

## Detection of AI-generated Writing in Students' Assignments: A Comparative Analysis of Some Tools' Reliability

BENARAB Imen Hanane<sup>1</sup> 

<sup>1</sup>Higher School of Management and Digital Economy, Algeria

Received: 25 / 07 / 2024

Accepted: 15 / 08 / 2024

Published: 30 / 09 / 2024

### Abstract

This article answers the question: *How reliable are current detection tools at identifying human or AI writing in students' assignments?* It aims to test the reliability of these tools through a comparative analysis of 10 of the most popular ones. This enabled us to assess the reliability and robustness of such tools in the face of various writing manipulations that some students may perform while producing their work to hide the artificial origin. We revealed the limitations of each tool taken individually and the need to combine several to overcome their deficiencies and use them to detect the presence of AI writings in students' work.

**Keywords:** Artificial Intelligence, Generative AI, AI writing detector, students' assignments, Open AI

### ملخص

هدف هذا المقال الذي يجيب عن السؤال: ما مدى موثوقية أدوات الكشف الحالية في تحديد الأصل البشري أو الاصطناعي للكتابة في واجبات الطلاب؟ إلى اختبار موثوقية هذه الأدوات من خلال تحليل مقارن بين 10 من التي تعد الأكثر شيوعاً منها. مكنا هذا من تقدير موثوقية وثبات مثل هذه الأدوات في مواجهة التلاعبات المختلفة في الكتابة، التي قد يقوم بها بعض الطلاب أثناء إنجاز أعمالهم بهدف إخفاء الأصل الاصطناعي فيها، وتوصلنا إلى محدودية كل أداة إذا ما أخذت على انفراد وضرورة الجمع بين العديد منها لمحاولة التغلب على هذه المعضلة والاستعانة بها في تقييم استخدام الكتابة بالذكاء الاصطناعي في أعمال الطلاب.

**الكلمات المفتاحية:** الذكاء الاصطناعي، الذكاء الاصطناعي التوليدي، كاشف الكتابة بالذكاء الاصطناعي، واجبات الطلاب، الذكاء الاصطناعي المفتوح.

**Email:** [ibenarab@esgen.edu.dz](mailto:ibenarab@esgen.edu.dz)

Atras Journal/ 2024, published by the University of Saida , Dr. Moulay Tahar, Algeria  
This is an Open Access Article under The CCBY License (<http://creativecommons.org/licenses/by/4.0/>)

## Introduction

Artificial Intelligence (AI) has developed rapidly in recent years and is making a significant contribution to various aspects of human life, including writing. Indeed, although the beginnings of AI date back several decades, its evolution has accelerated in recent years to give rise to generative AI (Kaswan et al., 2023).

ChatGPT is the best-known and most widely used AI language model today (Hoke, 2024). Capable of generating text on a large scale extremely quickly and autonomously, this technology is often used in students' work, whether for presentations, answers to exercises, or writing their dissertations.

This practice has raised the new challenge of differentiating between a text written by a human being and one generated by a machine (Grimes et al., 2023), which teachers frequently face when assessing their students. To address this concern, detection tools soon appeared with the promise of making this distinction possible (Warner, 2023). However, the reliability of these detectors is itself called into question in the face of the rapid evolution of AI language models and their ongoing adaptation (Elkhatat et al., 2023).

This study is part of a very topical theme, responding to current concerns linked to :

- The rise of generative AI and the emergence of high-performance models, such as ChatGPT, have profoundly transformed writing practices and increased the need for reliable detection tools.
- The easy and widespread access to these new technologies, which enable students to generate artificial content in response to any assignment imposed by the university to validate their learning, is increasingly raising questions about the integrity of these assignments.
- The need to find a reliable way of assessing students fairly while ensuring that, despite the possible use of AI, students have made the necessary effort to complete their work.
- The desire of many universities to equip themselves with a tool for detecting writing generated by AI, along with our desire to provide them with answers regarding the reliability of these tools.
- The ethical challenges posed today by the abusive use of AI in academic circles and the possibility of limiting it through detectors.

Through this study, we aim to:

- Compare the performance of the tools in correctly identifying the human or artificial origin of the texts.
- Examine the extent to which manipulation, through the humanization of the texts and paraphrasing, can mislead the detection tools and impact their reliability.
- Identify the situations that can influence the results provided by these tools and cause confusion for teachers when they assess the work done by their students.
- Formulate recommendations for the integration of these tools at the university to ensure fair assessment of students.

