Vol.8 No.1 (2025)

Journal of Applied Learning & Teaching

ISSN : 2591-801X

Content Available at : http://journals.sfu.ca/jalt/index.php/jalt/index

# AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini?

Muhammad Abid Malik[A]

Amjad Islam Amjad[B]

[A] *Academic Research Advisor, Shandong Vocational University of Foreign Affairs, Weihai, China*

[B] *School Education Department, Government of Punjab, Kasur, Pakistan*

## Keywords

AI-detection tool;
artificial intelligence;
ChatGPT;
large language model;
adversarial technique.

## Correspondence

m_abidmalik7@yahoo.com [A]

## Article Info

Received 11 October 2024
Received in revised form 9 November 2024
Accepted 25 December 2024
Available online 13 January 2025

**DOI:** https://doi.org/10.37074/jalt.2025.8.1.9

## Abstract

AI chatbots and LLMs have made a significant impact in a short time. Despite their benefits, they pose serious threats to academic integrity and ethics by generating human-like text, which is very hard to detect. Various AI-detection tools have been developed to tackle this issue. However, their effectiveness is questionable. This study investigates the performance of four AI-detection tools (Turnitin, ZeroGPT, GPTZero, and Writer AI) in detecting AI-generated text. That text was generated using three LLMs (ChatGPT, Perplexity, and Gemini). Furthermore, three adversarial techniques (edited through Grammarly, paraphrased through Quillbot, and 10%-20% editing by a human expert) were applied to see their effects on the performance of AI-detection tools. Turnitin turned out to be the most accurate and consistent one, with a 100% AI score even with the adversarial techniques. ZeroGPT and GPTZero also reported relatively high AI scores, especially with the original files and the first and third adversarial techniques. Among the three adversarial techniques, paraphrasing through Quillbot affected the performance of three AI-detection tools (ZeroGPT, GPTZero, and Writer AI) the most. Among the three LLMs, text generated through Perplexity was more accurately detected, while Gemini-generated text showed a relatively lower AI score. What was the most note-worthy was the fact that in many cases, even when the text was generated through the same LLM, and detected through the same AI-detection tool; different files showed different AI scores, further highlighting the inconsistencies among AI-detection tools.

## Introduction

Artificial intelligence (AI) has made an immense impact on human life in a short span of time (Jiang et al., 2022; Malik, 2024). AI chatbot is a software-based electronic system that emulates conversations by responding to recognized keywords or phrases. They have evolved to the extent that they are executing activities that usually necessitate human ability (Chandra et al., 2022; Nawaz & Gomes, 2019), such as reading and understanding language, analyzing data, identifying patterns, writing computer programming codes, developing new medicines, and solving complex problems (Bawack et al., 2021; Kaul et al., 2020; Liu & Li, 2024; Raisch & Fomina, 2023; Rusmiyanto et al., 2023).

AI chatbots and large language models (LLMs) have been extensively employed for different writing tasks (Huang et al., 2023; Lee & Yoon, 2021; Malik et al., 2024). Various programs and platforms like ChatGPT, Bing, Grammarly, and Hemingway employ AI to offer immediate feedback on spelling, style, and structure to improve the quality of work (Barbetta, 2023; Rasul et al., 2023; Reza et al., 2023). In addition, AI-powered chatbots and LLMs can enhance communication between the users and the software, facilitating immediate cooperation and feedback (Gill et al., 2024; Song & Song, 2023).

LLMs gained popularity and started to be used widely for text generation (Amjad et al., 2024; Hussain & Qazi, 2023; Malik, 2024; Malik et al., 2024). With their ability to think, synthesize, and generate text similar to human writing; they have been transforming academic and non-academic writing (Bates et al., 2020; Carobene et al., 2024; Dwivedi et al., 2021); and are widely used for summarization, literature review, developing manuscripts and to support research (Rasul et al., 2023; Xames & Shefa, 2023). As a result, it provides more opportunities to the researchers and scholars by assisting them in their academic and research work, and saving their time (Malik et al., 2024). Despite the advantages and benefits that ChatGPT brings with it, there are also multiple apprehensions and concerns about its negative use and adverse impact (Hutson, 2022; Malik et al., 2024). There have been fears about its negative influence on writing and cognitive skills. Shidiq (2023) said that "relying too much on ChatGPT can make individuals weak in thinking critically" (p. 354). Multiple studies have also expressed fears that overdependence on ChatGPT may degrade students' writing skills (Malik, 2024; Malik et al., 2024).

It started with the advent of ChatGPT in November 2022, owned by OpenAI (OpenAI, 2022). It can produce human-like text for different questions and contexts (Malik, 2024). It can also be employed for various other tasks, such as creating social media materials, computer coding, and responding to customer service questions (Kocoń et al., 2023; Taecharungroj, 2023).

Following the success of ChatGPT, more LLMs were developed (Javaid et al., 2023; Kooli & Yusuf, 2024; Onal & Kulavuz-Onal, 2024). Despite having similar purposes and objectives, they come with their own unique set of strengths and weaknesses. Perplexity, another LLM, was launched in 2022. It is extensively used for text generation and provides users with textual responses to their queries (Iorliam & Ingio, 2024). Like other LLMs, it has the ability to analyze, review, and write text with human-like characteristics. It is widely used for brainstorming ideas, developing an outline for the topic, and generating relevant citations and references (Tilwani et al., 2024).

Gemini is another well-known LLM that utilizes Google's advanced language models (Hasanein et al., 2024; Mainaly, 2023). With its ability to respond to diverse prompts, it can also produce text resembling human language, making it a valuable tool for various tasks and purposes (Guo et al., 2023). Due to these qualities, it is becoming increasingly popular, especially among students (Hasanein et al., 2024).

