International Journal for Educational Integrity

ORIGINAL ARTICLE Open Access

Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities

Ahmed M. Elkhatat[1*]

*Correspondence:
ahmed.elkhatat@qu.edu.qa

[1] Department of Chemical Engineering, Qatar University, PO Box 2713, Doha, Qatar

Abstract

Academic plagiarism is a pressing concern in educational institutions. With the emergence of artificial intelligence (AI) chatbots, like ChatGPT, potential risks related to cheating and plagiarism have increased. This study aims to investigate the authenticity capabilities of ChatGPT models 3.5 and 4 in generating novel, coherent, and accurate responses that evade detection by text-matching software. The repeatability and reproducibility of both models were analyzed, showing that the generation of responses remains consistent. However, a two-sample t-test revealed insufficient evidence to support a statistically significant difference between the text-matching percentages of both models. Several strategies are proposed to address the challenges posed by AI integration in academic contexts; one probable solution is to promote self-transcendent ideals by implementing honor codes. It is also necessary to consider the restricted knowledge base of AI language models like GPT and address any inaccuracies in generated references. Additionally, designing assignments that extract data from imaged sources and integrating oral discussions into the evaluation process can mitigate the challenges posed by AI integration. However, educators should carefully consider the practical constraints and explore alternative assessment methods to prevent academic misconduct while reaping the benefits of these strategies.

Keywords: Artificial intelligence, Plagiarisim, Academic integrity, ChatGPT

Introduction

Academic plagiarism has gained prominence in the academic sphere, as it has been detected in various student assignments, including reports, homework, projects, and more. Academic plagiarism can be characterized as using ideas, content, or structures without adequately attributing the source (Fishman 2009). Conventional Plagiarism tactics employed by students vary, with the most extreme form involving complete duplication of the source material. Alternative methods include partial paraphrasing by altering grammatical structures or replacing words with synonyms, utilizing online paraphrasing services to rephrase text (Elkhatat et al. 2021; Meuschke & Gipp 2013; Sakamoto & Tsuda 2019).

Recently, Artificial intelligence (AI) powered ChatGPT has emerged as a tool that assists students in generating customized content based on prompts, utilizing

natural language processing (NLP) techniques (Radford et al. 2018), posing potential risks related to cheating and plagiarism, with severe academic and legal consequences (Foltýnek et al. 2019). Despite the utility of ChatGPT in assisting students with essay writing and other academic tasks, concerns have been raised about the originality and appropriateness of the content generated by the chatbot for academic use (King & chatGpt 2023). Furthermore, ChatGPT has faced criticism for producing incoherent or inaccurate content (Gao et al. 2022; Qadir 2022), offering superficial information (Frye 2022), and possessing a limited knowledge base due to its disconnection from the internet and reliance on data available up to September-2021 (Williams 2022). Nevertheless, empirical evidence to substantiate these assertions remains scarce.

Academic plagiarism represents a breach of ethical conduct and is among the most grievous instances of research impropriety, as it imperils obtaining and evaluating competencies. Consequently, implementing measures to mitigate plagiarism is crucial for upholding academic integrity and precluding the perpetuation of such dishonest practices in students' subsequent academic and professional pursuits. (Alsallal et al. 2013; Elkhatat 2022; Foltýnek et al. 2020). Text-Matching Software Products (TMSPs) are potent tools academic institutions employ to identify plagiarism owing to their advanced text-matching algorithms and comprehensive databases encompassing web pages, journal articles, periodicals, and other publications. Moreover, some TMSPs databases index previously submitted student papers, enhancing their effectiveness in plagiarism detection (Elkhatat et al. 2021).

In light of concerns about ChatGPT responses, the present study aims to investigate ChatGPT's ability to generate novel, coherent, and accurate responses that evade detection by text-matching software, exploring the potential implications of using such AI-generated content in academic settings.

Background and literature review

AI has recently opened up numerous possibilities in the academic domain, transforming the educational landscape through various applications, such as NLP and autonomous systems (Norvig 2021). AI has been employed in education to create personalized student learning experiences, leveraging NLP and machine learning algorithms (Chen et al., 2012). The advent of AI-based tutoring systems has contributed to increasingly interactive and engaging student learning environments (Sapci & Sapci 2020). Furthermore, AI-based platforms have maintained academic integrity by detecting plagiarism and providing personalized feedback (Hinojo-Lucena et al. 2019). However, AI also poses potential risks related to cheating and plagiarism, with severe academic and legal consequences (Foltýnek et al. 2019). AI in higher education has led to concerns about academic integrity, as students may use AI tools to cheat and plagiarize, allowing students to tailor the content they create, potentially misusing AI for academic dishonesty (Cotton et al. 2023; Francke & Bennett 2019).

Recently, AI-powered ChatGPT has emerged as a tool that assists students in generating customized content based on prompts, utilizing NLP techniques (Radford et al. 2018). The original GPT model demonstrated the potential of unsupervised pre-training followed by supervised fine-tuning for a wide range of NLP tasks. Subsequently, OpenAI released ChatGPT (model 2), further improving the model's capabilities by scaling

up the architecture and employing a more extensive pre-training dataset (Radford et al. 2019). The subsequent release of ChatGPT (models 3 and 3.5) marked another milestone in the development of ChatGPT as it showcased remarkable performance in generating human-like text, achieving state-of-the-art results on multiple NLP benchmarks. The model's ability to generate coherent and contextually relevant text in response to prompts made it an ideal foundation for building ChatGPT, an AI-powered chatbot designed to assist users in generating text and engaging in natural language conversations (Brown et al. 2020; OpenAI 2022). The recently released ChatGPT (model 4) by OpenAI on March 14, 2023, marks a substantial NLP technology milestone. With advanced safety features and superior response quality, it outperforms its predecessors in addressing complex challenges. ChatGPT (model 4)'s extensive general knowledge and problem-solving aptitude empower it to handle demanding tasks with increased accuracy. Additionally, its creative and collaborative functionalities facilitate the generation, editing, and iteration of various creative and technical writing endeavors, such as composing songs, crafting screenplays, and adapting personalized writing styles. Notably, ChatGPT (model 4) is available through the Plus plan subscription, which costs $20 per month. Nonetheless, it is essential to recognize that ChatGPT (model 4)'s knowledge is limited to the cutoff date in September 2021 (OpenAI 2023).

The study of academic plagiarism and online cheating is a constantly evolving research field that has garnered significant attention in the academic community. Numerous published studies have developed algorithms and codes that effectively search for matched texts (Hajrizi et al. 2019; Pizarro V & Velásquez, 2017; Roostaee et al. 2020; Sakamoto & Tsuda 2019; Sánchez-Vega et al. 2013). Additionally, other studies have presented pedagogical strategies to mitigate plagiarism among students (Elkhatat et al. 2021; Landau et al. 2016; Yang et al. 2019).

Furthermore, the literature has potential risks related to cheating and plagiarism using AI-powered chatbots. These research efforts demonstrate the importance of addressing plagiarism using AI-powered chatbots in academia and the need for ongoing research and development. A recent article (Anders 2023) explored the ethical implications and potential misuse of AI technologies like ChatGPT in the educational context. The author discusses the necessity of a future-proofing curriculum to address the challenges posed by AI-assisted assignments and highlights vital concerns related to this emerging technology. A recent editorial published in Nurse Education in Practice (Siegerink et al. 2023) discussed the role of large language models (LLMs), specifically ChatGPT, in nursing education and addressed the controversy surrounding its listing as an author. They argue that ChatGPT cannot be considered an author due to the lack of accountability and the inability to meet the authorship criteria outlined by the International Committee of Medical Journal Editors (ICMJE) and the Committee on Publication Ethics (COPE). The authors suggest that LLMs like ChatGPT should be transparently mentioned in the writing process, especially in academic texts where arguments are central to the work. The norms regarding using such models in science and nursing education are still emerging, and transparency and a critical attitude are crucial moving forward. Another article (Alser & Waisberg 2023) confirms what was previously mentioned. The authors express concerns regarding the growing use of ChatGPT in academia and medicine, specifically addressing the issues of authorship and plagiarism. They argue that

