

ORIGINAL ARTICLE

Open Access



# Artificial intelligence, text generation tools and ChatGPT – does digital watermarking offer a solution?

Thomas Lancaster<sup>1\*</sup> 

\*Correspondence:  
[thomas@thomaslancaster.co.uk](mailto:thomas@thomaslancaster.co.uk)

<sup>1</sup> Department of Computing,  
Imperial College London,  
London, UK

## Abstract

Text generation tools, often presented as a form of generative artificial intelligence, have the potential to pose a threat to the integrity of the educational system. They can be misused to afford students marks and qualifications that they do not deserve. The emergence of recent tools, such as ChatGPT, appear to have left the educational community unprepared, despite the fact that the computer science community has been working to develop and improve such tools for years. This paper provides an introduction to text generation tools intended for a non-specialist audience, discussing the types of assessments that students can outsource, showing the type of prompts that can be used to generate text, and illustrating one possible watermarking technique that may allow generated text to be detected. A small-scale study into watermarking suggests that this technique is feasible and show technical promise but should not be relied on as a solution to widespread use of artificial intelligence based tools by students. Alternative solutions are needed, including encouraging the educational community to work with artificial intelligence rather than against it. As such, the paper concludes by discussing seven potential areas for further exploration.

**Keywords:** Artificial intelligence, Academic integrity, Generative AI, ChatGPT, Text generation, Digital watermarking

## Introduction

The rapid growth of artificial intelligence (AI) powered solutions provides great potential for society and for education. In the workplace AI can enable increased proficiency and output, with the automation of many tasks that are mundane or tedious. In education, students can be engaged through personalised learning and educators can deploy new and emerging pedagogical methods.

AI based systems can also bring with them problems that were not originally considered. Some may approach these problems as opportunities for change. Within the context of education, a key challenge is the ability of students to use generative artificial intelligence to produce original text which they then use to answer assignment questions. In many ways, this provides answers akin to those produced by contract cheating



providers, giving text that appears to be original but is not by the student. Unlike commercial contract cheating, generative AI allows such text can be produced almost instantly and at minimal or no cost to the student. With the right prompt to make text, a student could even produce a different answer to an essay style question for every member of their class with the click of a button.

When a student is generating answers, rather than working on solving problems for themselves, they are likely to be missing the opportunity to learn. The students may no longer be meeting the learning outcomes for a given assessment. Such an approach is likely to be considered as a breach of academic integrity principles in many institutions, although policies do not always explicitly consider the use of text generation tools.

Although language models are not themselves new, the launch by OpenAI of the ChatGPT text generation tool in November 2022 made their use accessible to a wider audience. Reportedly, ChatGPT gained over one million users within its first week, a scale of uptake previously unheard of for a technology product (Altman 2022). Although not aimed specifically at an academic audience, the potential of ChatGPT to generate academic text is there and many examples of this have appeared across social media.

In many ways, academic text generation using automated techniques can be considered as a successor to contract cheating, defined originally by Clarke and Lancaster (2006). With contract cheating, a student hires or uses a third party in order to dishonestly complete their assessments. Key research into contract cheating has considered the reasons student outsource work (Amigud and Lancaster, 2019), how this spans multiple disciplines (Lancaster, 2020), detection techniques (Rogerson 2017), the risks posed to students exposing themselves to third parties (Yorke et al., 2020), the size of the industry (Owings and Nelson 2014), the writers themselves (Walker 2020), and many other areas.

Text generation tools remove the need for a student to hire anyone for many tasks. The students can complete the work for themselves. Many technical publications exist discussing the continual development of such tools, but these are not being considered in this paper as they are out of scope for an academic integrity audience. Few papers considering the tools from an educational or academic integrity perspective exist that predate the release of ChatGPT. One key source is an opinion piece by Sharples (2022), which notes that current techniques for detecting plagiarism will be of little use against generated essays, a view not dissimilar to that expressed regarding contract cheating.