In the context described above, the following question arises: *How reliable are current detection tools at identifying human or AI writing in students' assignments?*

The following underlying questions may derive from this issue:

- How do AI detectors work, and on what basis do they judge whether a piece of writing is AI-generated or not?

- To what extent can these tools falsely consider a text written by a human to be the work of an AI and vice versa?
- To what extent can various text manipulations fool these tools?

As preliminary answers to these questions, we propose the following hypotheses:

- AI-generated writing detectors function based on the writing styles that are used to train the machines.
- These tools can consider a text written by a human as being generated by the AI and vice versa insofar as a person could have a style similar to that of the machine, and the continuous evolution of AI tends to bring the outputs of the AI as close as possible to those of humans.
- Since these detection tools are based on AI, it is possible to deceive them by understanding their functioning and finding ways to circumvent their mechanisms.

## Literature Review

### *Generative AI and its Implications*

Generative AI is a field of Artificial Intelligence that can produce original content through text, music, images, or video (Jovanović & Campbell, 2022). These authors identify three essential trends in this technology: the continuous improvement in performance through the use of more sophisticated neural network architectures; the expansion of possible applications beyond the simple generation of text and images; and the emergence of multimodal generative AI, which can analyse and generate data of various types in an integrated way. Indeed, it has developed to the point where it can now even match the creative capacities of humans (Rafner et al., 2023), and these authors suggest that we should consider hybrid human-IA interfaces to enhance creativity in all areas of knowledge.

Concerning scientific research, Bourg et al. (2024) have explored ways of using AI to minimise bias and reduce the burden of the peer review process, while also improving the quality and accessibility of open data, and increasing inclusion in scientific communication. They highlight the limited trust in its processes, which are inconsistent with scientific integrity, and the need to ensure its responsible use to advance open, fair, and trustworthy research (Bourg et al., 2024; Bozkurt, 2024). In this context, some editions require authors to declare the use of these intelligences in their writings, specifying the content generated by the machine (Flanagin et al., 2023; Hosseini & Holmes, 2023). According to these authors, they have even extended this rule to reviewers, forbidding them to use chatbots in their assessment to avoid bias resulting from AIs; this underlines the importance of human supervision and the need to make authors accountable to guarantee the integrity of their writing.

To examine the impact of generative AI-based writing tools on students' written production, Marzuki et al. (2023) conducted an interview study with teachers of English as a Foreign Language (EFL). The goal was to understand how these tools influence the content and organization of their students' writing. They showed that although these tools can enhance students' creativity and provide them with new ideas, over-reliance on them could be detrimental to developing their writing skills and their ability to write independently and critically.

In this context, and to assess teachers' ability to identify texts generated by AI, in particular those from OpenAI's ChatGPT, Fleckenstein et al. (2024) conducted a study with experiments involving 89 novice and 200 experienced teachers. They showed that both novices and experienced teachers had difficulty differentiating between texts generated by the AI and

those written by students and that sometimes, teachers rated AI-generated texts more positively than those written by students, raising concerns about the implications for fair assessment. They also point out that many teachers are overconfident in their ability to recognise AI-generated writing, which can lead to errors in judgement, especially since this technology can effectively mimic students' writing styles, making detection even more complex. They suggest developing strategies to train teachers to recognise AI output, maintain academic integrity and improve assessment methods in teaching.

This issue has even become a concern for many countries that are trying to regulate the use of generative AI (Lo, 2023), such as China, the USA, Europe and the UK, which, while emphasising innovation, are seeking to preserve privacy, ethics and transparency in the use of this technology. For its part, UNESCO proposes a guide to regulating its use in education and research (UNESCO, 2024). This guide focuses on protecting privacy and age adaptation and recommends a human-centered approach to the design and ethical validation of this AI while calling for the establishment of coherent rules to allow to exploration of the possibilities offered by generative AI in teaching and learning. Through this approach, UNESCO aims to ensure that the use of this technology in education and science is safe and meaningful.

In this context, a symposium at Harvard explored how generative AI is transforming education (Manning, 2024). The symposium assessed the impact of generative AI on intellectual property and teaching methods and showcased students' projects exploring the possibilities offered by this technology. This author highlights the profound implications of generative AI for education, which require academic institutions to adapt to meet the challenges posed by this innovation while seizing its opportunities.