### Increased fears of academic fraud and plagiarism due to LLMs

As computers, the internet, and easy access to digitalized materials made it easier to plagiarize (Malik et al., 2021), the rise of various LLMs is leading to a higher number and more sophisticated cases of plagiarism (Dwivedi et al., 2023; Malik, 2024; Motlagh et al., 2023; Rasul et al., 2023; Sullivan et al., 2023; Xames & Shefa, 2023). Their ability to generate human-like texts and difficulty in detecting them makes those LLMs ideal for those who are looking for shortcuts (Alsabhan, 2023; Malik, 2024; Malik et al., 2024). "LLMs therefore represent a clear potential threat to academic integrity as academic staff may be unable to identify the amount of content produced by a student correctly" (Perkins, 2023, p. 7). Malik et al. (2024) also said that ChatGPT would "increase plagiarism in academic writing due to its ease of use and ability to generate human-like text" (p. 9). In the same study, one participant pointed out how many school students in Trinidad and Tobago had been using ChatGPT to generate assignments for creative writing tasks (Malik et al., 2024). Haider et al. (2024) conducted a study to trace ChatGPT-generated papers on Google Scholar. They downloaded a sample of scientific papers with signs of GPT-use from the website. The study found that almost two-third of the sampled papers used ChatGPT fraudulently or did not declare its use. Most of them were from health science, computer science, and environmental studies. Liang et al. (2024) conducted a large-scale study across 950,965 papers that were published from January 2020 to February 2024 in the arXiv, bioRxiv, and Nature portfolio journals. They found that the use of LLMs in research papers was increasing over the years with the fastest growth observed in the field of computer science (up to 17.5%). In October 2024, it was reported on Retraction Watch that since September, Springer Nature had retracted over 200 papers due to malpractices including fraudulent or undeclared use of AI (Chawla, 2024). Many researchers have, therefore, called for increased vigilance and measures to detect such malpractices (Cingillioglu, 2023; Gustilo et al., 2024; Malik et al., 2024). Different steps, such as implementing strict policies and penalties for AI-generated text and using advanced AI-detection tools, have been recommended to tackle this issue (Elkhatat et al., 2023).

### Use of AI-detection tools to trace AI-driven plagiarism

With the advent of AI chatbots and LLMs, AI-detection tools have also become widespread to check AI-driven plagiarism (Alhijawi et al., 2024; Carobene et al., 2024; Dwivedi et al., 2023; Gustilo et al., 2024). These tools utilize different algorithms and AI approaches to examine the text and compare it with an extensive library of sources to detect any occurrences or patterns that may flag it as AI-generated (Nwohiri et al., 2021).

Turnitin, ZeroGPT, GPTZero, Copileaks, Writer AI, and Winston AI are some of the AI-powered tools for detecting AI-driven plagiarism (Ladha et al., 2023). They utilize advanced AI functions to identify AI-generated text (Odri & Yoon, 2023). The algorithm can analyze text in many languages and claim to accurately identify AI-generated text with a high precision rate (Ladha et al., 2023). These tools can identify plagiarism with or without different adversarial techniques with varying degrees of accuracy (Arabi & Akbari, 2022; Mitchell et al., 2023; Perkins et al., 2024). In addition, many of these tools have the ability to identify plagiarism across various languages, rendering them valuable in different professional and educational contexts (Fairooz et al., 2023). Some can also be integrated into learning management systems (LMS) and other platforms, making it easier to detect AI-driven plagiarism (Kumar, 2023). They also offer extensive data and analytics, helping stakeholders monitor plagiarism patterns and pinpoint AI-generated text (George & Wooden, 2023).

However, studies have shown mixed results regarding the accuracy and efficiency of those AI detection tools. Weber-Wulff et al. (2023) concluded that AI-detection tools were unreliable and did not show consistent results. It was further reinforced by Odri and Yoon (2023), who generated text through ChatGPT4 and detected AI-generated text using eleven different AI-detection tools. They found that most of them presented AI-generated text as human-written text. Elkhatat et al. (2023) carried out a study to examine the efficiency of five AI-detection tools for text generated by two versions of ChatGPT (3.5 and 4). They also reported the inconsistencies and false-positive scores of AI-detection tools. When Foster (2023) used Turnitin to detect a text for AI that was entirely generated by ChatGPT4, it showed a 0% AI score.

Ladha et al. (2023) conducted a study to explore the efficiency of different AI-detection tools, i.e. Copyleaks, Writer, and Content@scale. The study found that those AI-detection tools were inconsistent in distinguishing between AI and human-generated texts. Akram (2023) also tested the accuracy of six AI-detection tools (GPTZero, GPTkit, Originality, Writer, Sapling, and Zylalab) and found that they showed inconsistent results, with their accuracy ranging from 55.29% to 97%. In another study, Walters (2023) investigated the accuracy and effectiveness of 16 AI-detection tools for 42 essays produced by AI and humans. It found that most of the tools were inconsistent in detecting AI-generated text; however, Copyleaks, Originality.AI, and Turnitin were more accurate, efficient, and consistent. The study also found that registered and paid tools were more reliable and accurate than the free ones.

Chaka conducted a few studies investigating the effectiveness of different AI-detection software and tools in tracing AI-generated text. In 2023, one study compared the effectiveness of five AI-detection tools (GPTZero, OpenAI Text Classifier, Writer, Copyleaks, and GLTR) in identifying text generated by three AI chatbots (ChatGPT, YouChat, and Chatsonic). In that study, Copyleaks turned out to be the most effective in detecting AI-generated text (Chaka, 2023). In another study in which Chaka reviewed 17 journal articles, Crossplag was found to be the most effective AI-detection tool, followed by Copyleaks (Chaka, 2024a). In the same year, he conducted another study, this time evaluating the effectiveness of thirty freely available non-premium AI-detecting tools for detecting AI-generated text in the essays written by L1 and L2 university students. The study investigated the accuracy, false-positive, and true-negative rates of those AI-detecting tools. Only two of the tools (Copyleaks and Undetectable AI) were able to identify the essays as human writings accurately. The study concluded that most of the freely available AI-detecting tools showed inconsistent results and were not very effective (Chaka, 2024b).