The related area of paraphrasing tools also demonstrates some of the threats to academic integrity that text generation tools may pose, albeit on a much smaller scale. Roe and Perkins (2022) reviewed research into these, noting that the use of paraphrasers was often hard to identify and not explicitly covered in university policies. This work built heavily on previous studies, such as tests by Rogerson and McCarthy (2017) showing that paraphrasing tool output could confuse originality checking software, and similar results from investigations into the related area of essay spinning (Lancaster and Clarke, 2009).

This paper is intended to introduce modern AI based text generation tools to the educational community, with particular reference to ChatGPT. The use of these tools may be considered as a breach of academic integrity if their use has not been approved for assessment, since a student would be getting credit for work for which they have bypassed learning. The discussion will also consider examples of the style of assignment

solutions that can be generated. Aaronson (2022) has suggested that the digital watermarking of generated text may be possible and this may provide a level of protection against misuse. A more detailed explanation will follow later in the paper, but the digital watermarking method presented here involves part of the tool's textual output being generated in a systematic and repeatable way, rather than being random. This thus should allow the repeated elements of text to be detectable. Recommendations surrounding how educators may wish to consider reacting to the growth of text generation tools are also provided.

The paper is deliberately presented in a case study and example-based format, designed to be accessible to educators and the academic integrity community, rather than in a formal mathematical manner. The approach draws heavily on research talks and demonstration given by the author, as well as the style of examples they have previously shared on social media.

### **About ChatGPT**

In its simplest terms, ChatGPT is a chatbot. Type in something to say, ChatGPT will respond within the parameters available to it, then it is possible to have something resembling a conversation. As a chatbot, ChatGPT has a form of memory of earlier aspects of the conversation, so future responses can refer and build upon this. Such an approach means that it is possible to use ChatGPT to refine ideas.

Perhaps the best way to introduce ChatGPT is through a short conversation with it. The conversation, as shown in Table 1, is unedited, although the initial instruction has been carefully chosen. It should be noted that ChatGPT and its underlying Generative Pretrained Transformer (GPT) models are under continual development. All examples of ChatGPT output presented in this paper were obtained in December 2022.

ChatGPT self describes itself as a "language model". Despite the common use of the term artificial intelligence, ChatGPT responds in a predetermined way, based upon its trained model, the input data, earlier parts of the conversation and a random number, known as a seed. That does not mean that the replies are random. If the same input is given with the same seed number, ChatGPT will deliver the same replies. However, the seed number being used by ChatGPT at any time is not visible to users and this may change repeatedly during a chat session.

What this does mean is that if ChatGPT was asked the same questions as in Table 1 again, the responses would almost certainly be different. For example, Table 2 shows four further unedited responses to the first question from Table 1. Although there are elements of overlap between the different responses, there are differences in phrasing, in the examples used and in how the technology is described. Informal experimentation suggests also that the overlap is less pronounced with requests for longer answers or where the chatbot is allowed to be more creative.

As it is a language model, ChatGPT can generate answers that appear well-written, but which are factually incorrect. This may be because the data used to train ChatGPT was itself incorrect, but it may also be because ChatGPT is also trying to produce a particular form of words. For example, ChatGPT can generate realistic looking academic references to papers and journals that do not exist. It can produce false biographies of