ChatGPT does not meet the ICMJE guidelines for authorship, as it lacks accountability and approval of published work. The authors also conducted plagiarism checks on parts of writing contributed by ChatGPT, revealing instances of direct, paraphrasing, and source-based plagiarism. They discuss the potential biases in ChatGPT's outputs, as the model does not differentiate between sources based on the level of evidence and can be manipulated through user voting. The authors recommend against using ChatGPT in academia, and if its use is unavoidable, they suggest acknowledging the bot without granting authorship and paying attention to potential plagiarism and biases. Furthermore, in a study discussing the impact of AI tools like ChatGPT on scientific writing (Rozencwajg & Kantor 2023), the authors emphasize the benefits of AI-generated content, such as speed and efficiency, but also underscore the importance of maintaining accuracy and rigor. The authors used ChatGPT to create an editorial addressing AI's impact on scientific writing and the role of reviewers and editors. The model produces a well-organized, scientifically-sound text with references, showcasing its potential as a valuable tool for scientific writers. However, the authors caution against using AI-generated content without proper monitoring, as it may create biased or inaccurate content. In (Eke 2023), the author argued that the use of AI-powered text generators such as ChatGPT could potentially undermine academic integrity but also has the potential to revolutionize academia. The author suggests that OpenAI and other LLM creators should be willing to work with academia to use AI-powered text generators responsibly and that a multi-stakeholder endeavor is needed to co-create solutions to maintain academic integrity. Recently (Sadasivan et al. 2023) published a study investigating the reliability of current detection techniques in identifying AI-generated text. The researchers used a set of 10,000 text samples, half of which were generated by AI models, to evaluate the effectiveness of 10 different detection methods. The methods included traditional feature-based approaches, deep learning models, and combining both. The output of the research showed that while some of the detection techniques were effective in identifying AI-generated text, none of them were completely reliable. The researchers found that AI models have become sophisticated enough to generate text that is difficult to distinguish from human-generated text. They also concluded that more research is needed to develop better detection techniques that can keep up with the advancements in AI technology.

Despite extensive research on the concerns and risks of ChatGPT, no studies have yet examined the authenticity of ChatGPT Responses in terms of repeatability and reproducibility and the capability of ChatGPT (models 3.5 and 4) to generate multiple responses without being detected by text-matching software. Thus, the current study aims to investigate the authentic capabilities of ChatGPT (models 3.5 and 4) and to propose strategies for mitigating potential risks associated with using ChatGPT while ensuring academic authenticity.

Methodology

A prompt to write 100 words on the "Application of cooling towers in the engineering process." was provided to ChatGPT's chatbot (models 3.5 and 4). The chatbot's response was recorded and then regenerated twice more within the same chatbot to assess its repeatability in generating new and original responses. A new chatbot was then created,

and the same prompt was used to repeat the experiment and assess the reproducibility of the chatbot's ability to generate new and original responses. Each response was evaluated and coded based on the repeatability and reproducibility process. Table 1 displays 45 responses from the ChatGPT chatbot; 30 responses from ten chats using chatGPT model 3.5, each generated three times (first response and two regenerated responses), and 15 responses from five chats using ChatGPT model 4, each generated three times.

The 45 responses were uploaded one by one on SafeAssign of the Blackboard Learn (Blackboard 2023) platform, which allows students to submit assignments and check the text match percentage. Each response was checked for information quality and the text match percentage and source. Statistical analysis and capabilities tests were conducted using Minitab (Minitab 2023a).

Results and discussion

An examination of text match percentages and similarity origins for ChatGPT models 3.5 and 4: a comparative analysis of response authenticity

The results of the text match percentage of each of the forty-five generated ChatGPT models 3.5 and 4 responses are presented in Table 2. In addition to the overall similarity percentage, the table incorporates the origin of the similarity, such as previously generated ChatGPT responses, the students' Global database, which consists of other students' submissions via the Blackboard platform, and sources from the internet. Furthermore, Tables 3 and 4 illustrate the text similarity metrics for ChatGPT models 3.5 and 4, respectively. The findings indicate that, in the case of ChatGPT model 3.5, most of the similarities stem from previously generated ChatGPT responses, with a peak of 55% in certain instances. This is followed by the Global database, reaching up to 45%, and online resources, with a minimum similarity percentage of 30%. Conversely, responses produced by ChatGPT model 4 solely originated from prior ChatGPT responses, with one instance exhibiting a 40% similarity and an overall average of 12% similarity across

Table 1 Codes of the 45 responses from 15 chats, each generated three times using ChatGPT models 3.5 and 4

	First Response	Regeneration 1	Regeneration 2
Chat Process 1 (model GPT 3.5)	ChatGPT 1–1	ChatGPT 1–2	ChatGPT 1–3
Chat Process 2 (model GPT 3.5)	ChatGPT 2–1	ChatGPT 2–2	ChatGPT 2–3
Chat Process 3 (model GPT 3.5)	ChatGPT 3–1	ChatGPT 3–2	ChatGPT 3–3
Chat Process 4 (model GPT 3.5)	ChatGPT 4–1	ChatGPT 4–2	ChatGPT 4–3
Chat Process 5 (model GPT 3.5)	ChatGPT 5–1	ChatGPT 5–2	ChatGPT 5–3
Chat Process 6 (model GPT 3.5)	ChatGPT 6–1	ChatGPT 6–2	ChatGPT 6–3
Chat Process 7 (model GPT 3.5)	ChatGPT 7–1	ChatGPT 7–2	ChatGPT 7–3
Chat Process 8 (model GPT 3.5)	ChatGPT 8–1	ChatGPT 8–2	ChatGPT 8–3
Chat Process 9 (model GPT 3.5)	ChatGPT 9–1	ChatGPT 9–2	ChatGPT 9–3
Chat Process 10 (model GPT 3.5)	ChatGPT 10–1	ChatGPT 10–2	ChatGPT 10–3
Chat Process 11 (model GPT 4)	ChatGPT 11–1	ChatGPT 11–2	ChatGPT 11–3
Chat Process 12 (model GPT 4)	ChatGPT 12–1	ChatGPT 12–2	ChatGPT 12–3
Chat Process 13 (model GPT 4)	ChatGPT 13–1	ChatGPT 13–2	ChatGPT 13–3
Chat Process 14 (model GPT 4)	ChatGPT 14–1	ChatGPT 14–2	ChatGPT 14–3
Chat Process 15 (model GPT 4)	ChatGPT 15–1	ChatGPT 15–2	ChatGPT 15–3

Table 2 The results of the text match percentage of each of the forty-five generated ChatGPT responses

Response	Overall Matching	Origin of Matching		
		Previously Generated ChatGPT Responses	Students Global Database	Internet
ChatGPT 1–1 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 1–2 (model GPT 3.5)	17%	17%	0%	0%
ChatGPT 1–3 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 2–1 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 2–2 (model GPT 3.5)	34%	34%	0%	0%
ChatGPT 2–3 (model GPT 3.5)	16%	0%	0%	16%
ChatGPT 3–1 (model GPT 3.5)	30%	0%	0%	30%
ChatGPT 3–2 (model GPT 3.5)	42%	0%	34%	8%
ChatGPT 3–3 (model GPT 3.5)	21%	21%	0%	0%
ChatGPT 4–1 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 4–2 (model GPT 3.5)	10%	0%	0%	10%
ChatGPT 4–3 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 5–1 (model GPT 3.5)	16%	16%	0%	0%
ChatGPT 5–2 (model GPT 3.5)	16%	16%	0%	0%
ChatGPT 5–3 (model GPT 3.5)	68%	23%	45%	0%
ChatGPT 6–1 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 6–2 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 6–3 (model GPT 3.5)	18%	18%	0%	0%
ChatGPT 7–1 (model GPT 3.5)	17%	17%	0%	0%
ChatGPT 7–2 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 7–3 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 8–1 (model GPT 3.5)	0%	0%	0%	0%
ChatGPT 8–2 (model GPT 3.5)	18%	18%	0%	0%
ChatGPT 8–3 (model GPT 3.5)	55%	28%	16%	12%
ChatGPT 9–1 (model GPT 3.5)	19%	19%	0%	0%
ChatGPT 9–2 (model GPT 3.5)	55%	38%	13%	15%
ChatGPT 9–3 (model GPT 3.5)	55%	55%	0%	0%
ChatGPT 10–1 (model GPT 3.5)	44%	44%	0%	0%
ChatGPT 10–2 (model GPT 3.5)	14%	14%	0%	0%
ChatGPT 10–3 (model GPT 3.5)	16%	0%	0%	16%
ChatGPT 11–1 (model GPT 4)	0%	0%	0%	0%
ChatGPT 11–2 (model GPT 4)	0%	0%	0%	0%
ChatGPT 11–3 (model GPT 4)	0%	0%	0%	0%
ChatGPT 12–1 (model GPT 4)	0%	0%	0%	0%
ChatGPT 12–2 (model GPT 4)	0%	0%	0%	0%
ChatGPT 12–3 (model GPT 4)	40%	40% [a]	0%	0%
ChatGPT 13–1 (model GPT 4)	17%	17% [a]	0%	0%
ChatGPT 13–2 (model GPT 4)	25%	25% [a]	0%	0%
ChatGPT 13–3 (model GPT 4)	0%	0%	0%	0%
ChatGPT 14–1 (model GPT 4)	28%	28% [a]	0%	0%
ChatGPT 14–2 (model GPT 4)	23%	23% [a]	0%	0%
ChatGPT 14–3 (model GPT 4)	0%	0%	0%	0%
ChatGPT 15–1 (model GPT 4)	0%	0%	0%	0%
ChatGPT 15–2 (model GPT 4)	23%	23% [a]	0%	0%
ChatGPT 15–3 (model GPT 4)	26%	26% [a]	0%	0%