Some researchers also used different adversarial techniques to check how effective they were in evading the detection of AI-generated text. Mitchell et al. (2023) found that once AI-generated text was paraphrased using an automated paraphrasing tool (APT), the detection rate drastically reduced from 70.3% to 4.6%. Perkins et al. (2024) investigated the effectiveness of seven AI-detection tools (Turnitin AI detector, GPTZero, ZeroGPT, Copyleaks, Crossplag, GPT-2 Output Detector, and GPTKit) to detect AI-generated text in different human-written and AI-generated texts (generated by Bard, Claude 2, and GPT-4). They used six adversarial techniques (i.e. adding spelling errors, increasing burstiness, paraphrasing, decreasing complexity, writing as a non-native English speaker, and increasing complexity) to evade detection. The study showed that those adversarial techniques had different degrees of success in evading detection (drop in accuracy for different adversarial techniques: adding spelling errors 27%; increasing burstiness 24%; paraphrasing 21%; decreasing complexity 19%; writing as non-native English speaker 12%; increasing complexity 2%). Overall, Copyleaks and Turnitin were the top two AI-detection tools, while ZeroGPT was the least effective.

All these studies show inconsistent results for AI-detection tools in identifying AI-generated text. Not only are they inconsistent in detecting AI-generated text, but they also show inconsistencies in false-positive and true-negative results. However, as the LLMs and AI-detection tools are evolving at an incredible pace, it is important to continue conducting research to find the most efficient ones.

The current study further contributes to the existing literature in this area by detecting the text generated by three different LLMs (ChatGPT 3.5, Gemini, and Perplexity) through four AI-detection tools (Turnitin, ZeroGPT, GPTZero, and Writer AI). It not only checks the original files generated through those LLMs but also carries out three adversarial techniques to further check their efficiency and accuracy in detecting AI-generated text. More specifically, the study has the following research objectives.

- To investigate the effectiveness of four AI-detection tools (Turnitin, ZeroGPT, GPTZero, and Writer AI) for text generated through three LLMs (ChatGPT 3.5, Gemini, and Perplexity) without any adversarial technique
- To investigate the effectiveness of four AI-detection tools (Turnitin, ZeroGPT, GPTZero, and Writer AI) for text generated through three LLMs (ChatGPT 3.5, Gemini, and Perplexity) with a first adversarial technique (edited by Grammarly)
- To investigate the effectiveness of four AI-detection tools (Turnitin, ZeroGPT, GPTZero, and Writer AI) for text generated through three LLMs (ChatGPT 3.5, Gemini, and Perplexity) with a second adversarial technique (paraphrased by Quillbot)
- To investigate the effectiveness of four AI-detection tools (Turnitin, ZeroGPT, GPTZero, and Writer AI) for text generated through three LLMs (ChatGPT 3.5, Gemini, and Perplexity) with a third adversarial technique (10-20% editing by a human expert).

## Methods and materials

The main purpose of this study was to determine the effectiveness and accuracy of different AI-detection tools for identifying text generated through different LLMs. Furthermore, it investigated if the performance of those AI-detection tools was in any way affected by automated or human adversarial techniques. Figure 1 further explains the research process.

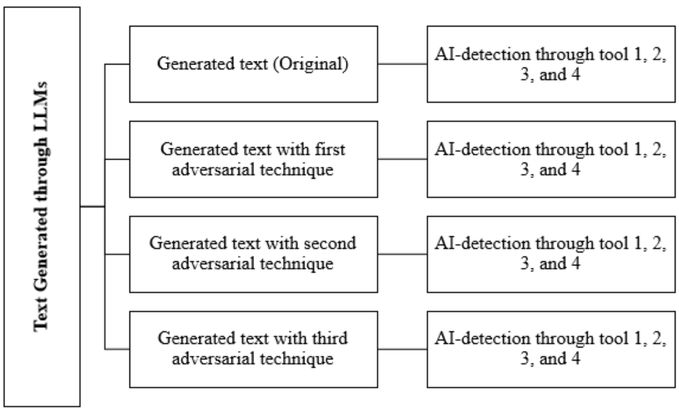


Figure 1. Research process.

### Selection of LLMs and text generation

After developing the process, we looked at different LLMs to generate text for this study. Both free and paid versions/ LLMs are available. Free ones are usually less advanced, while paid ones have better features and modelling abilities (Walters, 2023). However, we decided to use free versions for generating text as they are more easily accessible and more commonly used by the students. After running through different search engines, we opted to use ChatGPT 3.5, Gemini, and Perplexity to generate text. Although both Application Programming Interface (API) and websites are available for all three LLMs, we used websites for text generation. Further details about those LLMs are given in Table 1.

Table 1. Details and features of LLMs.

| Name | ChatGPT 3.5 | Perplexity | Gemini |
|---|---|---|---|
| Developed by | OpenAI | Perplexity AI | Google Deepmind |
| First launched in | November 2022 | August 2023 | March 2023 |
| Paid or Free | Free | Free | Free |
| Latest version | ChatGPT 4 | Perplexity LLM2 | Gemini 1.5 |
| Type of Modelling | Unimodalling (advanced, paid versions also have multimodally capabilities) | Unimodalling (some multimodal capabilities are available through subscriptions, etc.) | Multimodalling |
| Access | Both API and website | Both API and website | Both API and website |

Once the LLMs were selected, the study proceeded to develop a prompt to generate the text. As a pilot, we developed five prompts and generated texts from the LLMs mentioned above. After looking at the generated texts, we selected the following prompt: "Generate an essay for an undergraduate student about Higher Education in America: Challenges, Strategies, and Future Opportunities. The essay should be around 500 words long." To make the task manageable, the generated essays had a maximum word limit of 500 words each. 15 essays were generated at this stage (five each from ChatGPT, Perplexity, and Gemini).

### Further adversarial techniques

The study aimed not only to find the effectiveness and accuracy of AI-detection tools for text generated by different LLMs but also to see if (and how much) any software-based or human adversarial technique would affect their performance.