### ***AI Writing Detectors***

With the emergence of generative AI followed by the explosion of detectors on the market, much research has focused on these tools, how they work, and whether they deliver on their promises.

The study carried out by Schuster et al. in 2020 already looked at stylometry as a way of detecting writing generated by AI. This technique, which relies on the stylistic features of a text, was proposed at that time as a promising approach for detecting AI-generated texts. The researchers carried out a series of experiments using texts produced by the GPT-2 and GPT-3 language models, as well as those written by humans, to serve as a control group. They subjected them to detector tests based on stylometric analyses and concluded that these approaches had significant limitations in accurately identifying AI-generated texts. According to these researchers, the ability of AI models to effectively mimic human writing styles allows them to deceive detectors based on stylistic analysis, thus calling into question the reliability of these tools in accurately identifying AI-generated texts.

For their part, Sadasivan et al. (2023) carried out their study using the same previous language models in addition to Grover. They generated various texts and had them evaluated by Hugging Face, GPT-2 Output Detector, and Gltr. They also collected human texts for comparison. The results they obtained showed that these tools had an uneven performance, managing to correctly detect a large proportion of the text generated by the AI but also giving a lot of false positives, i.e., they wrongly identified the human text as being generated by the AI, which we think could be very unfair when a teacher has to assess a student's work.

In their work, Elkhatat et al. (2023) also aimed to evaluate the ability of detectors to differentiate between human writings and writings produced by AI models. They focused on

ChatGPT 3.5 and 4 to create 15 paragraphs on a specific topic. They also used five human texts as controls. They tested them through the OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag detectors. Their tests showed that the detection tools performed consistently on most of the texts generated by ChatGPT 3.5, although there were sometimes some uncertain classifications and false negatives. However, their performance was inconsistent on ChatGPT 4 texts, where there were more uncertain classifications and false negatives, and on human texts, with many false positives and uncertain classifications.

The results of the study by Walters (2023), who tested 16 detection tools, also show that although some tools achieve high detection rates, they generate many false positives; that others with better overall performance struggle to detect certain types of AI-generated text and that no tool was able to achieve perfect results, highlighting the continuing challenges in this area.

In his latest research, Chaka (2024) looked at the various existing AI detection approaches and tools. He also carried out an integrative hybrid analysis to critically evaluate the strengths and weaknesses of these tools. He concluded that current detectors had an uneven performance with highly variable detection rates and that no single tool was able to detect all types of text generated by AI reliably.

These studies described in the literature review have all focused on comparing AI-generated texts and human texts. In our work, we will try to go further. We will analyse the robustness of detection tools when they are faced with different text manipulations. This approach will enable us to assess the reliability of these tools under more realistic conditions of use, as students would do.

## **Methods and Materials**

Given the nature of our study and the recency of the theme, we relied on a rich documentary source to try to understand this new phenomenon that is the subject of our research. Then, we opted for the comparative analysis to base our observations on different tools and to provide well-founded and solid answers to our problem.

### ***Research Instruments***

Through our study, we aim to test the AI writing detection tools to assess their reliability as instruments that teachers can use to evaluate the work submitted by their students. There are many tools available, all with the same promise of distinguishing texts produced by different AI language models from those written by humans. Among these detectors, we have selected a sample of 10, some of which are chargeable and some of which are free. For the paid tools, we will use the free trials available, which will give us an idea of their detection potential. These are :

- Free tools: QuillBot, ZeroGPT, ContentatScale, Copyleaks, GPTZero, Scribbr, Sapling.
- Paid tools: Originality.ai, Winston ai, Ai Detector Pro (AIDP).

We chose these tools not only because they are among those designed to detect writing generated by ChatGPT, which is now the model most used by students for writing (Gabriele, 2024), but also because they are among the most reputed and most cited when we looked for the ranking of the best AI detectors (Hartshorne, 2024; Roza, 2024; Driessen, 2024).

Our research relies on a comparative analysis of different AI-generated text detection tools. We will use ChatGPT-4 as our primary model for text generation, given its status as the most widely adopted tool in this domain. We will submit its outputs to various detection tools and compare them with texts generated by ourselves, as well as those resulting from different

manipulations. We will capture the results in screenshots and insert them into our work, followed by comments.