[a] *Responses from ChatGPT (model 4) were regenerated from previous responses of ChatGPT (model 4)*

Table 3 Statistics of the Text matching of ChatGPT (model GPT 3.5) responses

Variable	Mean	Minimum	Maximum
Overall matching	19%	0%	68%
ChatGPT previously generated responses	13%	0%	55%
Students global database	4%	0%	45%
Internet	4%	0%	30%

Table 4 Statistics of the Text matching of ChatGPT (model GPT4) responses

Variable	Mean	Minimum	Maximum
Overall matching	12%	0%	40%
ChatGPT previously generated responses	12%	0%	40%
Students global database	0%	0%	0%
Internet	0%	0%	0%

the responses. Moreover, responses from ChatGPT (model 4) were regenerated from previous responses of the same model, and none of the responses were regenerated from ChatGPT (model 3.5), indicating the implementation of distinct algorithms and techniques in ChatGPT (model 4).

Evaluating ChatGPT Models' performance in adhering to academic integrity standards: a capability assessment Using Ppk and Ppm Indices in an educational context

The acceptable range of plagiarism percentages in educational contexts is subject to variability across institutions, disciplines, and assignment types. While certain academic institutions impose stringent policies, permitting no more than 10% similarity in assignments, others may accept a similarity below 15%, especially in the context of journal submissions. However, a similarity exceeding 25% is generally regarded as a high percentage of plagiarism, which raises concerns about academic integrity and may result in severe penalties. (Jones & Sheridan 2014; Scanlon 2003). In light of these considerations, the present study sought to evaluate the capacity of ChatGPT to generate responses with a text-matching percentage of less than 10% (as a strict AI capability) and 25% (as a maximum acceptance limit capability), using the MatLab platform.

Capability indices, such as Ppk and Ppm, are statistical measures that provide insight into a process's performance by assessing its ability to meet specifications. Ppk (Process Performance Index) is a measure that indicates how well a process is performing relative to its specification limits. It takes into account both the process mean and variability. A higher Ppk value suggests that the process is more capable, producing fewer defective products and staying within the specified limits. A Ppk value greater than 1.33 is generally considered satisfactory, indicating that the process is capable and has a minimal variation with the specification limits (Bothe 1998). Ppm (Parts per Million) is another metric representing the number of defective parts in a batch of one million units. Lower Ppm values indicate a higher process capability, as fewer defective responses are generated. Ppm can be linked to the process capability indices (Cp and Cpk), which estimate the number of defects a process might generate. Capability tests calculate both expected

and observed PPM in capability testing. The expected PPM in capability testing is a long-term estimate using the standard deviation, while the observed PPM is a direct measurement of the current process performance, and it is the actual number of defective units in a sample divided by the total sample size(Minitab 2023b; Montgomery 2020).

In the strict AI capability test (10%) of ChatGpt (model 3.4), as shown in Fig. 1, the Ppk value of -0.14 is substantially below the acceptable threshold of 1.33. This finding indicates that the performance is unsatisfactory, exhibiting considerable variation and deviation from the target of generating responses with less than 10% text matching. The expected and observed Ppm < LSL values are 665,305.85 and 333,333.33, respectively. These figures represent the number of responses (in this case, responses with less than 10% text matching) per million generated, signifying that the expected overall capability of ChatGPT (model 3.4) to generate responses with less than 10% text matching is 66.5%; however, the observed capability stands at only 33.33%. The discrepancy between the observed and expected capabilities implies that ChatGPT (model 3.4) performance can be better evaluated when a larger volume of generated responses is considered. Figures 2, 3, 4, and 5 measure the capability of each source of the matching (ChatGPT previously generated responses, students' global database, and the internet). The summary of the capabilities of these sources is shown in Table 5.

For the maximum acceptance limit capability (25%), the values of capability indices PPK and PPM for the overall and each source of the matching (ChatGPT previously generated responses, students' global database, and the internet) are shown in Figs. 5, 6, 7, 8, respectively. The summary of the capabilities of these sources is shown in Table 6.

The authentic capability of ChatGPT (model 4) was assessed for 10% and 25% text matching, as displayed in Fig. 9 and Fig. 10. The Ppk values of -0.27 and -0.35 are significantly below the acceptable threshold of 1.33, indicating unsatisfactory performance characterized by substantial variation and deviation from the target. The expected and observed capabilities at 10% text matching stand at 53.3% and 78.9%, respectively, while the expected and observed capabilities at 25% text matching are 73.3% and 85.3%, respectively. These results might suggest an enhanced capability of ChatGPT model 4 compared to ChatGPT model 3.5 in generating authentic responses. However, a two-sample t-test hypothetical analysis is needed to decide on this enhanced performance, which will be discussed in the following section.

Comparing the authenticity of responses between ChatGPT (model 3.5) and (model 4): a Two-Sample T-Test Analysis

The two-sample t-test is a statistical method used to compare the means of two independent samples to determine whether they have a significant difference. This test is employed when the population variances are assumed equal and the samples are normally distributed (Field, 2013). In the present study, a two-sample t-test was conducted using Minitab (Minitab 2023a) to compare the text-matching percentages of ChatGPT (model 3.5) and (model 4). The null hypothesis (H_0) posits that there is no difference between the means of the two samples (i.e., the difference is equal to 0), while the alternative hypothesis (H_1) asserts that the difference is less than 0. The test yielded a p-value of 0.085. Although the p-value suggests a potential difference between the two samples, it is greater than the conventional significance level (α) of 0.05. Consequently, at a 95%

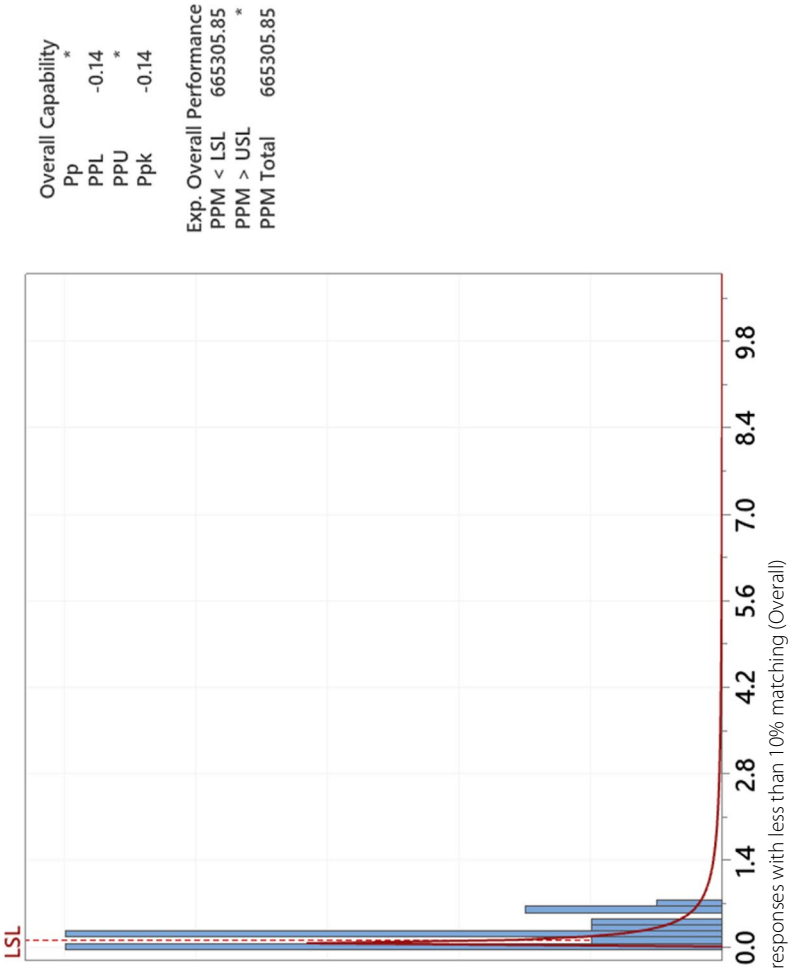

Fig. 1 The capability of ChatGPT (model 3.5) to generate responses with less than 10% matching (Overall)

