Personal experience with AI-generated peer reviews: a case study
Nicholas Lo Vecchio
Research Integrity and Peer Review 2025
DOI: 10.1186/s41073-025-00161-3

Additional file 1

Table of Contents

Appendix A. AI detection tests

I ran the peer review reports through several online AI detectors, using various controls. The detection tools accurately predicted the pure ChatGPT output and known human output, and predicted that the review reports were machine.

As has been widely reported in papers [17, 27, 28, 29, 30], press [48, 49, 50] and by OpenAI itself [51], early iterations of such commercial detectors have shown notorious unreliability. Their development is ongoing (and improvable [31]) and their use merits caution (including by specialists who have excessively relied on a single detector [22, 52]). When properly controlled, AI detection tools may offer relative, not absolute, indications that help to contextualize a text's origin.

AI detector	ChatGPT output	Reviewer 1		Reviewer 2		My article
		Peer review report	**Response to author**	**Peer review report**	**Response to author**	
AI Detector Pro	98% chance AI	98% chance AI	2% chance AI	98% chance AI	3% chance AI	2% chance AI
Content at Scale	"Human probability: **X** Reads like AI!" "Unfortunately, it reads very machine-like"	"Human probability: Hard to tell" "Unfortunately, it reads very machine-like"	"Human probability: Passes as Human!"	"Human probability: **X** Reads like AI!" "Unfortunately, it reads very machine-like"	"Human probability: Passes as Human!"	"Human probability: Passes as Human!"
Copyleaks	100% AI content	73.5% AI content	0% AI content	60% AI content	0% AI content	0% AI content
Crossplag	ChatGPT 3.5: AI Content Index 80% – 100% ChatGPT 4: AI Content Index 0% – 100%	AI Content Index 100% "This text is mainly written by an AI"	AI Content Index 0% "This text is mainly written by a human"	AI Content Index 82% "This text is mainly written by an AI"	AI Content Index 0% "This text is mainly written by a human"	AI Content Index 0% "This text is mainly written by a human"
DetectGPT	Probability Breakdown: Human 0% 84% – 97% probability AI depending on text input	Probability Breakdown: Human 0% Mixed 7% AI 93%	Probability Breakdown: Human 78% Mixed 3% AI 19%	Probability Breakdown: Human 3% Mixed 4% AI 93%	Probability Breakdown: Human 75% Mixed 4% AI 21%	Probability Breakdown: Human 100% Mixed 0% AI 0%
GLTR	higher likelihood	higher likelihood	lower likelihood	higher likelihood	lower likelihood	lower likelihood
GPTZero	95% – 98% probability AI generated depending on text input	Probability Breakdown: Human 24% Mixed 32% AI 44%	Probability Breakdown: Human 93% Mixed 6% AI 0%	Probability Breakdown: Human 83% Mixed 15% AI 2%	Probability Breakdown: Human 90% Mixed 10% AI 0%	Probability Breakdown: Human 91% Mixed 9% AI 0%
Originality AI (*% = confidence level of prediction*)	AI Detection Score 0% Original 100% AI	AI Detection Score 0% Original 100% AI	AI Detection Score 22% Original 78% AI	AI Detection Score 0% Original 100% AI	AI Detection Score 1% Original 99% AI	AI Detection Score 81% Original 19% AI
Sapling	Fake: 98% – 100%	Fake: 100%	Fake: % variable depending on input	Fake: 93.5%	Fake: % variable depending on input	Fake: 0.0% – 2.4%
Winston AI	ChatGPT 3.5: Human Score 0% ChatGPT 4: Human Score 0% – 7%	Human Score 5%	Human Score 100%	Human Score 32%	Human Score 100%	Human Score 100%

Notes:
– Most of the above tests were run on February 23, 2024. DetectGPT results were added on November 24, 2024.
– GPTKit, Plagiarism Check, Scribbr, Undetectable AI, Writer and Writefull were initially tested but were excluded from the analysis due to their inability to consistently detect pure ChatGPT output.
– Such tests provide comparative indications only. Originality AI gives false positives for some human control text, outliers indicating lower reliability of this tool. I did not include the Sapling "% fake" predictions for the reviewer responses due to their inconsistency depending on the text inserted.