### **Research Procedures**

In this study, we will assess the reliability of detection tools on both AI-generated texts and those created from our writing. Additionally, we will evaluate the robustness of these tools against various forms of text manipulation. We will carry out our tests in 2 stages:

#### *- Stage 1: Testing the raw writings*

We have chosen the topic of ‘*Digital reputation of companies*’ since it is our area of expertise. We are going to start by asking ChatGPT4 to provide us with a definition of the concept of a *company's e-reputation*. Then, we're going to write a paragraph ourselves, based solely on our knowledge without using any technology.

In this first stage, we will submit these writings as they are to the detectors.

#### *- Stage 2. Testing the manipulated texts*

In the second stage, we will test the detectors against the previous texts but with a battery of manipulations, which we will carry out in two series:

##### *First series of manipulations: On the text generated by ChatGPT*

- The first time, we will submit the text initially generated by ChatGPT to ChatGPT itself, instructing it to rewrite it in human style, varying the rhythm and structure of the sentences and using a simple structure.

- On the second occasion, we will take the initial text generated by ChatGPT and submit it to the AIUndetect humanizer tool, which is designed to convert AI-generated texts into more human-like writing.

##### *Second series of manipulations: On the human text*

- We will first submit our initial text to ChatGPT, asking it to improve the writing without adding anything or changing its meaning.

- In a 2nd run, we will mix 2/3 of our initial text with 1/3 of the text generated by ChatGPT.

## **Results**

### ***-Results of Stage One: The raw writing tests***

At this first stage, we asked chatGPT4 to write us a definition of the company's e-reputation, and it came up with the following definition:

*The e-reputation of a company refers to its digital image or perception among online communities and stakeholders. It encompasses how the company is perceived based on online information, including social media, reviews, news articles, and other digital platforms. E-reputation reflects not only the company's brand image but also public sentiment, trustworthiness, and credibility in the digital sphere. Its role is crucial in influencing consumer decisions, investor confidence, and overall business success. Managing e-reputation involves monitoring online mentions, responding to feedback, and strategically enhancing positive perceptions while mitigating negative ones to maintain a favorable digital presence and foster trust.*

Then we wrote our definition of this concept, as follows:

*A company's e-reputation is its reputation on the Internet based on information about itself and others that circulates rapidly and widely online. This reputation can be positive or negative and constitutes an intangible asset for companies, serving as a reference for the company's reputation and as an indicator of confidence for its various stakeholders. It also represents an aggregation of the different reputations that the company may have with its*

stakeholders, built on their expectations and their actual experiences with it, which they exchange over time on the Internet, leading to a collective perception of what the company is.

We then submitted these two texts as they are to the various detectors selected and took screenshots of the evaluation given by each. We present the results of this first stage in Figure One:





Detection tools used	Text entirely generated by ChatGPT	Text entirely written by the author
	100% of text is likely AI-generated	0% of text is likely AI-generated
	Your Text is AI/GPT Generated 100% AI GPT*	Your Text is Human written 0% AI GPT*
	Human Probability Results 😄 READS LIKE AI! ⚠️	Human Probability Results 😞 HARD TO TELL! ?
	AI Content Detected	✓ This is human text
	human mixed ai 0% 0% 100%	human mixed ai 0% 0% 100%
	100% Chance that your text is generated by AI	44% Chance that your text is generated by AI
	Fake: 100.0%	Fake: 100.0%
	0% Original 100% AI	5% Original 95% AI
	Human Score 0%	Human Score 0%
	Probability of AI 98/100	Probability of AI 98/100

Figure 1. The raw writings' tests

In this first case, the results should all be 100% AI generated for the 2<sup>nd</sup> column of the table and

0% AI generated for the last column.

From the evaluation results of the various tools selected, we can see that when the text is entirely generated by ChatGPT, all the tools detect it.

However, for the text that we have entirely written, without any use of technology, we have :

- Only three of the ten tools confirmed this (QuillBot, ZeroGPT, and Copyleaks).
- Two tools were unable to confirm with certainty the human origin of the text (Content at Scale and Scribbr).
- The remaining five tools (GPTZero, Sapling, Originality.ai, Winston ai, Ai Detector Pro) give false positives, falsely assuming that the text was produced by an AI, with rates ranging from 95% to 100%. We would point out that three of these detectors are chargeable.

### **Results of Stage Two: The manipulated writing tests**

In this second stage, we will test the reliability of the detectors when we manipulate the texts. To do this, we have carried out two series of manipulations as described previously in the research procedures.