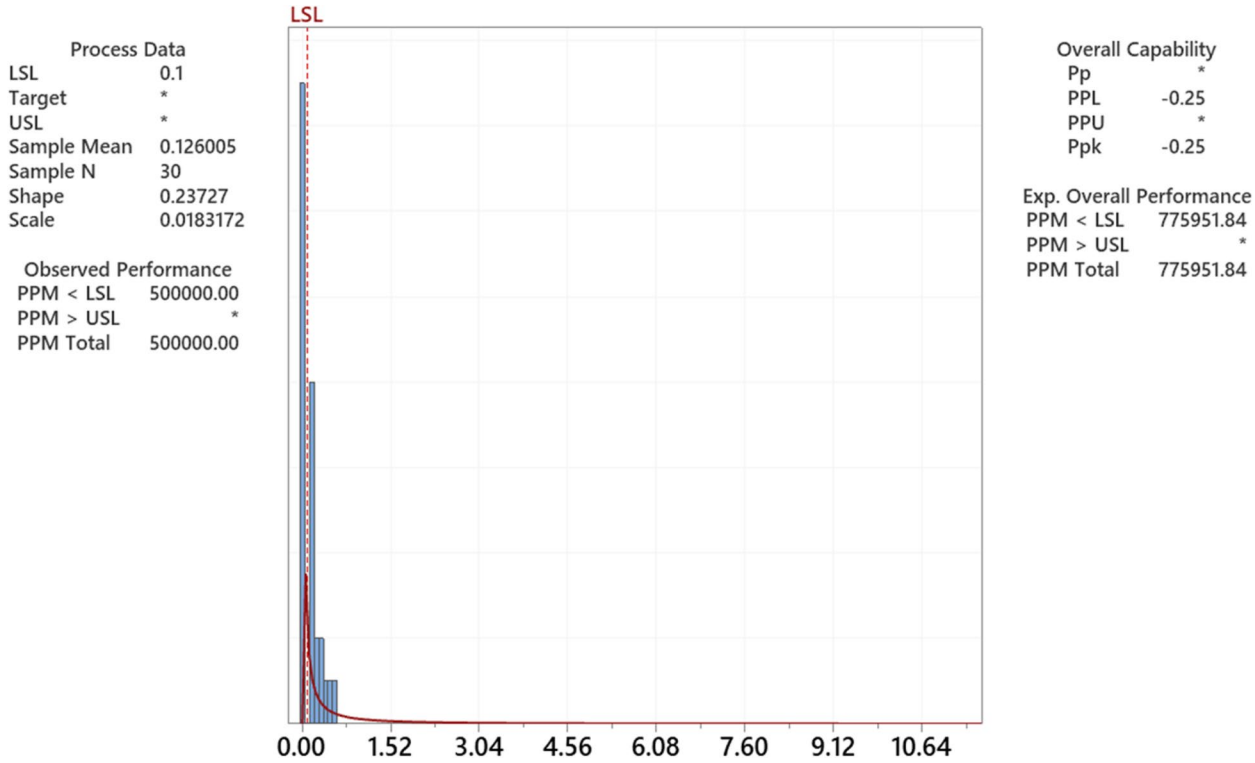

Fig. 2 The capability of ChatGPT (model 3.5) to generate responses with less than 10% matching (ChatGPT previously generated responses)

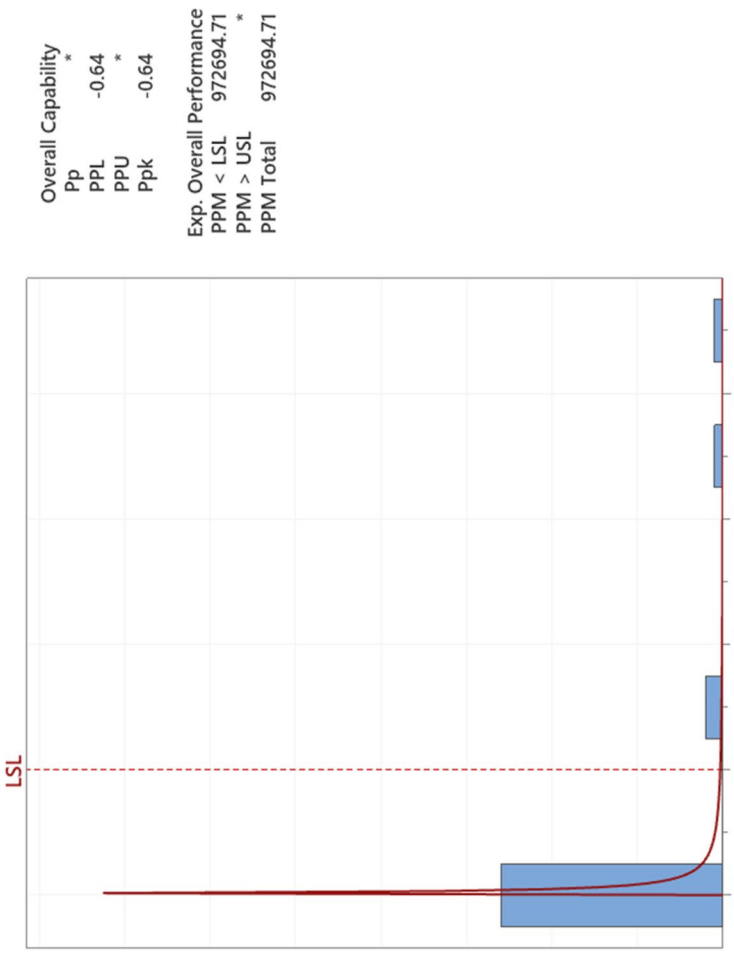

Fig. 3 The capability of ChatGPT (model 3.5) to generate responses with less than 10% matching (Students Global Database)

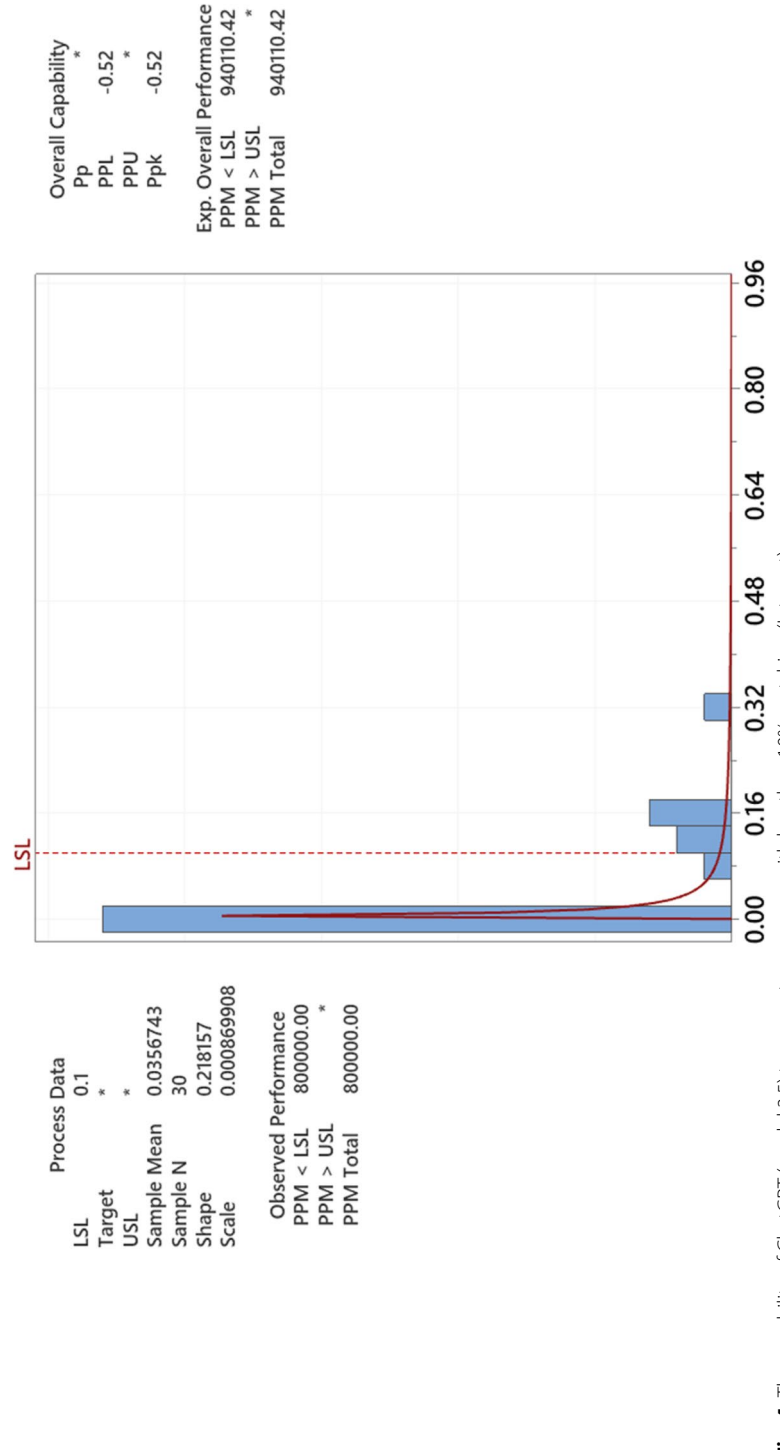

Fig. 4 The capability of ChatGPT (model 3.5) to generate responses with less than 10% matching (Internet)

