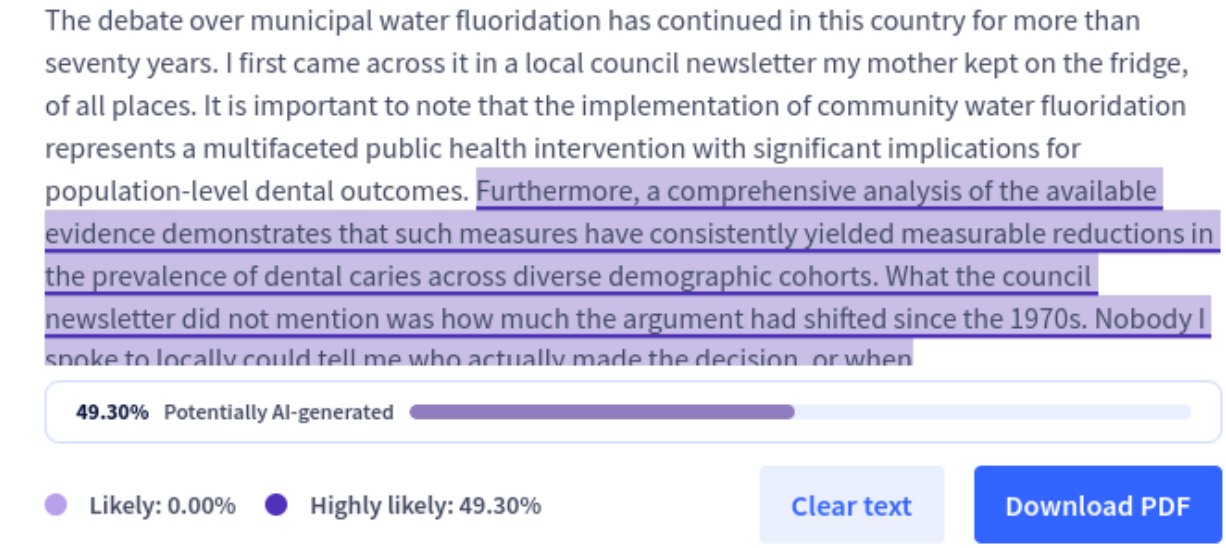
AI Detection accuracy and method: evidence, numbers, and competitor comparison

Unlike most popular AI detectors, our tool doesn't produce a general document score. It checks every sentence, deciding the probability of it being AI-generated, and flags parts with AI traces in the report. How much AI presence is too much is up to you to judge.



“20% AI-generated” result leaves you wondering whether it is a whole machine-generated paragraph or just small edits across the document. Highlighted lines give you a conversation-starting point and clear evidence to make data-driven decisions.

How accurate is the result?

99.23% AI detection accuracy

We checked 33,699 sentences in 709 documents generated by seven frontier AI models and correctly classified 99.23% of them as AI, measured per sentence.

1.04% false positives (sentence level)

Of 280,781 sentences written by people, we misclassified 1.04% as AI.

0.00% false positives (document level)

With a sentence-by-sentence check approach, a misclassified line doesn't condemn the whole paper. This way, at the document level, our false-positive rate drops to zero. None of the 2,805 human-written documents were wrongly classified as AI.

Why two false-positive figures

Metric	Result	What it means
Recall (sentence)	**99.23%**	Of all AI-written sentences, we correctly flag 99.23%. Measures how much AI text we catch.
False positives (sentence)	**1.04%**	Of all human-written sentences, we wrongly flag 1.04%. Measures how often we unfairly accuse human text.
Recall (document)	**100%**	Of 709 fully AI-written documents, we correctly identified every one.
False positives (document)	**0.00%**	Of 2,805 human-written documents, zero were wrongly classified as AI.

Was part of an essay generated by ChatGPT, or was it the detector catching formulaic-sounding extracts in a handful of sentences? With a general metric per document, it is hard to tell.