#### **Results of the first series of manipulations (On the text generated by ChatGPT)**

For this first series of manipulations, we instructed ChatGPT to rewrite the initial text it had generated by simplifying the structure and integrating a more human style. The text thus produced is:

*Business e-reputation is the corporate image or brand awareness that clients, suppliers, customers, and online communities have of a business. It means the overall picture of the company defined by the online information: posts on social media, customer reviews, articles, blogs, and others. E-reputation implies more than just the recognition of the company brand by the public and their perception, trust, and credibility in the website environment. This reputation determines consumer behavior, investors' confidence and general business outcomes. It is the process of active control of all opinions and mentions in the electronic media with a view of controlling negative feelings and enhancing the positive sentiments to ensure customers are more trustful with business*

We also submitted the initial text generated by ChatGPT to AIUndetect, an AI humanizer tool that offers free trials. The text rewritten by this tool is as follows:

*E-reputation denotes the way we perceive a company in the online environment among digital communities and stakeholders. It is the reflection of its visibility, drawing from things such as the company's social networks, forums, news announcements, and other digital spaces. The e-reputation demonstrates not only the brand image of the company but also its emotional aspects, such as how trustworthy and credible the brand is in the digital space. It is a significant element in influencing consumer decisions, investor confidence, and business success in general. Managing the e-reputation involves to follow online mentions, to react to feedback, to address issues, and to create positive perceptions to maintain a favorable online presence.*

We presented these two manipulated texts to the detectors and we present the evaluation results in the following figure:




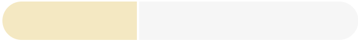









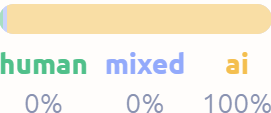
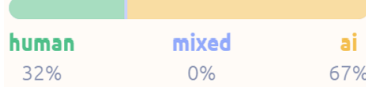










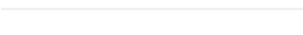
Detection tools used	ChatGPT's Text humanised by ChatGPT	ChatGPT's Text humanised by AIUndetect
	38% of text is likely AI-generated 	66% of text is likely AI-generated 
	Your Text is Human written 	Your Text is Human written 
	Human Probability Results ☹️ HARD TO TELL! 	Human Probability Results ☹️ HARD TO TELL! 
	AI Content Detected	✓ This is human text
		
	100% Chance that your text is generated by AI 	100% Chance that your text is generated by AI 
	Fake: 100.0%	Fake: 0.0%
	0% Original 100% AI	64% Original 36% AI
	Human Score 	Human Score 
	Probability of AI 	Probability of AI 

Figure 2. Tests of the manipulated ChatGPT's writing

In the present case, the attempt to make the texts more similar to those that could be written by humans does not change the fact that the writings are entirely generated by AIs. As a result, the evaluations by the detectors should give 100% for the AI and 0% for the human for the two columns of the table. But we can see from the results that :

For the text humanised using ChatGPT, only three of the 10 tools changed their assessment, either by hesitating about the use of AI (QuillBot and Content at Scale) or by falsely invalidating it (ZeroGPT).

However, when it comes to the text humanised by a dedicated tool, the detectors seem more confused. Except Scribbr, all the others show either hesitations about whether AI has been used or not, or false negatives by falsely assuming that these texts are human.

These humanizer tools therefore call into question the reliability of the detectors, and even though they are chargeable in the majority of cases, students can use them, posing a real problem for teachers who have to assess their work.

### ***Results of the second series of manipulations (On the human text)***

For this last series of manipulations, we asked ChatGPT to improve the text we had initially written without making any changes in the meaning or additions. It suggested the following rewrite:

*An organization's e-reputation is its digital image shaped by widespread online information, encompassing self-generated content and external perceptions. This reputation, whether positive or negative, serves as a vital intangible asset, referencing the company's credibility and acting as a confidence indicator for stakeholders. It consolidates various perceptions held by stakeholders, formed through their expectations and interactions over time. These exchanges across digital platforms collectively shape how the company is perceived. Managing e-reputation involves vigilantly monitoring and strategically responding to online feedback to maintain a favorable digital presence and enhance trustworthiness.*

We have also created a mixed text by combining our original text with the one produced by ChatGPT, as follows:

