

GUIDE

AI detectors and their limits

Published tests show AI detectors miss much AI text and flag human writing, especially from non-native speakers. Use them as one weak signal, never as proof.

Updated November 13, 2026

Contents

What OpenAI's own test found ····· 5

What the Stanford study found ····· 5

What this means in practice ····· 5

The better question ····· 5

Sources ····· 5

AI detectors promise to answer one question: did a machine write this? The published evidence says they answer it badly, in both directions. Here is what the public record shows.

What OpenAI's own test found

In January 2023 OpenAI released a free classifier for AI-written text. Its own evaluation on a challenge set of English texts found it correctly identified 26% of AI-written text as "likely AI-written", while incorrectly labeling human-written text as AI-written 9% of the time. It was unreliable on texts under 1,000 characters, recommended for English only, and OpenAI said it should not be used as a primary decision-making tool. On July 20, 2023 OpenAI withdrew the classifier, citing its low rate of accuracy.

What the Stanford study found

Researchers including Weixin Liang and James Zou tested seven widely used GPT detectors, published in the journal Patterns in July 2023 (DOI 10.1016/j.patter.2023.100779). On 91 TOEFL essays written by non-native English speakers, the detectors mislabeled more than half as AI-generated; the average false-positive rate was 61.3%. All seven detectors unanimously flagged 19.8% of the human essays, and at least one detector flagged 97.8% of them. On US eighth-grade essays the detectors were near-perfect, which shows the bias runs against simpler, non-native-style writing. The mechanism is perplexity: detectors score how predictable the word choices are, and plainer writing scores as more machine-like.

What this means in practice

- A high detector score is not evidence. Human writers, especially non-native speakers and anyone writing plainly, get flagged.
- A low score is not evidence either. AI text edited by a person, or produced by a model the detector never saw, can pass.
- If you run a classroom, a hiring process or a publication, do not make a detector the deciding factor. Ask for drafts, revisions and sources instead.
- If a tool you are evaluating ships a "AI detector" as a feature, weigh it the same way: a weak signal, marketed as a feature.

The better question

The useful question is not "did AI write this" but "is this accurate, original and useful". Google's guidance on generative AI content focuses on exactly that: review for accuracy before publishing, add value for the reader, and do not scale thin pages. Judge the text on those terms and the detector's opinion matters much less.

Sources

All sources below were fetched and checked on 2026-11-13. That is the access date, not a claim about publication date.

- OpenAI: New AI classifier for indicating AI-written text, January 31, 2023, accessed 2026-11-13.

- Stanford Electrical Engineering: GPT detectors are biased against non-native English writers, May 2023, accessed 2026-11-13.
- Stanford SCALE Initiative: GPT detectors are biased against non-native English writers, July 2023, accessed 2026-11-13.

Frequently asked questions

Can an AI detector prove a text was written by AI?

No. OpenAI's own classifier caught only 26% of AI-written text in its evaluations and was withdrawn in July 2023 for low accuracy. Treat any detector score as a weak signal.

Can a detector prove a text was written by a human?

No. Stanford researchers found detectors flagged more than half of human-written TOEFL essays as AI-generated. A clean score is not proof of human authorship either.

Sources

openai.com — https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/
ee.stanford.edu — https://ee.stanford.edu/james-zou-et-al-warn-objectivity-ai-detectors
scale.stanford.edu — https://scale.stanford.edu/publications/gpt-detectors-are-biased-against-non-native-english-writers