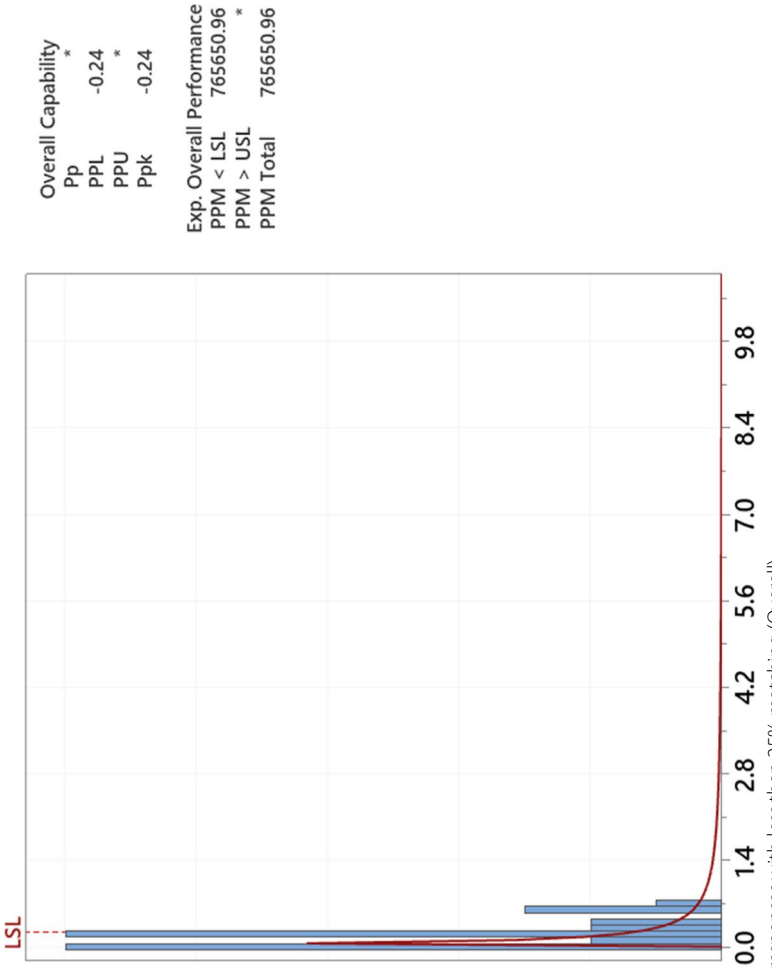

Fig. 5 The capability of ChatGPT (model 3.5) to generate responses with less than 25% matching (Overall)

Table 5 Summary of ChatGPT (model 3.5) to generate responses with less than 10% matching

Source of Matching	Observed Capability	Expected Capability
Overall	33.33%	66.5%
ChatGPT previously generated responses	50%	77.6%
students' global database	86.7%	97.3%
Internet	80%	94%

confidence level, we fail to reject the null hypothesis, indicating insufficient evidence to support a statistically significant difference between the text-matching percentages of ChatGPT (model 3.5) and (model 4).

Assessing repeatability and reproducibility of ChatGPT Models 3.5 and 4 in Generating authentic responses

To evaluate the chatbot's repeatability in generating novel and original responses, the initial response was recorded and regenerated twice more within the same chatbot session. After that, a new chatbot was created, and the same prompt was utilized to replicate the experiment, assessing the reproducibility of the chatbot's capacity to generate new and original responses.

The repeatability and reproducibility of ChatGPT (model 3.5) in generating authentic responses were examined using a Boxplot, as depicted in Fig. 11. The results indicate that the generation of responses by ChatGPT (model 3.5) remains consistent, regardless of whether the response is created within the same chatbot session or initiated by a new chat input. Similarly, the repeatability and reproducibility of ChatGPT (model 4) in generating authentic responses were assessed using a Boxplot, as illustrated in Fig. 12. The findings reveal that the generation of responses by ChatGPT (model 4) is also consistent, irrespective of whether the response is produced within the same chatbot session or prompted by a new chat input, akin to ChatGPT (model 3.5).

Strategies for mitigating risk and ensuring authenticity using ChatGPT

The integration of AI in academic contexts has presented new challenges regarding addressing academic misconduct. While technology can be utilized to invigilate students during exams, it is not as effective in preventing misconduct in take-home assignments. In order to address the misuse of AI, including GhatGPT, several strategies can be implemented.

1) Firstly, emphasizing the negative consequences of cheating and plagiarism and promoting self-transcendent ideals through implementing honor codes can help effectively reduce instances of academic misconduct.
2) The second point pertains to the restricted knowledge base of GhatGPT that is constrained by data that extends until September 2021 for both versions 3.5 and 4, and is not connected to the internet. Consequently, educators may develop assignments based on information that is not accessible to the model. Curiously, a response generated by ChatGPT model 4 demonstrated a lack of awareness regarding the mod-

Fig. 6 The capability of ChatGPT (model 3.5) to generate responses with less than 25% matching (ChatGPT previously generated responses)

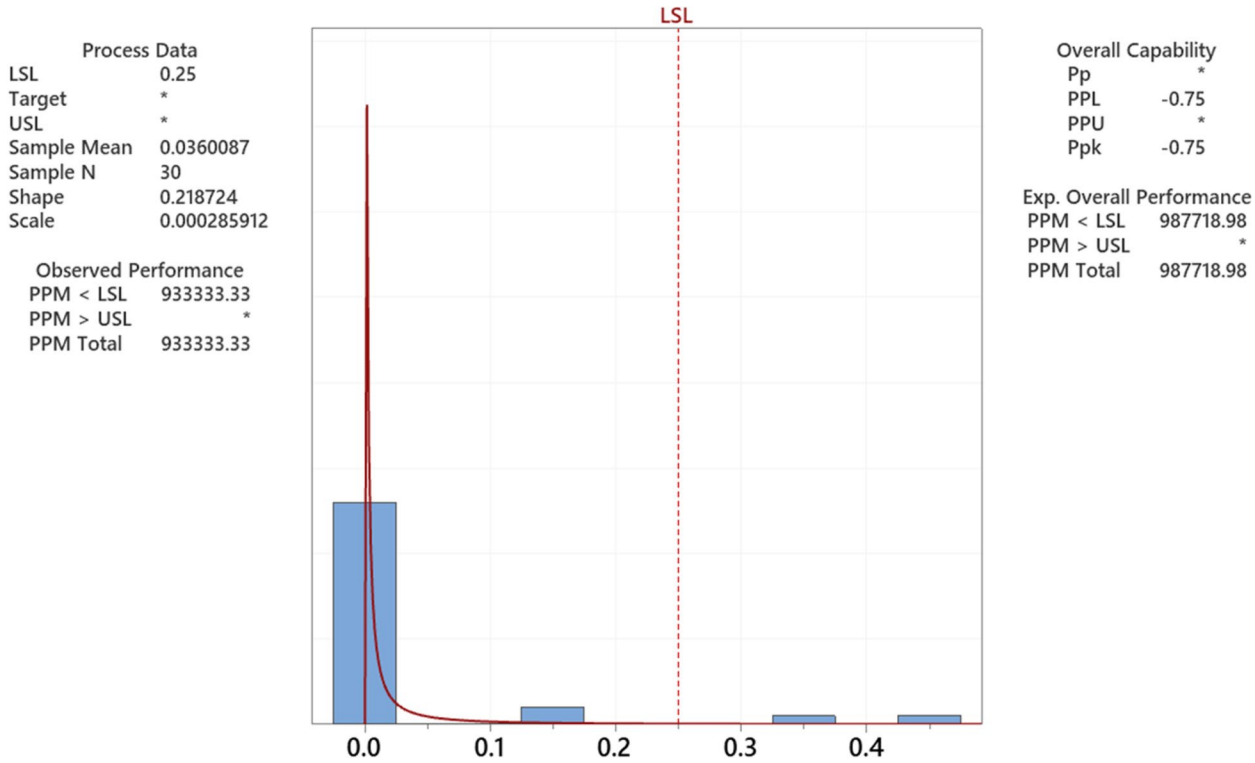

Fig. 7 The capability of ChatGPT (model 3.5) to generate responses with less than 25% matching (Students Global Database)

