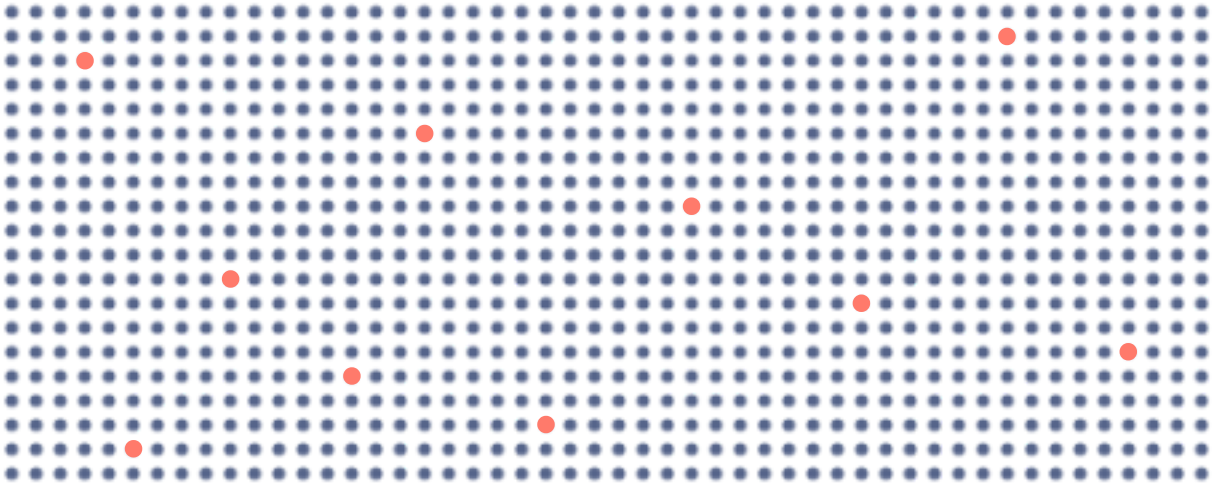
ricoeberle.ch/en/blog/ai-detectors-are-not-proof/

99 percent accuracy. That sounds like a reliable AI detector. Wanting a tool that says “this was AI” at the push of a button is understandable, and vendors advertise with numbers like this. Let’s do the maths anyway.

One percent sounds small until you scale it

Let’s assume “99 percent” means this: only one percent of honest texts are wrongly flagged as AI. A school runs 1000 essays through a detector. Every one of them was written honestly. A one percent false alarm rate means ten students fall under suspicion, even though they did nothing wrong.

1000 honest essays



1 % false alarms = 10 wrongly suspected students

Worked example with an optimistic number

False alarms grow with volume. The more texts you check, the more innocent students get flagged, even if the percentage stays the same. Each case is a difficult conversation with trust at stake.

And 99 percent is a deliberately optimistic figure.

In practice the tools are worse

In 2023, a research team led by Weber-Wulff tested fourteen detection tools. Their verdict: the tools are neither accurate nor reliable. Above all, they tended to classify AI text as human-written, so they missed it. A detector can therefore be wrong in both directions: it suspects honest students and lets AI text through.

Worse in practice

14 tools tested

“neither accurate nor reliable”

Weber-Wulff et al., 2023

Written in a second language

flagged as AI especially often

Liang et al., 2023

OpenAI

shut down its own detector

OpenAI, 2023

A second study, by Liang and colleagues, points to a fairness problem. English texts by people writing in a language that isn’t their first are flagged as AI especially often. The study looked at English essays, including TOEFL essays. It says nothing about other languages.

The researchers suggest that detectors may unintentionally penalise writers with more constrained linguistic expression. They explicitly caution against using these tools in evaluative or educational settings.

OpenAI, the company behind ChatGPT, shut down its own detector in July 2023. The reason: its low rate of accuracy.

A reason to look closer, not proof

At most, a detector result is a reason to look more closely. Other things do more in everyday teaching. In the video I recommend three:

1. **Make the writing process visible.** Drafts, notes and intermediate versions become part of the submission. Students who developed a text themselves can show how they got there.
2. **Talk with the student about the text.** Not an interrogation, but part of the assessment: Why this structure? What was hard?
3. **Design tasks that require their own thinking.** For example, by connecting them to your own lessons, personal experience or a decision that has to be justified.

The maths for your own staff room

The calculation above works in any staff meeting. How many written assignments are handed in at your school each semester? Take one percent of that, the optimistic figure. The result is the number of conversations with honest students you would have to have.

After that, it is worth asking which tasks in your own teaching would make a detector unnecessary.

Trust doesn’t come from software. It comes from good tasks.

Frequently asked questions

How reliable are AI detectors?

Not very. In 2023, a research team led by Weber-Wulff tested fourteen detection tools and concluded that they are neither accurate nor reliable. OpenAI shut down its own detector in July 2023, citing its low rate of accuracy.

Is an AI detector result proof?

No, at most it is a reason to look more closely. Even at an optimistic false positive rate of one percent, ten out of 1000 honestly written essays would be flagged wrongly.

Are AI detectors biased against non-native writers?

For English, yes: according to a study by Liang and colleagues (2023), English texts by people writing in a language that isn't their first are flagged as AI especially often. The study did not look at other languages. The researchers caution against using these tools in evaluative or educational settings.

Sources

1. Weber-Wulff et al.: Testing of detection tools for AI-generated text, International Journal for Educational Integrity (2023)
   https://link.springer.com/article/10.1007/s40979-023-00146-z
2. Liang et al.: GPT detectors are biased against non-native English writers, Patterns (2023)
   https://arxiv.org/abs/2304.02819
3. OpenAI: New AI classifier for indicating AI-written text (2023, discontinued)
   https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

ABOUT ME

Rico Eberle

I work in e-learning, hold a degree in business economics (FH) and sit on the municipal council of Dübendorf, Switzerland. I also chair the foundation board of WBK Dübendorf, a continuing education foundation. I make the learning nuggets on my own: topic, research, text and video. Questions and disagreement are welcome.

Contact: ricoeberle.ch · LinkedIn: linkedin.com/in/rico-eberle

LICENCE

Text, transcript and video by Rico Eberle are licensed under CC BY 4.0 (music and sound effects in the video excluded). You may share and adapt the content, including for teaching, as long as you credit the source. creativecommons.org/licenses/by/4.0/

SUGGESTED CITATION

“AI detectors are not proof”, Rico Eberle, ricoeberle.ch, 2026, https://ricoeberle.ch/en/blog/ai-detectors-are-not-proof/, CC BY 4.0