Occasionally, every detector mismarks the sentences that resemble robotic patterns as potentially AI. On their own, they don't prove the document's origin. What matters is the whole paper's shape, the consistency of highlighted text, and its structure. Usually, an AI-generated essay won't contain scattered machine-sounding phrases. It will have flagged whole chunks of writing, if not the majority of the text.
So, we don't hurry into classifying the whole paper. We analyze every sentence and highlight those containing AI traces. You set the threshold for what proportion of flagged content is alarming, and we give you evidence to decide whether it crosses your limit.

- Initially, we classify every sentence. Hence, the primary metrics we provide are per sentence.
- Our competitors provide one verdict per document, and their metrics are per document. To compare like for like, we translated our sentence measure into a document-level figure.
- At the document level, our error rate drops to zero, as with our approach, one misclassified sentence doesn't condemn the whole paper.

The threshold you set

Checking thesis papers and editing a social media post are two different tasks that require different approaches to measurement. Sentence-by-sentence analysis allows us to provide a flexible threshold rather than a fixed operating point.
You choose which setting fits your tasks today. High sensitivity is recommended for early intervention, and a more conservative setting is appropriate when the stakes are high and the consequences serious. The same scan supports both, with no re-processing needed.

Content flagged as AI	AI recall	95% CI*	False positives	95% CI	Suggested use
<5% of sentences	**100%**	99.5–100	**5.45%**	4.67–6.36	Screening only
>10% of sentences	**100%**	99.5–100	**2.10%**	1.63–2.70	High sensitivity
>20% of sentences	**100%**	99.5–100	**0.64%**	0.41–1.01	Balanced
>30% of sentences	**100%**	99.5–100	**0.25%**	0.12–0.51	Conservative
>50% of sentences	**100%**	99.5–100	**0.00%**	0.00–0.14	**Disciplinary review**

*(*CI=confidence interval)*

Measured on 709 AI documents and 2,805 human documents, July 2026.

Which AI models we detect

When trained to recognize one model, the detector fails to catch the other provider's output. Here is evidence that we are not just a ChatGPT detector. The total spread across seven flagship models is 1.56 points, which means we classify them with 98.31%-99.87% accuracy.

Provider	Model	Documents	Sentences	Detected	Recall
Google	Gemini 3.1 Pro	116	4,578	4,572	**99.87%**
	Gemini Flash 3.5	93	4,311	4,245	**98.47%**
	Provider total	**209**	**8,889**	**8,817**	**99.19%**
Anthropic	Claude Sonnet 4.6	100	5,035	5,020	**99.70%**
	Claude Opus 4.7	100	5,576	5,553	**99.59%**
	Provider total	**200**	**10,611**	**10,573**	**99.64%**
OpenAI	GPT-5.4	100	7,421	7,365	**99.25%**
	GPT-5.5	100	3,709	3,668	**98.89%**
	Provider total	**200**	**11,130**	**11,033**	**99.13%**
xAI	Grok 4.3	100	3,069	3,017	**98.31%**
All	**7 models**	**709**	**33,699**	**33,440**	**99.23%**

Unedited model output, generated and tested in July 2026. Recall measured per sentence.

As new models emerge and are continually upgraded, we retrain our detector to stay up to date. You can follow our recent upgrades here.

How we checked

To minimize false positives, the detector must be trained and tested on purely human-written text. AI presence in it affects the accuracy rate and test results. Here is how we avoided that.

Human text corpus from pre-AI era

Our main corpus is the British Academic Written English (BAWE) collection, the real university coursework gathered between 2004 and 2007. It is 15 years before ChatGPT was launched; hence, it physically cannot contain AI traces. BAWE contains 2,688 documents (277,033 sentences), to which we added our own internal human corpus of 117 documents (3,748 sentences). 280,781 purely human-written sentences in total were used to test our AI checker.

AI text corpus from frontier models

Seven LLM models, roughly 100 documents from each, 33,699 sentences in total. Each text used for testing is a pure AI output, with no paraphrasing or manual editing.