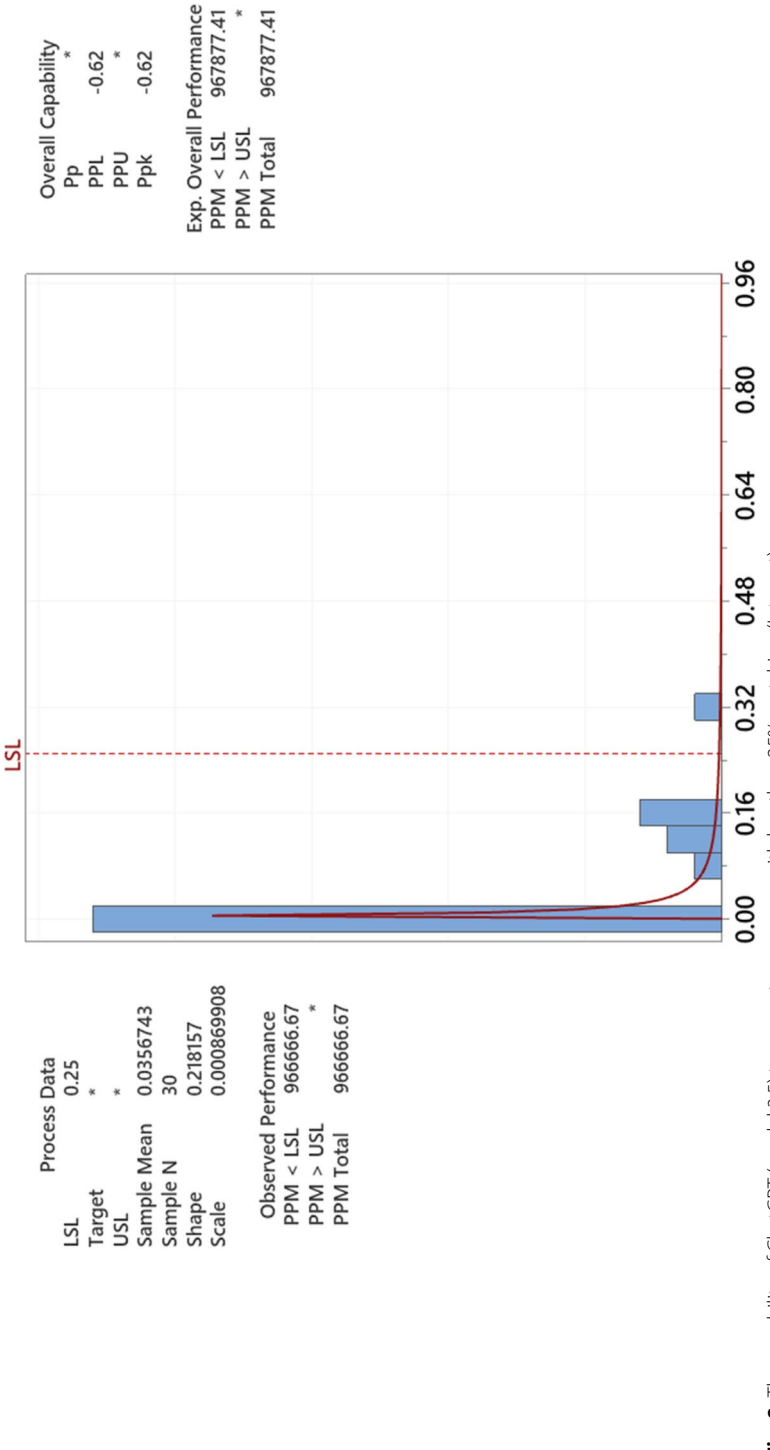

Fig. 8 The capability of ChatGPT (model 3.5) to generate responses with less than 25% matching (Internet)

Table 6 Summary of ChatGPT (model 3.5) to generate responses with less than 25% matching

Source of Matching	Observed Capability	Expected Capability
Overall	73.3%	76.6%
ChatGPT previously generated responses	83.3%	84.4%
students' global database	93.3%	98.8%
Internet	96.7%	96.8%

el's release, affirming that its knowledge is confined to September 2021. However, it should be acknowledged that this strategy may not be sustainable in the long term as updates to chatbots may overcome these limitations.

3) Thirdly, it is essential to note that these chatbots may not generate accurate references for the information they provide. In the academic realm, students must include appropriate references for all information in their assignments. To affirm this, ChatGPT (model 3.5) was prompted to provide five research papers on cooling towers in this study, It was discovered that the chatbot provided references in APA format, yet none of these references existed. Furthermore, ChatGPT (model 4) provides DIO hyperlinks to these references, however, they are linked to articles other than the ones cited.
4) Fourthly, since chatbots do not allow the upload or reading of imaged data such as graphs and charts, assignments should be designed to extract data from these sources.
5) Finally, integrating oral discussion into the evaluation process can offer insights into students' comprehension and awareness of the submitted work. However, it is essential to acknowledge the potential challenges in implementing oral assessments, such as logistical difficulties with large classes or language barriers for students whose first language differs from the course's instructional language. Educators should weigh the benefits of this approach against the practical constraints and seek alternative assessment methods that provide a fair evaluation of students' understanding while minimizing the risk of academic misconduct.

Conclusion

In conclusion, this study assessed the authenticity capabilities of ChatGPT models 3.5 and 4 in generating responses with less than 10% and 25% text matching, observing that ChatGPT model 4 might have a higher capability in generating authentic responses compared to model 3.5. However, a two-sample t-test revealed insufficient evidence to support a statistically significant difference between the text-matching percentages of both models. The repeatability and reproducibility of both models were also analyzed, showing that the generation of responses remains consistent in both cases. Notably, responses from ChatGPT model 4 were not regenerated from model 3.5, suggesting distinct algorithms and techniques in the newer model. Research findings have shown that ChatGPT models 3.5 and 4 can generate unique, coherent, and accurate responses that can evade text-matching software, presenting a

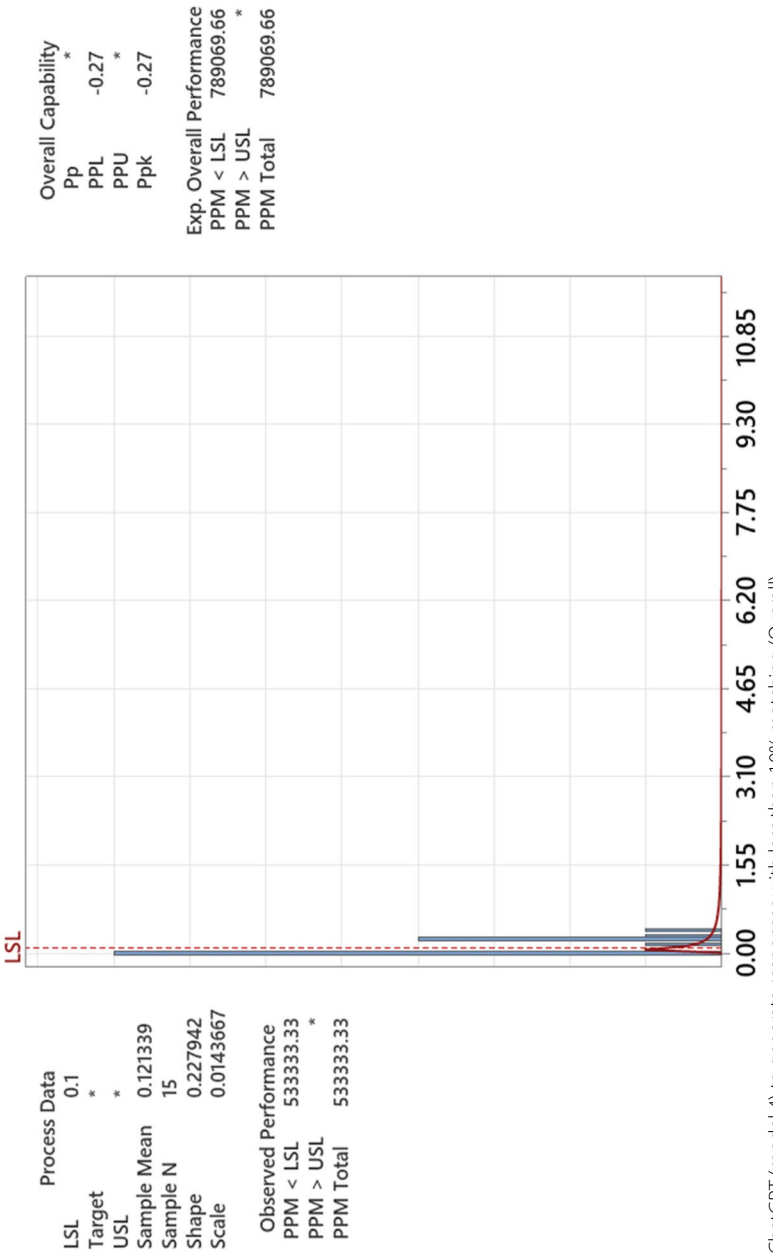

Fig. 9 The capability of ChatGPT (model 4) to generate responses with less than 10% matching (Overall)