*A company's e-reputation is its reputation on the Internet based on information about itself and others that circulates rapidly and widely online. This reputation can be positive or negative and constitutes an intangible asset for companies, serving as a reference for the company's reputation and an indicator of confidence for its various stakeholders. It also represents an aggregation of the different reputations that the company may have with its stakeholders, built on their expectations and their actual experiences with it, which they exchange overtime on the Internet, leading to a collective perception of what the company is. This digital image, whether lauded or maligned, forms a nebulous asset, acting as a yardstick for the company's ethos and a gauge of stakeholder trust. It amalgamates varied stakeholder perceptions, crafted from their lofty expectations and real-time interactions shared over the web, resulting in a mosaic of the company's identity.*

We submitted these manipulated texts once again to the detectors for evaluation, and the results were as follows:





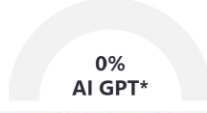
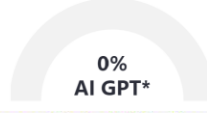





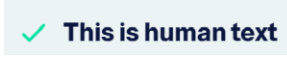

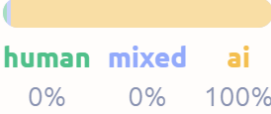
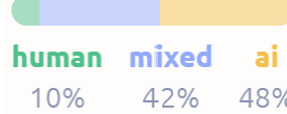


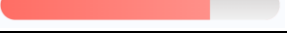








Detection tools used	The author's text enhanced with ChatGPT	A mix of the author's text and ChatGPT's text
 QuillBot	100% of text is likely AI-generated 	100% of text is likely AI-generated 
 ZeroGPT	Your Text is Human written 	Your Text is Human written 
 Content at Scale	Human Probability Results 😄 READS LIKE AI! 	Human Probability Results 😞 HARD TO TELL! 
 Copyleaks	AI Content Detected 	✓ This is human text 
 GPTZero	 human mixed ai 0% 0% 100%	 human mixed ai 10% 42% 48%
 Scribbr	100% Chance that your text is generated by AI 	75% Chance that your text is generated by AI 
 Sapling	Fake: 0.0%	Fake: 100.0%
 Originality.ai	0% Original 100% AI	1% Original 99% AI
 winstonai	Human Score 	Human Score 
 AIDP	Probability of AI 	Probability of AI 

Figure 3. Tests of the manipulated human writing

In this latter case, the score for the second column should be 0% for the AI and 100% for the human since the human only relied on the AI as a tool to improve his writing and was not, strictly speaking, substituted by it to carry out his work. For the last column, the scores should be 66 to 67% for the human and 33 to 34% for the AI, given that the mixed text included 2/3 human input and 1/3 from the AI.

However, analysis of the results shows that for eight tools out of 10, the AI-enhanced text is considered to be entirely generated by this intelligence, and only two tools (ZeroGPT

and Sapling) consider that it is still of human origin. This situation poses a problem if a student makes an effort to write a piece of work and then uses this technology to improve it, as there is a risk that the detectors will falsely assess it as AI-generated.

The same problem arises for mixed text where no detector has given the right proportion for the AI contribution. Some of the tools did not detect AI-generated writing at all (ZeroGPT and Copyleaks), while others hesitated or falsely attributed all the text to AI (QuillBot, Sapling, and Winston ai).

## **Discussion**

Through the results of the various tests, we can answer the research questions by saying that the detection tools are not reliable. They can falsely consider a text written by a human as AI-generated and vice versa. Additionally, text manipulations can deceive these tools. Indeed, when we submitted the original texts to the detectors- one generated entirely by ChatGPT and the other written by humans- we found that all the tools successfully identified the text produced by ChatGPT. However, only a few were able to confirm the human origin, while most showed hesitations or gave false positives, incorrectly stating that human-written texts were generated by AI.

These results are in line with those of Sadasivan et al. (2023), Elkhatat et al. (2023), Walters (2023), and Chaka (2024) that we discussed in our literature review, where these authors came to the same conclusion that the tools were highly reliable for detecting texts produced by AI, but that they had limitations when the texts are entirely human.

We tried to take the research a step further by testing the detectors against various manipulations of the texts. Indeed, instead of presenting a piece of writing entirely generated by the AI, which as we have seen is easily detected by the tools, a student could try to mislead them in various ways. We have observed that the problem of the reliability of the detectors also arises when the texts are manipulated.

We have seen that the use of humanizers now available on the Internet distorts the results and calls into question the reliability of the detectors. The same is true for the instructions that can be given to ChatGPT to rewrite its text differently by imitating the human style, by giving it a model of an author or a journalist for example. Students can use these kinds of manipulations to make an AI-generated piece of writing more like a human one and obtain validation from the detectors. In this way, a teacher who uses these detectors before assessing the student could wrongly give him a good assessment when he has entrusted his work entirely to AIs.

