

LEARNING NUGGET E-LEARNING & AI 15 SEPTEMBER 2026

AI detectors are not proof

Why even a one percent false alarm rate hits many honest students, what research shows about AI detectors, and what works better in the classroom.

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

RICO EBERLE · LEARNING NUGGET

99 %

accuracy.

Sounds reliable.

**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/

Ever since language models started writing entire essays, schools have been looking for a tool that settles the question at the push of a button: was this written by AI? Detector vendors promise exactly that, often backed by impressive percentages. The trouble is not only that those numbers rarely hold up in practice. It starts with the maths itself.

One percent sounds small until you scale it

Say a detector is right 99 percent of the time. Your school runs 1000 essays through it, and every single one was written honestly. A one percent false alarm rate then means ten students fall under suspicion, even though they did nothing wrong.



Even at 99 percent accuracy, ten out of 1000 honest essays come under suspicion.

The key point: 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.

What the research shows

Real-world results are considerably worse. In 2023, a research team led by Weber-Wulff tested fourteen detection tools. Their verdict: the tools are neither accurate nor reliable.

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



Three research findings on AI detectors.

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, and they explicitly caution against using these tools in evaluative or educational settings.

Even OpenAI, the company behind ChatGPT, shut down its own detector in July 2023, citing its low rate of accuracy.

What works better

A detector result is not proof. At most, it is a reason to look more closely. In everyday teaching, these three recommendations from the video go further:

- 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 as an interrogation, but as 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.

Not proof.

WHAT HELPS INSTEAD

- 1 Make the process visible: drafts, notes
- 2 Talk with the student about the text
- 3 Design tasks that require their own thinking



Three approaches that achieve more than a detector score.

An exercise for your staff meeting

My suggestion: estimate together how many written assignments are handed in at your school each semester. Then apply a one percent false alarm rate, the optimistic figure. The number you get is how many conversations you would have to have with honest students. After that, ask a better question: which tasks in our 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 99 percent accuracy, 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 THE AUTHOR

Rico Eberle

Rico Eberle is an e-learning expert, business economist (FH) and municipal councillor in Dübendorf, Switzerland. He chairs the foundation board of WBK Dübendorf, a continuing education foundation. In the learning nuggets he explains research on learning, AI and digital sovereignty, briefly and with sources.

Contact: ricoeberle.ch · LinkedIn: [linkedin.com/in/rico-eberle](https://www.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