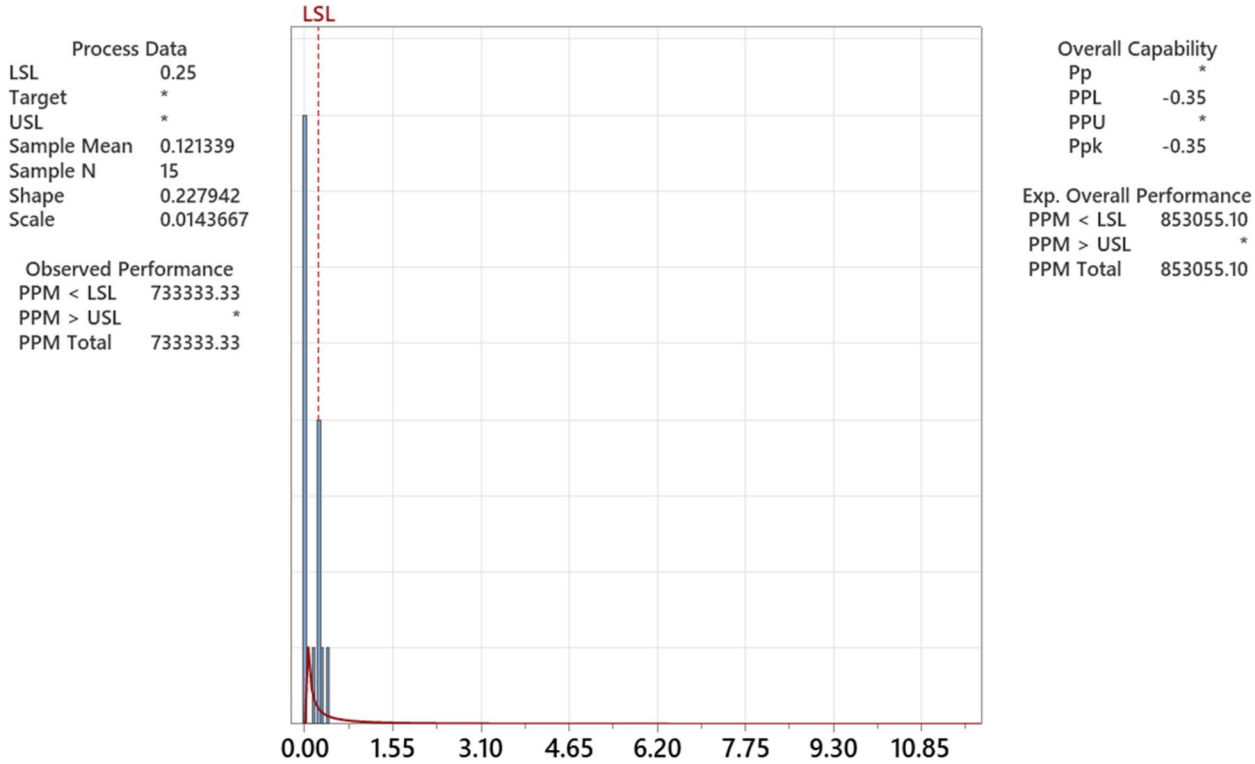

Fig. 10 The capability of ChatGPT (model 4) to generate responses with less than25% matching (Overall)

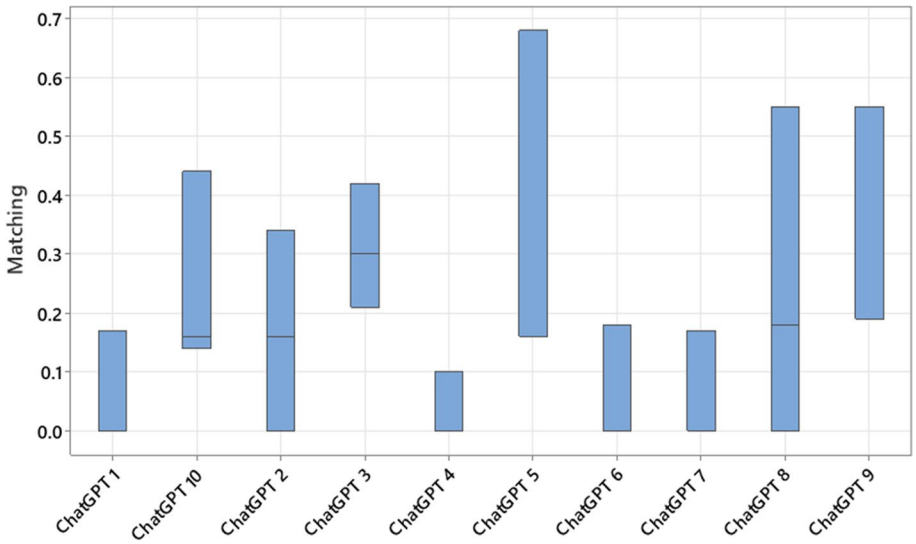

Fig. 11 Boxplot of total matching for the ten ChatGPT (model 3.5) responses

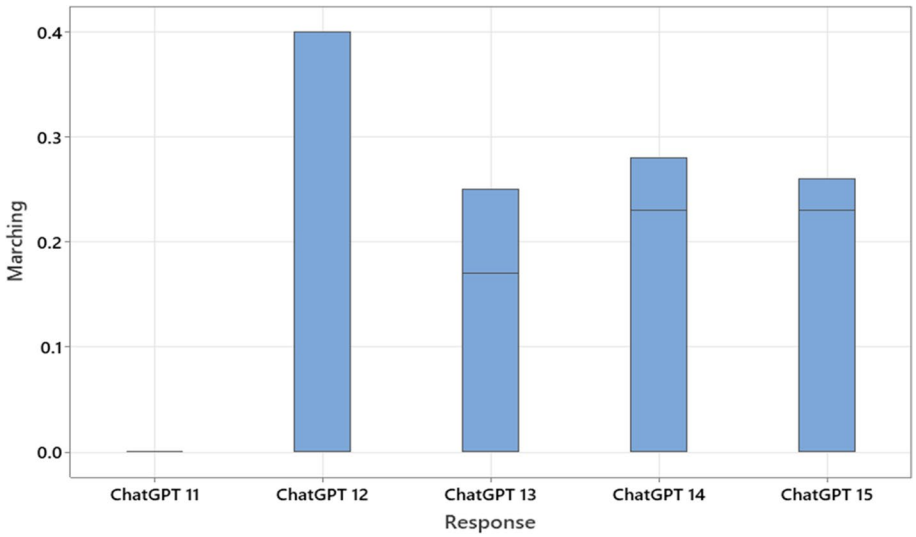

Fig. 12 Boxplot of total matching for the ten ChatGPT (model 4) responses

potential risk for academic misconduct. Therefore, assessors must acknowledge these limitations and actively seek alternative assessment methods to maintain academic integrity while leveraging AI integration's benefits. Several strategies can be employed to address the challenges posed by AI integration in academic contexts. These strategies include promoting self-transcendent ideals through implementing honor codes, considering the restricted knowledge base of ChatGPT, addressing inaccuracies in generated references, designing assignments to extract data from imaged sources, and integrating oral discussions into the evaluation process. However, educators must weigh the benefits of these strategies against practical constraints and seek alternative assessment methods to minimize the risk of academic misconduct.

Abbreviations

AI	Artificial intelligence
COPE	Committee on Publication Ethics.
Cpk	Capability Process Index
ICMJE	International Committee of Medical Journal Editors
LLM	Large Language Model
NLP	Natural Language Processing
PPK	Process Performance Index of Capability test
PPM	Part Per Million of Capability test

Acknowledgements
The publication of this article was funded by the Qatar National Library.