We were also able to observe that the detectors could induce errors of judgment even in cases where the AI is used solely as a writing aid. When we asked the AI to improve our text, most of the detectors considered the result as AI-generated. Similarly, when there was a mix where a more significant proportion of the writing was of human origin, most of the detectors considered that it was entirely created by the AI. In these two cases, the work was essentially human and the AI provided only assistance and not the entire text.

## **Pedagogical Implications**

After our study, we were able to see that teachers cannot use AI detection tools reliably to judge the origin of a text and assess on their basis the work submitted by students with certainty. Indeed, having tested various situations, we saw that these tools' accuracy was only proven when the texts were entirely written by the AI. However, as soon as we manipulated the texts, whether through reformulations, changes in style, or rewritings using humanizer tools,

the detectors lost their accuracy. It is these rewritings or the mix between technology and human writing that corresponds most to reality. If a student had to produce a piece of work, it would be difficult to forbid him from using IA technologies, and at the same time, unfair to under- or over-assess him by relying on a detection tool and the assessment it makes of his work.

Each of the tools we have selected, whether gratis or chargeable, has shown its limitations, either through hesitations when judging specific texts, or through false positives or false negatives depending on the case, and none of them has succeeded in being consistent across all the tests we have carried out. It is therefore risky to rely on a single tool to carry out this type of assessment and to be able to judge with certainty the origin of a text.

To attempt to detect the use of AI in student work, we advise universities wishing to equip themselves with detectors to consider that these tools are not infallible and thus take their evaluation as indicative only. The teacher should not rely entirely on this tool and should investigate further, for example, by subjecting the student to an oral test to determine how much personal effort he has put into his work. Furthermore, a combination of tools should be considered rather than a single tool, as even the best-paying tools are not consistent in their assessment, and it is by combining several that it would be possible to reduce the risk of error.

## **Conclusion**

This paper tried to assess how reliable the current AI writing detection tools are at identifying human or AI writing in students' assignments to guide teachers and universities wishing to equip themselves with such tools to assess students' work somewhat in the current context dominated by the use of AI in general and generative AI in particular.

To do this, we conducted a comparative analysis based on 10 of the most reputable and widely used detection tools. Some of these tools are chargeable and some are gratis. We submitted to these detectors ChatGPT-generated texts and human texts. We also submitted to them manipulated texts, mixed texts, and AI-improved texts. These manipulations correspond to the reality of AI use today.

Through the results of the various tests, we could note the ability of all these tools to detect texts written entirely by AI effectively. However, all of them, without exception, lack precision when the texts are human or manipulated in such a way as to make them more similar to human style, or when the AI has been asked to help improve the writing, showing hesitations in classification, as well as false negatives or false positives.

We concluded that today, no single tool is yet capable of estimating with certainty and precision the percentage of writing generated by AI in students' works. This fact calls into question the possibility of a fair and equitable assessment of these works based on a unique tool and underlines the need to combine several to make up for this shortcoming. In this way, teachers will be able to use the detectors to help them assess their students by getting an idea of the originality of their writing without basing the entire assessment on these detectors' results.

## **About the Author**

**Imen Hanane BENARAB**, PhD in Commercial Sciences, a graduate of the School of Higher Commercial Studies- EHEC. Currently Lecturer A at the Higher School of Management and Digital Economy- ESGEN-Kolea (Tipaza). Teaching at this school since 2012, with in-depth expertise in e-business, management and information and communication technologies (ICT), the author is passionate about research and digital technology. After defending a doctoral thesis on digital reputation management in the Web

2.0 era, she introduced the *e-reputation* module in ESGEN's education program. Through her work, she aims to pass on her knowledge and advanced thinking in the fields of management and digital technology. <https://orcid.org/0009-0001-5070-6584>

### **Declaration of AI Tools Use**

As part of writing my article on the comparative analysis of AI text detection tools, and given the nature of this research, I had to use several AI tools. I used ChatGPT to generate the texts produced by the machine I needed to test. In addition, I used ten detection tools to evaluate these generated texts. To humanise the content generated by ChatGPT prior to these tests, I also used the AIUndetect humanizer tool.

It is important to note that using these tools can introduce standardised patterns that are characteristic of AI-generated content. As a result, a certain proportion of the content could reflect linguistic structures specific to AI. However, all of the intellectual content and analysis presented in this article remain the fruit of my personal work and reflection.