After talking to the students and looking at the literature, it was decided to carry out three adversarial techniques that we thought were the most commonly used to evade detection: editing the generated text through Grammarly, paraphrasing it through Quillbot, and editing by a human expert.

Grammarly is a software that is commonly used to edit and correct language. It suggests changes to the user, which the user can accept or dismiss. We decided to accept all the changes Grammarly suggested. It has both paid and free versions; however, this study used the premium/ paid version. One set of all fifteen originally generated essays was edited through Grammarly.

Quillbot, developed in 2017, can suggest language editing, paraphrasing, and sentence completion. Students frequently use it to paraphrase plagiarized text to avoid similarity index detection. Its recent versions use AI. Quillbot has both free and paid versions. For this study, we used the paid version.

The third adversarial technique involved human editing. An English language expert was asked to read one set of text generated by each of the three LLMs and make 10% to 20% editing. Overall, 60 essays were generated for this study (15 original and 45 with three adversarial techniques).

### Selection of AI-detection tools and the process

Finally, we proceeded to select AI-detection tools for this study. Since the inception of LLMs, different software and tools have been developed to track AI-generated text (Chaka, 2023; Elkhatat et al., 2023; Ladha et al., 2023). We made a list of the available AI-detection tools and then decided to use four of them (Turnitin, ZeroGPT, GPTZero, and Writer AI) in this study.

Turnitin is a well-renowned software used to check similarity indexes. In 2023, it also added features to detect AI-generated text. ZeroGPT, GPTZero, and Writer AI are AI-based tools that can be used not only to detect AI-generated text but also for text generation, summarization, and editing.
All 60 pieces of generated texts (15 original and 45 with adversarial techniques) were checked through these four AI-detection tools.

## Results

This section discusses the performance of four AI-detection tools in accurately detecting AI-generated text with and without adversarial techniques.

### Comparing the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (original)

First, all the original essays generated through ChatGPT, Perplexity, and Gemini were checked through Turnitin, ZeroGPT, GPTZero, and Writer AI for AI-detection. The results (Table 2) show that Turnitin was able to detect all the files perfectly with 100% detection. ZeroGPT and GPTZero also had relatively high AI-detection rates. ZeroGPT was able to detect Perplexity-generated text with 100% accuracy. Even for ChatGPT and Gemini, its AI scores were quite high (for ChatGPT, range 97%-100%, average 99.4%; for Gemini, range 82%-100%, average 95.4%). GPTZero had a 100% accuracy rate for both Perplexity and Gemini; however, for ChatGPT, its accuracy rate was slightly lower (range 92%-100%, average 97.2%). Writer AI performed quite poorly for all three LLMs with very low average AI scores (for ChatGPT, range 32%-35%, average 34.6%; for Perplexity, range 31%-35%, average 33%; for Gemini, range 26%-28%, average 26.8%).

Table 2. Comparison of the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (original).

| AI-generated essay (Original) | Turnitin (%) | ZeroGPT (%) | GPTZero (%) | Writer AI (%) |
|---|---|---|---|---|
| ChatGPT1 | 100 | 100 | 92 | 34 |
| ChatGPT2 | 100 | 97 | 100 | 35 |
| ChatGPT3 | 100 | 100 | 98 | 32 |
| ChatGPT4 | 100 | 100 | 98 | 37 |
| ChatGPT5 | 100 | 100 | 98 | 35 |
| Perplexity1 | 100 | 100 | 100 | 35 |
| Perplexity2 | 100 | 100 | 100 | 34 |
| Perplexity3 | 100 | 100 | 100 | 33 |
| Perplexity4 | 100 | 100 | 100 | 31 |
| Perplexity5 | 100 | 100 | 100 | 32 |
| Gemini1 | 100 | 82 | 100 | 28 |
| Gemini2 | 100 | 100 | 100 | 26 |
| Gemini3 | 100 | 100 | 100 | 26 |
| Gemini4 | 100 | 96 | 100 | 28 |
| Gemini5 | 100 | 99 | 100 | 26 |

### Comparing the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (with the first adversarial technique - edited through Grammarly)

In the second stage, all AI-generated essays with the first adversarial technique were checked through those four AI-detection tools. The results (Table 3) show that Turnitin was again able to detect all the files with the first adversarial technique with a 100% detection rate. ZeroGPT and GPTZero detected Perplexity-generated files perfectly and had a quite high detection rate for ChatGPT and Gemini (ZeroGPT: for ChatGPT range 97%-100%, average 99%; for Gemini range 81%-100%, average 95.2%; GPTZero: for ChatGPT range 98%, average 98%; for Gemini range 90%-100%, average 98%). Writer AI once more performed quite poorly in detecting AI-generated content with the first adversarial technique (for ChatGPT, range 30%-33%, average 32%; for Perplexity, range 29%-32%, average 31%; for Gemini, range 24%-27%, average 25.8%).

### Comparing the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (with second adversarial technique-paraphrased through Quillbot)

In the third step, all AI-generated essays with the second adversarial technique were checked for AI-generated text. The results (Table 4) show that even when all AI-generated text was paraphrased through Quillbot, Turnitin was able to

Table 3. Comparison of the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (with the first adversarial technique- edited through Grammarly).

| AI-generated essay (Edited by Grammarly) | Turnitin (%) | ZeroGPT (%) | GPTZero (%) | Writer AI (%) |
|---|---|---|---|---|
| ChatGPT1 | 100 | 100 | 98 | 32 |
| ChatGPT2 | 100 | 97 | 98 | 30 |
| ChatGPT3 | 100 | 100 | 98 | 33 |
| ChatGPT4 | 100 | 98 | 98 | 33 |
| ChatGPT5 | 100 | 100 | 98 | 32 |
| Perplexity1 | 100 | 100 | 100 | 32 |
| Perplexity2 | 100 | 100 | 100 | 32 |
| Perplexity3 | 100 | 100 | 100 | 31 |
| Perplexity4 | 100 | 100 | 100 | 29 |
| Perplexity5 | 100 | 100 | 100 | 31 |
| Gemini1 | 100 | 81 | 100 | 26 |
| Gemini2 | 100 | 100 | 100 | 26 |
| Gemini3 | 100 | 100 | 90 | 24 |
| Gemini4 | 100 | 96 | 100 | 27 |
| Gemini5 | 100 | 99 | 100 | 26 |