**Table 1** Conversation with ChatGPT

---

Author	I am going to ask you a series of questions. Answer them in a way that is understandable by a non-specialist audience of educators. Limit each response to one paragraph The first question is: What is ChatGPT?
ChatGPT	ChatGPT is a large language model that was trained by OpenAI. It is capable of generating human-like text based on the input it receives, allowing it to hold conversations with people. It is designed to be able to understand and respond to a wide range of topics, and can be used in a variety of applications, such as chatbots and language translation tools
Author	How was ChatGPT developed?
ChatGPT	ChatGPT was developed using a technique called unsupervised learning, where the model is trained on a large dataset of text without being specifically told what to look for or how to interpret the data. This allows the model to learn to generate text that is similar in style and content to the text it was trained on, allowing it to hold conversations and generate responses that are natural and human-like
Author	How does ChatGPT work?
ChatGPT	ChatGPT works by using machine learning algorithms to process and analyze large amounts of text data. When the model is given an input, it uses this information to generate text that is similar in style and content to the text it was trained on. The model is able to generate responses that are human-like because it has been trained on a vast amount of data, allowing it to understand and use a wide range of words, phrases, and sentence structures. This allows it to hold conversations with people and generate responses that are natural and relevant to the conversation
Author	How much data was ChatGPT trained on?
ChatGPT	As a large language model, ChatGPT was trained on a vast amount of data in order to be able to understand and generate human-like text. The exact amount of data that was used to train ChatGPT is not publicly available, but it is likely to be in the tens or hundreds of gigabytes. This amount of data is necessary to train a model of ChatGPT's size and capabilities, as it allows the model to learn to generate text that is natural and human-like
Author	What type of sources were included in the training data?
ChatGPT	As a large language model, ChatGPT was trained on a wide variety of text data from many different sources. This includes books, websites, and other publicly available text data. Because I am a language model, I am able to process and understand a wide range of topics, so the training data is likely to be diverse and cover many different subjects. However, I do not have access to the internet and am unable to provide specific details about the training data that was used to develop ChatGPT

---

**Table 2** Responses to question with different random seed

---

ChatGPT is a language model developed by OpenAI. It is a type of artificial intelligence that is designed to be able to generate human-like text in response to the input it receives. It is based on the GPT-3 model, which is one of the most advanced language models currently available. ChatGPT is specifically designed to be used in chat applications and other conversational interfaces, where it can generate responses to user input in a way that is natural and engaging

ChatGPT is a large language model trained by OpenAI. It uses a powerful neural network to generate human-like text, allowing it to hold conversations and answer questions on a wide range of topics. It has been trained on a vast amount of text data, giving it a deep understanding of language and the ability to provide accurate and relevant responses to users

ChatGPT is a large language model trained by OpenAI. It uses deep learning algorithms to generate human-like responses to natural language inputs. ChatGPT is designed to be able to generate a wide range of responses and hold complex conversations, making it useful for applications such as natural language processing, language translation, and text summarization

ChatGPT is a natural language processing (NLP) model developed by OpenAI. It is a variant of the GPT-3 model, which is a powerful language processing tool that is able to generate human-like text. ChatGPT is specifically designed to be used in chat applications, where it can help users generate responses to messages in a conversational way. It uses machine learning techniques to understand the context of a conversation and generate appropriate responses. This can help improve the user experience in chat applications and make the conversation feel more natural

---

real people. There are situations where the use of the information produced may be dangerous. Despite attempts by OpenAI to implement safeguards, it has been possible to get ChatGPT to produce ethically questionable responses, write fake news stories or

contribute to the spread of conspiracy theories, for example, by asking it to roleplay a situation or by providing what-if scenarios.

The paper will continue by considering the ways in which systems like ChatGPT can potentially be misused by students in an academic setting.

### **Academic assessment use cases for ChatGPT**

The information in this section is provided for reference purposes for the educational community, based on experiments conducted by the author. It is not intended as a guide to students as to how they can cheat, but many students will be able to figure this out for themselves and some will have devised other methods. By its very nature, this section is also far from complete. There are many types and styles of assignments out there for which answers can be produced.

ChatGPT is a chatbot for automatically generating text. This means that it cannot easily generate other forms of media such as images, although alternative tools exist for such purposes. ChatGPT can also generate the prompts, or guidelines, for image generation tools. Perhaps, most crucially, for technically adept students, ChatGPT can generate source code and other forms of instructions for computers. This not only allows it to complete computer programming assignments, but also to generate runnable code to provide answers to other types of assessments.

Table 3 provides examples of a selection of assessment related tasks that ChatGPT can be used for. It is always worth remembering that, if a student is unhappy with a response, they can always run a prompt again, or they can provide follow-up questions or instructions to generate an answer that meets their requirements.

The discussion of potential academic misconduct surrounding ChatGPT use is subject to nuance. A student may be wishing to use this as a form of automated copy editor or grammar checker, to make text they have written but which they know is imperfect, with the intention of smoothing this out into a more academic style. This may represent a legitimate reason to use such technology, rather than be forming an attempt to cheat.