### **Statement of Absence of Conflict of Interest**

The author declares that there are no conflicts of interest regarding the publication of this article. No financial, personal, or professional interests have influenced the research, writing, or conclusions of this work.

## References

- Bourg, C. et al. (2024). Generative AI for trustworthy, open, and equitable scholarship. *An MIT Exploration of Generative AI*. <https://doi.org/10.21428/e4baedd9.567bfd15>.
- Bozkurt, A. (2024). GenAI et al.: Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. *Open Praxis*, 16(1). 1-10. doi: 10.55982/openpraxis.16.1.654
- Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning & Teaching*, 7(1),1-12.
- Driessen, K. (2024). Best AI detector: Free & premium tools compared <https://www.scribbr.com/ai-tools/best-ai-detector/>
- Elkhatat, A.M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1). 1-16. <https://doi.org/10.1007/s40979-023-00140-5>
- Flanagin, A., Kendall-Taylor, J., & Bibbins-Domingo, K. (2023). Guidance for authors, peer reviewers, and editors on the use of AI, language models, and chatbots. *JAMA*. 330(8). 702–703.
- Fleckenstein, J., Meyer, J., Jansen, T., Keller, S.D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*. Vol. 6. <https://doi.org/10.1016/j.caeai.2024.100209>.
- Gabriele, N. (2024). *Peut-on se fier aux détecteurs d'IA afin de contrôler un texte ou du contenu web ?* Available at <https://scribweb.fr/fiabilite-detecteurs-ia/>
- Grimes, M., Von Krogh, G., Feuerriegel, S., Rink, F., & Gruber, M. (2023). From scarcity to abundance: Scholars and scholarship in an age of generative Artificial Intelligence. *Academy of Management Journal*, 66(6), 1617-1624.
- Hartshorne, D. (2024). *The 6 best AI content detectors in 2024*. Available at <https://zapier.com/blog/ai-content-detector/>
- Hoke, T. (2024). *Relying on generative AI in scholarly work has its pitfalls*. ASCE. Available at <https://www.asce.org/publications-and-news/civil-engineering-source/civil-engineering-magazine/issues/magazine-issue/article/2024/03/relying-on-generative-ai-in-scholarly-work-has-its-pitfalls>
- Hosseini, M., & Holmes, K. (2023). The evolution of library workplaces and workflows via generative AI. *College & Research Libraries*, 84(6), 836-842. doi: 10.5860/crl.84.6.836
- Jovanović, M. & Campbell, M. (2022). Generative Artificial Intelligence: Trends and prospects. *Computer*, 55(10). 107-112.
- Kaswan, K.S., Dhattewal, J.S., Malik, K., & Baliyan, A. (2023). Generative AI: A review on models and applications. *International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*. November. Greater Noida. India. doi:10.1109/ICCSAI59793.2023.10421601
- Lo, L. S. (2023). AI policies across the globe: Implications and recommendations for libraries. *IFLA Journal*. 49(4). 645-649.

- Manning, A.J. (2024). What is original scholarship in the age of AI? *The Harvard Gazette*. Available at <https://news.harvard.edu/gazette/story/2024/05/how-is-generative-ai-changing-education-artificial-intelligence/>
- Marzuki, M., Widiati, U., Rusdin, D., Darwin, R., & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education*, 10(2). <https://doi.org/10.1080/2331186X.2023.2236469>
- Rafner, J., Beaty, R.E., Kaufman, J.C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11), 1836-1838.
- Roza, N. (2024). *The best AI content detection tools and services for 2024*. Available at <https://nikolaroza.com/best-ai-content-detectors-tools-apps/>
- Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected?* ArXiv, abs/2303.11156.
- Schuster, T., Schuster, R., Shah, D.J., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(302). 1-18. doi:10.1162/coli\_a\_00380
- UNESCO. (2024). *Guidance for generative AI in education and research*. Available at <https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research>
- Walters, W. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, 7(1), 20220158. <https://doi.org/10.1515/opis-2022-0158>
- Warner, A. (2023). *The false promise of generative AI detectors*. Available at <https://multilingual.com/the-false-promise-of-generative-ai-detectors/>

#### Cite as

Benarab, I.H. (2024). Detection of AI-Generated Writing in Students' Assignments: A Comparative Analysis of Some Tools' Reliability. *Atras Journal*, 5 (Special Issue), 271-286.