Table 4. Comparison of the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (with second adversarial technique- paraphrased through Quillbot).

| AI-generated essay (Edited through Quillbot) | Turnitin (%) | ZeroGPT (%) | GPTZero (%) | Writer AI (%) |
|---|---|---|---|---|
| ChatGPT1 | 100 | 32 | 23 | 14 |
| ChatGPT2 | 100 | 57 | 78 | 14 |
| ChatGPT3 | 100 | 29 | 23 | 13 |
| ChatGPT4 | 100 | 66 | 82 | 12 |
| ChatGPT5 | 100 | 81 | 44 | 15 |
| Perplexity1 | 100 | 64 | 99 | 14 |
| Perplexity2 | 100 | 28 | 100 | 13 |
| Perplexity3 | 100 | 68 | 100 | 16 |
| Perplexity4 | 100 | 42 | 92 | 13 |
| Perplexity5 | 100 | 63 | 92 | 11 |
| Gemini1 | 100 | 38 | 91 | 7 |
| Gemini2 | 100 | 37 | 46 | 10 |
| Gemini3 | 100 | 37 | 82 | 8 |
| Gemini4 | 100 | 36 | 43 | 10 |
| Gemini5 | 100 | 11 | 43 | 15 |

detect all the files perfectly with 100% detection. However, the ability of the other three AI-detection tools was greatly affected due to this adversarial technique (ZeroGPT: for ChatGPT range 29%-81%, average 53%; for Perplexity range 28%-68%, average 53%; for Gemini range 11%-38%, average 31.8%; GPTZero: for ChatGPT range 23%-82%, average 50%; for Perplexity range 92%-100%, average 96.6%; for Gemini range 43%-91%, average 61%; Writer AI: for ChatGPT range 12%-15% average 13.6%; for Perplexity range 11%-16%, average 13.4%; for Gemini range 7%-15%, average 10%).

### Comparing the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (with third adversarial technique- 10%-20% editing by a human expert)

In the last analysis, all AI-generated essays with the third adversarial technique were checked by the four AI-detection tools (Table 5). Turnitin again performed perfectly by being able to detect all the AI-generated text with 100% accuracy despite the adversarial techniques. ZeroGPT was also able to detect contents generated by Perplexity with 100% accuracy. Both ZeroGPT and GPTZero were able to detect other files with high level of accuracy (ZeroGPT: for ChatGPT range 96%-100%, average 98%; for Gemini range 56%-99%, average 81.2%; GPTZero: for ChatGPT range 74%-100%, average 88.6%; for Perplexity range 83%-100%, average 96.2%; for Gemini range 91%-100%, average 97.8%). Writer AI once more performed below-par in accurately detecting AI-generated content with the third adversarial technique (for ChatGPT range 15%-29%, average 20.6%; for Perplexity range 23%-28%, average 25.8%; for Gemini range 18%-22%, average 20%).

The results revealed quite a few things. First of all, Turnitin performed the best among all four AI-detection tools that had been tested in this study. Even different adversarial techniques were not able to affect its ability to detect AI-generated text accurately. ZeroGPT and GPTZero were also able to detect AI-generated text with a high rate of accuracy; however, Writer AI was not that effective in detecting AI-generated text.

Among the LLMs, the essays generated through Perplexity were the most easily and accurately detected; on the other hand, AI-detection tools reported relatively low AI scores for the essays generated through Gemini. Of the three adversarial techniques, paraphrasing by Quillbot had the biggest impact on the AI-detection tools' ability to detect AI-generated text.

## Discussion and conclusion

AI-driven chatbots and LLMs have made a significant impact in a relatively short span of time (Jiang et al., 2022; Malik, 2024). Academic and research writing is one of the areas that has been influenced the most due to their ability to generate human-like text for different subjects, levels, and contexts (Carobene et al., 2024; Dwivedi et al., 2021; Malik et al., 2024). This has proven to be a double-edged sword which,

Table 5. Comparison of the accuracy of AI-detection tools in identifying AI-generated text from ChatGPT, Perplexity, and Gemini (with third adversarial technique- 10%-20% editing by a human expert).

| AI-generated essay (Edited by a human expert) | Turnitin (%) | ZeroGPT (%) | GPTZero (%) | Writer AI (%) |
|---|---|---|---|---|
| ChatGPT1 | 100 | 96 | 91 | 20 |
| ChatGPT2 | 100 | 97 | 91 | 19 |
| ChatGPT3 | 100 | 100 | 87 | 15 |
| ChatGPT4 | 100 | 100 | 100 | 29 |
| ChatGPT5 | 100 | 97 | 74 | 20 |
| Perplexity1 | 100 | 100 | 100 | 27 |
| Perplexity2 | 100 | 100 | 100 | 23 |
| Perplexity3 | 100 | 100 | 100 | 28 |
| Perplexity4 | 100 | 100 | 98 | 23 |
| Perplexity5 | 100 | 100 | 83 | 28 |
| Gemini1 | 100 | 88 | 100 | 22 |
| Gemini2 | 100 | 91 | 100 | 22 |
| Gemini3 | 100 | 72 | 100 | 19 |
| Gemini4 | 100 | 56 | 91 | 19 |
| Gemini5 | 100 | 99 | 98 | 18 |

on one side, facilitates saving time and improving the writing but, on the other, assists in AI-driven plagiarism (Alsabhan, 2023; Dwivedi et al., 2023; Malik et al., 2024; Motlagh et al., 2023; Xames & Shefa, 2023). AI-driven plagiarism is much harder to detect due to its advanced features and modeling abilities. Although many AI-detection tools have been developed to trace AI-generated text, their efficiency, accuracy, and consistency are questionable (Elkhatat et al., 2023; Weber-Wulff et al., 2023; Walters, 2023). However, as more advanced LLMs and AI-detection tools and versions are being developed, more studies need to be carried out using different tools and their versions.