Authors' contributions
Ahmed M. Elkhatat: Conceived and designed the analysis; Collected the data; Contributed data; Performed the analysis; and Wrote the paper.

Funding
Open Access funding provided by the Qatar National Library. The Qatar National Library funded the publication of this article according to Springer Nature and Qatar National Library Open Access agreement. The Qatar National Library provides open Access funding. https://www.springernature.com/gp/librarians/open-research-for-librarians/sn-oa-agreements/qatar.

Availability of data and materials
All data and materials are available.

Declarations

Competing interests
The authors declare that they have no conflict of interest.

Received: 23 January 2023 Accepted: 11 June 2023
Published online: 01 August 2023

References

Alser M, Waisberg E (2023) Concerns with the usage of ChatGPT in Academia and Medicine: A viewpoint. Am J Med Open. https://doi.org/10.1016/j.ajmo.2023.100036
Bothe D (1998) Measuring Process Capability: Techniques and Calculations for Quality and Manufacturing Engineers. J Manuf Syst 1(17):78
Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
Foltýnek T, Meuschke N, Gipp B (2019) Academic Plagiarism Detection. ACM Comput Surv 52(6):1–42. https://doi.org/10.1145/3345317
Foltýnek T, Meuschke N, Gipp B (2020) Academic Plagiarism Detection. ACM Comput Surv 52(6):1–42. https://doi.org/10.1145/3345317
Hajrizi E, Zylfiu B, Menxhiqi L (2019) Developing a system for detecting the same content within the UBT academic institution, including special characters. IFAC-PapersOnLine 52(25):264–268. https://doi.org/10.1016/j.ifacol.2019.12.493
Jones M, Sheridan L (2014) Back translation: an emerging sophisticated cyber strategy to subvert advances in 'digital age' plagiarism detection and prevention. Assess Eval High Educ 40(5):712–724. https://doi.org/10.1080/02602938.2014.950553
King MR, chatGpt. (2023) A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. Cell Mol Bioeng 16(1):1–2. https://doi.org/10.1007/s12195-022-00754-8
Landau JD, Druen PB, Arcuri JA (2016) Methods for Helping Students Avoid Plagiarism. Teach Psychol 29(2):112–115. https://doi.org/10.1207/s15328023top2902_06
Montgomery DC (2020) Introduction to statistical quality control. John Wiley & Sons
Pizarro VG, Velásquez JD (2017) Docode 5: Building a real-world plagiarism detection system. Eng Appl Artif Intell 64:261–271. https://doi.org/10.1016/j.engappai.2017.06.001
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
Sakamoto D, Tsuda K (2019) A Detection Method for Plagiarism Reports of Students. Procedia Computer Science 159:1329–1338. https://doi.org/10.1016/j.procs.2019.09.303
Sánchez-Vega F, Villatoro-Tello E, Montes-y-Gómez M, Villaseñor-Pineda L, Rosso P (2013) Determining and characterizing the reused text for plagiarism detection. Expert Syst Appl 40(5):1804–1813. https://doi.org/10.1016/j.eswa.2012.09.021
Scanlon PM (2003) Student online plagiarism: how do we respond? Coll Teach 51(4):161–165
Yang A, Stockwell S, McDonnell L (2019) Writing in your own voice: An intervention that reduces plagiarism and common writing problems in students' scientific writing. Biochem Mol Biol Educ 47(5):589–598. https://doi.org/10.1002/bmb.21282
Alsallal, M., Iqbal, R., Amin, S., & James, A. (2013, 16–18 Dec. 2013). Intrinsic Plagiarism Detection Using Latent Semantic Indexing and Stylometry. 2013 Sixth International Conference on Developments in eSystems Engineering,

Anders, B. A. (2023). Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? *Patterns*, *4*(3). https://doi.org/10.1016/j.patter.2023.100694
Blackboard. (2023). *Blackboard Learn Platform*. https://www.blackboard.com/en-eu/teaching-learning/learning-management/blackboard-learn
Chen, Chiang, Storey (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165. https://doi.org/10.2307/41703503
Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1–12. https://doi.org/10.1080/14703297.2023.2190148
Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, *13*. https://doi.org/10.1016/j.jrt.2023.100060
Elkhatat, A. M. (2022). Practical randomly selected question exam design to address replicated and sequential questions in online examinations. *International Journal for Educational Integrity*, *18*(1). https://doi.org/10.1007/s40979-022-00103-2
Elkhatat, A. M., Elsaid, K., & Almeer, S. (2021). Some students plagiarism tricks, and tips for effective check. *International Journal for Educational Integrity*, *17*(1). https://doi.org/10.1007/s40979-021-00082-w
Fishman, T. (2009, 28–30 September 2009). *"We know it when we see it" is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright* 4th Asia Pacific Conference on Educational Integrity, University of Wollongong NSW Australia.
Francke, E., & Bennett, A. (2019). The Potential Influence of Artificial Intelligence on Plagiarism: A Higher Education Perspective. European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2019),
Frye, B. L. (2022). Should Using an AI Text Generator to Produce Academic Writing Be Plagiarism? *Fordham Intellectual Property, Media & Entertainment Law Journal*. https://ssrn.com/abstract=4292283
Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. https://doi.org/10.1101/2022.12.23.521610
Hinojo-Lucena, F.-J., Aznar-Díaz, I., Cáceres-Reche, M.-P., & Romero-Rodríguez, J.-M. (2019). Artificial Intelligence in Higher Education: A Bibliometric Study on its Impact in the Scientific Literature. *Education Sciences*, *9*(1). https://doi.org/10.3390/educsci9010051
Meuschke, N., & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, *9*(1). https://doi.org/10.21913/IJEI.v9i1.847
Minitab. (2023a). https://www.minitab.com/en-us/
Minitab. (2023b). *Expected overall performance for Normal Capability Analysis*. Minitab® 20. Retrieved 23 March from https://support.minitab.com/en-us/minitab/20/help-and-how-to/quality-and-process-improvement/capability-analysis/how-to/capability-analysis/normal-capability-analysis/interpret-the-results/all-statistics-and-graphs/expected-overall-performance/
Norvig, S. R. P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson https://www.pearson.com/en-us/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000003500/9780137505135?tab=accessibility
OpenAI. (2022). *Introducing ChatGPT*. Retrieved March 21 from https://openai.com/blog/chatgpt/
OpenAI. (2023). *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses*. Retrieved March 22 from https://openai.com/product/gpt-4
Qadir, J. (2022). Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. *TechRxiv. Preprint*. https://doi.org/10.36227/techrxiv.21789434.v1
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
Roostaee, M., Sadreddini, M. H., & Fakhrahmad, S. M. (2020). An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes. *Information Processing & Management*, *57*(2). https://doi.org/10.1016/j.ipm.2019.102150
Rozencwajg, S., & Kantor, E. (2023). Elevating scientific writing with ChatGPT: A guide for reviewers, editors... and authors. *Anaesth Crit Care Pain Med*, *42*(3), 101209. https://doi.org/10.1016/j.accpm.2023.101209
Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? https://doi.org/10.48550/arXiv.2303.11156
Sapci, A. H., & Sapci, H. A. (2020). Artificial Intelligence Education and Tools for Medical and Health Informatics Students: Systematic Review. *JMIR Med Educ*, *6*(1), e19285. https://doi.org/10.2196/19285
Siegerink, B., Pet, L. A., Rosendaal, F. R., & Schoones, J. W. (2023). ChatGPT as an author of academic papers is wrong and highlights the concepts of accountability and contributorship. *Nurse Educ Pract*, *68*, 103599. https://doi.org/10.1016/j.nepr.2023.103599
Williams, C. (2022). Hype, or the future of learning and teaching? 3 Limits to AI's ability to write student essays. *The University of Kent's Academic Repository, Blog post*. https://kar.kent.ac.uk/99505/

Publisher's Note

Dr. Elkhatat a distinguished academic with a Ph.D. in Chemistry, Six Sigma Black Belt, and Section Head of Research Planning at Qatar University, excels in research, pedagogy, and teaching. Obtained patents and published in respected scientific journals. A skilled digital designer with 15 years of experience, Dr. Elkhatat also contributes to teaching, lab safety, and infrastructure management. He consistently stays current with the latest analytical technologies. Beyond his academic and professional pursuits, Dr. Elkhatat engages in extracurricular activities, they judge science competitions, foster innovation, and promote science through mass media. As a YouTube content creator, he's garnered over 2 million views sharing scientific content globally. ORCID: 0000-0003-0383-939X.