Table 3. Results of AI detection tests

Appendix B. ChatGPT simulations

I simulated the use of ChatGPT (3.5 and 4) by feeding versions of my original paper into the model and using the review report questions as prompts [53]. Using the corpus analyzer Sketch Engine, I ran concordances of some of the disproportionately used words and phrases across both the review reports and the ChatGPT simulations [54]. This testing indicated to me that ChatGPT 3.5 was the LLM most likely used to generate the reviews. All related materials are accessible on my website (www.nicospage.eu/publications/ai-peer-review).

The content of the recommendations overlapped between the review reports and the ChatGPT simulations, though not uniformly due to the random nature of each machine response. Globally, the points raised in the review reports are raised in the simulation responses: citing more sources ("existing literature") with less self-citation; providing more "background information"; structure, especially of the introduction and conclusion, and with more signposting and synthesis of points; explanation of the methodological and theoretical basis [26, 53, 54] – all generic aspects, reducible to writing tips, that could be addressed in any paper. One take on this overlap would be that the LLM output aligned with reviewers' independent assessment; a more critical take would be that the reviewers were primed by the LLM responses and signed on to them.

With regard to the text patterns, comparison between the review reports and the ChatGPT simulations revealed many disproportionate uses of the same vocabulary, including multi-token text strings. Table 4 provides some examples (inexhaustive list) [26, 53, 54].

Multi-token text strings	Individual types
articulation of the theoretical and methodological	abrupt
benefit from	broader
deeper understanding	challenging
especially when introducing	coherence
fall short	disjointed
fully explore	elucidate
fully grasp	enhance
further enrich the reader's understanding	enrich
help readers	findings
making it/them challenging	gaps
more robust	progression
persuasiveness of the analysis	rigo(u)r
the author should consider	seamless(ly)
there are instances	solidify
to improve (the) clarity	transition

Table 4. Partial list of matching text patterns found across review reports and ChatGPT simulations

It is striking to note the matching text patterns between my review reports (1772 words) and the tiny corpus (977 words) in the ChatGPT review simulation appended to Donker's letter in *The Lancet Infectious Diseases* [35]: "evaluate the validity of the findings," "reliability of the results/findings," "enhance the credibility," "stronger foundation," "simpler language," "well-structured," along with various other disproportionately repeating lexical types such as "insights," "lack," "contextualize/contextualization," "comprehensive," "thorough," "additionally" [26, 35]. On content, two of the main critiques align as well, as worded in the Donker simulation: "The methodology used in this study is not clearly explained" and "The literature review is not comprehensive" [35; compare to 26]. Whether analyzing work on gay language or infectious disease, ChatGPT says the same things over and over again.

Appendix references

See the main article for other references.

48. Heikkilä M. How to spot AI-generated text. MIT Technology Review. 19 December 2022. www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/
49. Coffey L. Professors cautious of tools to detect AI-generated writing. Inside Higher Ed. 9 February 2024. www.insidehighered.com/news/tech-innovation/artificial-intelligence/2024/02/09/professors-proceed-caution-using-ai#
50. Topinka R. The software says my student cheated using AI. They say they're innocent. Who do I believe? The Guardian. 13 February 2024. www.theguardian.com/commentisfree/2024/feb/13/software-student-cheated-combat-ai
51. OpenAI. New AI classifier for indicating AI-written text. 31 January 2023; updated 20 July 2023 [discontinuation]. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/
52. Gao CA, Howard FM, Markov NS, Dyer E, Ramesh S, Luo Y, Pearson AT. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj Digital Medicine. 2023;6:75.
53. OpenAI. ChatGPT 3.5 and ChatGPT 4 responses generated 3 February 2024. Raw output and synthesis document accessible at www.nicospage.eu
54. Sketch Engine concordances of ChatGPT text and anonymous review reports. Accessible at www.nicospage.eu