This study further contributes to this area by checking the efficiency and accuracy of four AI-detection tools (Turnitin, ZeroGPT, GPTZero, and Writer AI) in detecting AI-generated text from the essays generated through three different LLMs (ChatGPT, Perplexity, and Gemini). This study not only checks the original files generated by those LLMs but also uses three different adversarial techniques (edited by Grammarly, paraphrased by Quillbot, and 10%-20% editing by a human expert). The study found that the four AI-detection tools showed inconsistent AI scores - from very high by Turnitin (almost perfect) to very low by Writer AI. It further endorses the findings of the previous studies about the inconsistent performance of AI-detection tools (Ladha et al., 2023; Perkins et al., 2024; Weber-Wulff et al., 2023).

Turnitin reported a 100% AI score even with the three adversarial techniques. In their study, Perkins et al. (2024) found Turnitin to be the second most accurate AI-detection tool. In another study, Walters (2023) also found Turnitin to be one of the more accurate and consistent AI-detection tools; however, in another study that was carried out to detect AI-generated text in files generated through ChatGPT4, Turnitin showed 0% AI score (Foster, 2023). This is not a surprise, as multiple studies have pointed toward the inconsistencies in AI-detection tools (Ladha et al., 2023; Weber-Wulff et al., 2023).

However, there were also inconsistencies in the AI-detection of the original files and with the first adversarial technique (e.g. in the essays generated through Gemini, ZeroGPT reported an AI score of 82% for one file while in two others generated through the same LLM, it was 100%); ZeroGPT and GPTZero were able to report relatively high average AI scores. However, their accuracy decreased, and the inconsistencies increased when the second and third adversarial techniques were applied (much more with the third adversarial technique, as Quillbot paraphrased the entire text compared to human editing, which was restricted to 10%-20% of the text). Once the AI-generated text was paraphrased through Quillbot, the AI scores decreased considerably. Multiple studies have also shown that paraphrasing or rephrasing is one of the most effective adversarial techniques to evade AI-detection (Mitchell et al., 2023; Perkins et al., 2024). However, another interesting observation was that after the second adversarial technique, the AI score ranges for the files generated through the same LLM also became much bigger (ZeroGPT: for ChatGPT range 29%-81%; for Perplexity range 28%-68%; for Gemini range 11%-38%; GPTZero: for ChatGPT range 23%-82%; for Perplexity range 92%-100%; for Gemini range 43%-91%). These differences may be due to different algorithms and patterns used by both Quillbot (through which text was paraphrased) and AI-detection tools (that were used to trace AI-generated text). Although all the AI-detection tools work on the same principle, their algorithms, approaches, and the dataset upon which they were trained might be different (Nwohiri et al., 2021).

One interesting finding was the relatively minimal effect of editing by a human expert. It may be attributed to the fact that the editor was asked to improve the text, rephrase it, or improve sentence structure from 10 to 20 per cent of the entire text. Studies have shown that adding errors in AI-generated text is a more effective way of evading AI-detection (Perkins et al., 2024). When edited by the human expert, those algorithms or patterns may not be altered or disturbed prominently (there were some changes/corrections here and there, which may not disturb the algorithm or the pattern). AI scores after adversarial techniques may be linked to how and how much those patterns and algorithms were disturbed (making them harder to detect). Amongst the LLMs, Perplexity provides the most accurate detection, even after adversarial techniques. Again, it may be attributed to its unique algorithm, patterns, and the dataset that it was trained upon.

Despite all the inconsistencies, it is important to note that at least three AI-detection tools had reasonably high average AI scores. Even in the text paraphrased by Quillbot, those three tools were able to report an AI score of 50% or above (with the exception of ZeroGPT for Gemini, which reported an average AI score of 31.8%;). Writer AI, although with a smaller range, reported low AI scores for AI-generated texts, proving to be the most inefficient and inaccurate among the four AI-detection tools tested in this study. However, what was really concerning was the fact that even when the files were generated by the same LLM (with or without adversarial technique) and checked through the same AI-detection tool; their AI scores showed big ranges. It raises further questions about the consistencies of AI-detection tools.

This study shows that despite certain inconsistencies, Turnitin, and to a lesser extent, GPTZero and ZeroGPT can be used for AI-detection as they can indicate AI-generated text by reporting relatively high AI scores. However, due to relatively high inconsistencies (as indicated by high ranges) even when the same LLM, same AI-detection tool, and the same adversarial technique are used; such indications should not be taken as a final verdict, and further checks and investigations should be carried out before labelling a text as AI-generated or otherwise.

## Further research

As LLMs and AI-detection tools are evolving at an unprecedented pace, it is important to continue conducting studies with different LLMs and AI-detection tools and their latest versions to check their accuracy and reliability. This way, we can find more reliable and accurate ones that can be used for AI detection. It is also important to employ a wider variety of adversarial techniques such as deliberating incorporating errors, translating and retranslating, and using multiple paraphrasing software to gauge their effects on the performance of different AI-detection tools.

During data analysis, we also noticed the differences in the quality of text generated by three LLMs; however, as it falls out of the scope of the study, we did not focus on that. Further studies may be conducted to see the quality of text generated through different LLMs.

## References

Akram, A. (2023). *An empirical study of AI-generated text detection tools*. arXiv preprint arXiv:2310.01423.

Alhijawi, B., Jarrar, R., AbuAlRub, A., & Bader, A. (2024). *Deep learning detection method for large language models-generated scientific content*. arXiv preprint arXiv:2403.00828.

