

Overview for authors, AI detection tools

V01, 21/07/2026

Summary

This summary about AI detection tools is based on a limited testing of such tools and on published benchmarks (arXiv publications highlighted below). We also mention a resource page from Wikipedia that explains how humans can spot AI-generated text and list URLs to some AI detectors and citation checkers.

Key points:

- No tool alone can detect AI written text in a reliable way; this is also highlighted and acknowledged by the commercial tool providers.
- Uploading manuscripts from other authors to AI detection tools raises issues of data privacy and copyright; reviewers and editors might need to ask for permission from authors to analyse manuscripts with AI detection tools. Our [AI policy page](#) can be a start but we recommend discussing this use case with editors.
- LLMs are improving quickly at writing better and more human-sounding texts. AI detection tools are equally quickly developing to keep up with this process. Overall, this leads to a very short half-life of quality assessments of AI detection tools.
- Humans, especially when experienced with LLM-written text, can be very good at correctly identifying LLM-written text; trained humans can surpass the best available AI detection tools.

“RAID: A shared benchmark for robust evaluation of machine-generated text detectors” (<https://doi.org/10.48550/arXiv.2405.07940>)

Background:

- Robust benchmark for AI detection tools
- 6 million LLM-generated texts
- high ranking in this benchmark is advertised by many commercial AI detector tools.

Details:

- So-called perplexity can be a suitable measure to differentiate human-written from LLM-written text (“Perplexity is how unexpected, or surprising, each word in a piece of text is, when looked at from the perspective of a particular language model or LLM.” <https://www.pangram.com/blog/why-perplexity-and-burstiness-fail-to-detect-ai>)
- Perplexity is typically higher for human-written text. But different LLMs show very different perplexity values and might be harder to detect (Table 3 in the manuscript).
- Detectors can be tuned and their false-positive rate can be improved. But false-positive rates remain a serious concern (Table 4 and Fig. 4).
- Detectors can achieve remarkable detection accuracy for texts from a given LLM and/or for generated texts of a certain domain. But detection accuracy can dramatically drop in untrained areas (Table 5 and Fig. 6).
- Some strategies to hide LLM-generated text from detection can be very effective (for example, [homoglyphs](#); Table 6).

Conclusion:

- The up-to-date results from this benchmark for different detection tools are available: <https://raid-bench.xyz/leaderboard>
- Unsophisticated attempts at cheating with LLMs seem to be easily detectable (unsophisticated in the sense: text is 100% AI written, most common LLMs used, no attempts to increase perplexity by humanising text, no evasive strategies like using homoglyphs, ...). But significant false positive rates and the easily available methods to avoid detection question the reliability of AI detection that only relies on AI detection tools.

“People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text” (<https://doi.org/10.48550/arXiv.2501.15654>)**Background:**

- Human annotators were hired to decide if provided texts were AI- or human-generated
- Experienced LLM users excelled at detecting AI-generated text
- Such experienced annotators relied on specific cues but also picked up more complex signs of AI-generation and adapted to changes of LLMs for text generation

Details:

- Annotators with little LLM experience confidently decided whether a text is AI- or human-generated. However, their detection accuracy was as bad as a coin toss. Experienced LLM users have a similar confidence but score high true positive rate (92.7 %) and low false positive rate (4.0 %; Table 1).
- Human annotators correctly recognized AI-generated text due to vocabulary and sentence structure. Human-generated text was correctly recognized due to a wide range of cues. (Fig. 3)
- A majority vote from 5 experienced annotators correctly identified AI-generated text in 299 out of 300 cases.

Conclusion:

- It seems encouraging that humans can be so good at detecting AI-generated text. However, this was only true for experienced LLM users. Moreover, the accuracy of 299/300 correct annotations was achieved by a majority vote of 5; hardly a scalable approach. But on the other hand, 3-5 reviewers are close to a typical reviewing situation in an academic journal.

“Wikipedia:Signs of AI writing”

(https://en.wikipedia.org/wiki/Wikipedia:Signs_of_AI_writing)

Key points:

- This page is a detailed “field guide to help detect undisclosed AI-generated content *on Wikipedia*: while some of the signs may be broadly applicable, some may not apply in a non-Wikipedia context”; it could be a helpful resource if one wants to become an “expert human annotators” as described in the arXiv publication mentioned above.
- The accompanying “[Wikipedia:AI or not quiz](#)” can be used to test your skills and to compare it with other people’s high scores. Automated detectors (shown at the end of “Share your score”) score between 10/10 and 7/10.

AI detectors, examples

URLs and prices were collected on 09/04/2026. Prices are annual prices for the cheapest professional licence.

- GPT Zero: <https://gptzero.me/de> (155.88 USD)
 - How it works: <https://gptzero.me/news/how-ai-detectors-work/>
- Grammarly: <https://www.grammarly.com/> (144 EUR)
 - How it works: <https://www.grammarly.com/blog/ai/how-do-ai-detectors-work/#2>
- Originality: <https://originality.ai/> (155.40 EUR)
- Winston ai: <https://gowinston.ai/> (120 USD)
- QuillBot: <https://quillbot.com/> (99.96 EUR)
- Phrasly.ai: <https://go.phrasly.ai/> (104.28 CHF)
- Binoculars: <https://github.com/ahans30/Binoculars>
 - Open, freely available model that scored high in the RAID benchmark
 - You can test it yourself here: <https://huggingface.co/spaces/tomg-group-umd/Binoculars>

Citation checker

These citation checkers are for free.

Hallucinated citations can be a clear sign of misuse of AI during manuscript preparation. But even citation checkers can cause false positive signals and their detection results need additional verification.

- SwanRef: <https://www.swanref.org/>
- CheckIfExist: <https://zabbonat.github.io/References-Validation/>