Per-sentence analysis

Each document submitted for checking is split into sentences. The detector classifies each of them, and the test results are counted at the sentence level. The document-level score is calculated based on sentence analysis, and the final verdict is defined by the threshold you set, from 5% to 50%.

Data transparency

Every rate we publish is reported with a 95% Wilson confidence interval. We reveal our sample size, 280,781 sentences, making the 1.04% false-positive rate figure meaningful rather than demonstrative.

Versioned and updated

AI models are being upgraded, and so are our AI detector and this report. Every figure and piece of data on this page is marked with a model version and the date when the test was run. We keep improving the tool, our test methods, and this report regularly.

Element	Description
Segmentation	Every document is split into sentences, and each sentence is classified independently as AI or human. We do not produce a single whole-document verdict — the model shows which specific sentences look AI-written.
AI test set	Roughly 100 documents per model across 7 frontier LLMs, 33,699 sentences total. All text is unedited model output — no paraphrasing or manual revision.
Human test set	Two corpora, 2,805 documents and 280,781 sentences in total. BAWE (2,688 docs / 277,033 sentences, FPR 1.05%) and our internal human corpus (117 docs / 3,748 sentences, FPR 0.61%). Combined FPR is 1.04%.
Contamination control	Our primary corpus BAWE was collected in 2004–2007, fifteen years before ChatGPT, so it physically cannot contain AI-written text. This is the main reason our FPR figure is trustworthy in a way most published numbers are not.
Document-level derivation	A document is flagged as AI if more than a set percentage of its sentences are flagged. We report thresholds from 5% to 50% so an institution can choose its own tolerance. The 0.00% FPR figure is at the >50% threshold.
Statistics	All rates reported with 95% Wilson confidence intervals. The sample size (280,781 human sentences) is what makes a 1.04% figure meaningful rather than noise.
Versioning	Figures are as of July 2026. Performance shifts as new LLMs launch — re-run quarterly.

How this compares

Most detectors on the market claim 98-99.98% accuracy. However, those numbers, including ours, are marketing figures, based on the vendors' own tests.
What distinguishes our study is that we publish how many human-written content samples we tested, a metric none of the competitors disclose. This matters because AI presence in human-claimed texts distorts false-positive rate figures, making them look better than they are and reducing test precision.

Detector	Claimed Accuracy	Claimed False Positive Rate	Human Sample Tested	Source (Vendor Data)
PlagiarismCheck.org	99.23% recall, sentence	0.00% document	2,805 documents 280,781 sentences	This page
Pangram	99.98%	~0.004%	not published	pangram.com

Originality.ai	99% (Lite) / 99%+ (Turbo)	0.5% / 1.5%	not published	originality.ai
GPTZero	99%	<1%	not published	gptzero.me
Copyleaks	99%+	<0.2%	not published	copyleaks.com
Turnitin	98%	<1%	not published	turnitin.com
Winston AI	99.98%	not stated	not published	gowinston.ai

All figures retrieved in July, 2026. Independent studies frequently report different numbers from what vendors claim. ZeroGPT states 98% accuracy, while independent tests show 74%-80%.

Accuracy rate by itself is not representative unless it is backed by a research method and transparent study figures. This is why we publish our corpora, sample sizes, confidence intervals, and threshold curve.

What we haven't tested yet

There are some gaps in the 2026 results we openly warn you about and plan to fill in the next tests.

- **Paraphrased and humanized text.** All the figures above apply to unedited AI output. We have not yet published the data for text run through tools designed to disguise AI use, so we won't claim parity with competitors until revealing the numbers.
- **Hybrid content.** Mixed documents containing part human-written, part AI text are the most common real-world case and exactly what our sentence-level analysis is designed for. However, we haven't benchmarked it yet, and this is the next test we run.
- **Independent tests.** All published figures are internal, not verified by independent study – yet. We are working with researchers to change it, and the offer below stays open.

Free access for researchers

We invite researchers and independent evaluators to open testing. Reach out to get free access to our AI detector, with no review of your results before publication.

Contact us to get free access for independent testing!