Alsabhan, W. (2023). Student cheating detection in higher education by implementing machine learning and LSTM techniques. *Sensors, 23*(8), 4149. https://doi.org/10.3390/s23084149

Amjad, A. I., Aslam, S., & Tabassum, U. (2024). Tech-infused classrooms: A comprehensive study on the interplay of mobile learning, ChatGPT and social media in academic attainment. *European Journal of Education, 59*(2). e12625. https://doi.org/10.1111/ejed.12625

Arabi, H., & Akbari, M. (2022). Improving plagiarism detection in text documents using hybrid weighted similarity. *Expert Systems with Applications, 207*, 118034. https://doi.org/10.1016/j.eswa.2022.118034

Barbetta, P. M. (2023). Remedial and compensatory writing technologies for middle school students with learning disabilities and their classmates in inclusive classrooms. In *Preventing school failure: Alternative education for children and youth* (pp. 1-12). https://doi.org/10.1080/1045988X.2023.2259837

Bates, T., Cobo, C., Mariño, O., & Wheeler, S. (2020). Can artificial intelligence transform higher education? *International Journal of Educational Technology in Higher Education, 17,* 1-12. https://doi.org/10.1186/s41239-020-00218-x

Bawack, R. E., Fosso Wamba, S., & Carillo, K. D. A. (2021). A framework for understanding artificial intelligence research: Insights from practice. *Journal of Enterprise Information Management, 34*(2), 645-678. http://dx.doi.org/10.1108/JEIM-07-2020-0284

Carobene, A., Padoan, A., Cabitza, F., Banfi, G., & Plebani, M. (2024). Rising adoption of artificial intelligence in scientific publishing: Evaluating the role, risks, and ethical implications in paper drafting and review process. *Clinical Chemistry and Laboratory Medicine (CCLM), 62*(5), 835-843. https://doi.org/10.1515/cclm-2023-1136

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning & Teaching, 6*(2), 94-104. https://doi.org/10.37074/jalt.2023.6.2.12

Chaka, C. (2024a). 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. https://doi.org/10.37074/jalt.2024.7.1.14

Chaka, C. (2024b). Accuracy pecking order–How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. *Journal of Applied Learning and Teaching, 7*(1), 1-13. https://doi.org/10.37074/jalt.2024.7.1.33

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To be or not to be… human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. *Journal of Management Information Systems, 39*(4), 969-1005. https://doi.org/10.1080/07421222.2022.2127441

Chawla, D. S. (2024, October 15). *Springer Nature journal has retracted over 200 papers since September*. Retraction Watch. https://retractionwatch.com/2024/10/15/springer-nature-journal-has-retracted-over-200-papers-since-september/

Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. *The International Journal of Information and Learning Technology, 40*(3), 259-268. http://dx.doi.org/10.1108/IJILT-03-2023-0043

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71,* 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

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), 17. https://doi.org/10.1007/s40979-023-00140-5

Fairooz, F., Jayasundara, A., & Udara, N. (2023). Using artificial intelligence tools in language learning in tertiary education in Sri Lanka: A challenge to academic integrity? In *Annual International Conference On Business Innovation (ICOBI)* (pp. 281-294).

Foster, A. (2023). *Can GPT-4 fool TurnItIn? Testing the limits of AI detection with prompt engineering*. https://digital.kenyon.edu/cgi/viewcontent.cgi?article=1041&context=dh_iphs_ai

George, B., & Wooden, O. (2023). Managing the strategic transformation of higher education through artificial intelligence. *Administrative Sciences, 13*(9), 196. https://doi.org/10.3390/admsci13090196

Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., ... & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots. *Internet of Things and Cyber-Physical Systems, 4,* 19-23. https://doi.org/10.1016/j.iotcps.2023.06.002

Guo, D., Chen, H., Wu, R., & Wang, Y. (2023). AIGC challenges and opportunities related to public safety: A case study of ChatGPT. *Journal of Safety Science and Resilience, 4*(4), 329-339. https://doi.org/10.1016/j.jnlssr.2023.08.001

Gustilo, L., Ong, E., & Lapinid, M. R. (2024). Algorithmically-driven writing and academic integrity: Exploring educators' practices, perceptions, and policies in AI era. *International Journal for Educational Integrity, 20*(1), 3. https://doi.org/10.1007/s40979-024-00153-8

Haider, J., Söderström, K. R., Ekström, B., & Rödl, M. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review, 5*(5), 1-16. https://doi.org/10.37016/mr-2020-156

Hasanein, A., Sobaih, A., & Elshaer, I. (2024). Examining Google Gemini's acceptance and usage in higher education. *Journal of Applied Learning and Teaching, 7*(2), 1-9. https://doi.org/10.37074/jalt.2024.7.2.5

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society, 26*(1), 112-131. https://www.jstor.org/stable/48707971

Hussain, A., & Qazi, K. A. (2023). Textual Alchemy: AI, authorship and the shifting paradigms of interpretation. *Rupkatha Journal on Interdisciplinary Studies in Humanities, 15*(4). https://doi.org/10.21659/rupkatha.v15n4.08

Hutson, M. (2022). Could AI help you to write your next paper? *Nature, 611*(7934), 192-193. https://doi.org/10.1038/d41586-022-03479-w

Iorliam, A., & Ingio, J. A. (2024). A comparative analysis of generative artificial intelligence tools for natural language processing. *Journal of Computing Theories and Applications, 2*(1), 91-105. http://dx.doi.org/10.62411/jcta.9447

Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3*(2), 100115. https://doi.org/10.1016/j.tbench.2023.100115

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence?. *Discover Artificial Intelligence, 2*(1), 4. https://doi.org/10.1007/s44163-022-00022-8

Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointestinal Endoscopy, 92*(4), 807-812. https://doi.org/10.1016/j.gie.2020.06.040

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., ... & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion, 99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kooli, C., & Yusuf, N. (2024). Transforming educational assessment: Insights into the use of ChatGPT and large language models in grading. *International Journal of Human-Computer Interaction*, 1-12. https://doi.org/10.1080/10447318.2024.2338330

Kumar, R. (2023). Faculty members' use of artificial intelligence to grade student papers: A case of implications. *International Journal for Educational Integrity, 19*(1), 9. https://doi.org/10.1007/s40979-023-00130-7

Ladha, N., Yadav, K., & Rathore, P. (2023). AI-generated content detectors: Boon or bane for scientific writing. *Indian Journal of Science and Technology, 16*(39), 3435-3439. https://doi.org/10.17485/IJST/v16i39.1632

Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International Journal of Environmental Research and Public Health, 18*(1), 271.https://doi.org/10.3390/ijerph18010271

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., ... & Zou, J. Y. (2024). *Mapping the increasing use of LLMs in scientific papers*. arXiv preprint arXiv:2404.01268.

Liu, J., & Li, S. (2024). Toward artificial intelligence-human paired programming: A review of the educational applications and research on artificial intelligence code-generation tools. *Journal of Educational Computing Research, 62*(11). http://dx.doi.org/10.1177/07356331241240460

Mainaly, S. (2023). Bing, bard, and brainstorming: A triadic tenor of AI pedagogy. *Journal of Global Literacies, Technologies, and Emerging Pedagogies, 9*(2), 1595-1613. https://jogltep.com/wp-content/uploads/2023/12/14.0-Shiva-Mainaly-final.pdf

Malik, M. A. (2024). Challenges and opportunities about ChatGPT in higher education: A qualitative study about university teachers in Pakistan. *Voyage Journal of Educational Studies, 4*(2), 315-324. https://doi.org/10.58622/vjes.v4i2.166

Malik, M. A., Amjad, A. I., Aslam, S., & Fakhrou, A. (2024). Global insights: ChatGPT's influence on academic and research writing, creativity, and plagiarism policies. *Frontiers in Research Metrics and Analytics, 9*, 1-12, 1486832. https://doi.org/10.3389/frma.2024.1486832

Malik, M. A., Mahroof, A., & Ashraf, M. A. (2021). Online university students' perceptions on the awareness of, reasons for, and solutions to plagiarism: The development of the AS&P model to combat plagiarism. *Applied Sciences, 11*(24), 1-14, 12055. https://doi.org/10.3390/app112412055

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). *DetectGPT: Zero-shot machine-generated text detection using probability curvature* (arXiv:2301.11305). arXiv. http://arxiv.org/abs/2301.11305

Motlagh, N. Y., Khajavi, M., Sharifi, A., & Ahmadi, M. (2023). *The impact of artificial intelligence on the evolution of digital education: A comparative study of OpenAI text generation tools including ChatGPT, Bing Chat, Bard, and Ernie*. arXiv preprint arXiv:2309.02029.

Nawaz, N., & Gomes, A. M. (2019). Artificial intelligence chatbots are new recruiters. *International Journal of Advanced Computer Science and Applications, 10*(9), 1-5. https://dx.doi.org/10.2139/ssrn.3521915

Nwohiri, A., Opemipo, J. O. D. A., & Ajayi, O. (2021). AI-powered plagiarism detection: Leveraging forensic linguistics and natural language processing. *Fudma Journal of Sciences, 5*(3), 207-218. http://dx.doi.org/10.33003/fjs-2021-0503-700

Odri, G. A., & Yoon, D. J. Y. (2023). Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. *Orthopaedics & Traumatology: Surgery & Research, 109*(8), 103706. https://doi.org/10.1016/j.otsr.2023.103706

Onal, S., & Kulavuz-Onal, D. (2024). A cross-disciplinary examination of the instructional uses of ChatGPT in higher education. *Journal of Educational Technology Systems, 52*(3), 301-324.

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2). 1-24. https://doi.org/10.53761/1.20.02.07

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education, 21*(1), 1-25. https://doi.org/10.1186/s41239-024-00487-w

Raisch, S., & Fomina, K. (2023). Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review,* 1–24. https://doi.org/10.5465/amr.2021.0421

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., ... & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching, 6*(1), 41-56. https://doi.org/10.37074/jalt.2023.6.1.29

Reza, M., Laundry, N., Musabirov, I., Dushniku, P., Yu, Z. Y., Mittal, K., ... & Williams, J. J. (2023). *ABScribe: Rapid exploration of multiple writing variations in human-AI co-writing tasks using large language models*. arXiv preprint arXiv:2310.00117.

Rusmiyanto, R., Huriati, N., Fitriani, N., Tyas, N. K., Rofi'i, A., & Sari, M. N. (2023). The role of artificial intelligence (AI) in developing English language learner's communication skills. *Journal on Education, 6*(1), 750-757. http://dx.doi.org/10.31004/joe.v6i1.2990

Shidiq, M. (2023). The use of artificial intelligence-based Chat-GPT and its challenges for the world of education; From the viewpoint of the development of creative writing skills. In *Proceeding of International Conference on Education, Society, and Humanity, 1*(1), 360-364. https://ejournal.unuja.ac.id/index.php/icesh

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14,* 1-14. https://doi.org/10.3389/fpsyg.2023.1260843

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching, 6*(1), 1-10. https://doi.org/10.37074/jalt.2023.6.1.17

Taecharungroj, V. (2023). "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on Twitter. *Big Data and Cognitive Computing, 7*(1), 35. https://doi.org/10.3390/bdcc7010035

Tilwani, D., Saxena, Y., Mohammadi, A., Raff, E., Sheth, A., Parthasarathy, S., & Gaur, M. (2024). *REASONS: A benchmark for retrieval and automated citations of scientific* *sentences using public and proprietary LLMs*. arXiv preprint arXiv:2405.02228.

Walters, W. H. (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

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), 1-39. https://doi.org/10.1007/s40979-023-00146-z

Xames, M. D., & Shefa, J. (2023). ChatGPT for research and publication: Opportunities and challenges. *Journal of Applied Learning and Teaching, 6*(1), 390-395. https://doi.org/10.37074/jalt.2023.6.1.20