A human-led editing stage may also take place. Students may not always want to submit answers as generated. They may choose to run the text through other systems, such as existing paraphrasing tools or grammar checkers, to make this look less like the originally generated version.

A level of critique of generated information is also necessary. This information provided by generative AI may be factually inaccurate and so this should always be checked. ChatGPT does have flaws as its knowledge of recent events is limited and it appears often unable to produce accurate references. But there are also ways to augment the knowledge that ChatGPT has and there are bound to be technically minded students looking for ways around the limitations of ChatGPT.

### **Detecting ChatGPT use with digital watermarking**

A typical educational response to technology being used to facilitate academic dishonesty has been to develop detection technology. Similarity checking software has been used to find plagiarism, although this is unlikely to identify the varied and originally form text generated by ChatGPT. The launch of ChatGPT saw multiple companies compete to release AI detection software, a largely mysterious approach where input text

**Table 3** ChatGPT used to assist with assessment production

<b>Assessment Type</b>	<b>Example Prompt</b>	<b>Commentary (And Output)</b>
<i>Essay</i>	Describe and discuss the role of the unconscious mind in decision making, including its influence on our choices, behaviors, and emotions. How does the unconscious mind differ from the conscious mind, and how do the two interact in the decision-making process? What are some examples of the unconscious mind influencing decision making, and how can we become more aware of its influence in our own lives? Use examples and research from psychology to support your arguments	ChatGPT was used to generate a sample essay question. The output is 467 words long, beginning with the following: <i>"The unconscious mind is a part of the mind that contains the thoughts, memories, and desires that are not currently in our conscious awareness. It is the part of the mind that operates automatically and without our conscious control, influencing our decisions, behaviors, and emotions."</i>
<i>Reflection</i>	You are a student returning from a work placement at Meta, where you worked as a software engineer. Write a 500 word reflective analysis on your work and its impact on your future career development	415 words starting: <i>"As a software engineer at Meta, I had the opportunity to work on a number of exciting projects that have had a profound impact on my future career development. During my time at the company, I was able to gain valuable experience working on both front-end and back-end projects, which has helped me to develop a more well-rounded set of skills."</i>
<i>Marketing</i>	You are a marketing student who is working on the launch of a new home security system which uses machine learning to reduce crime. Produce sample sales copy for the new system including a headline, a benefit statement and responses to possible objections	285 words total, starting: <i>"Headline: Introducing the SmartSafe: the most advanced home security system on the market! Benefit Statement: With the SmartSafe, you can enjoy peace of mind knowing that your home is protected by the latest in machine learning technology. Unlike traditional security systems, the SmartSafe is constantly learning and adapting to your home's unique patterns, so it can effectively identify and deter potential intruders."</i> Also listed three objections and responses
<i>Rewriting</i>	Rewrite the following text to improve the written English and so it is different enough that it does not look like plagiarism: Schools are increasingly using machine learning to improve the education experience for students (note truncated – original text generated by ChatGPT – 103 words total)	New text starting: <i>"As schools continue to incorporate machine learning into their classrooms, the education experience for students is becoming increasingly personalized."</i> (109 words total)
<i>Programming</i>	Write a program in Haskell that creates and displays a simple animation. The program should allow the user to specify the number of frames in the animation, as well as the dimensions of each frame. The program should generate each frame by randomly selecting and positioning a set of graphical objects, such as circles, squares, or triangles, on the screen. The program should then display the frames in sequence, creating the illusion of movement. The user should be able to control the playback speed and direction of the animation	Generates code that can be copied and pasted Can also ask follow-up question "Explain how the code works" to get a guide starting: <i>"The code uses the gloss library to create a simple animation in Haskell. It defines a type called Frame to represent a single frame in the animation, which consists of a list of graphical objects and their positions."</i>

**Table 3** (continued)

<i>Assessment Type</i>	<i>Example Prompt</i>	<i>Commentary (And Output)</i>
<i>Presentation Slides</i>	Create a PowerPoint presentation about the ethics of air traffic control systems. Generate a title slide, plus appropriate topics and bullet points for the remaining slides. Then generate a Python program which I can run to produce the PowerPoint presentation. The code should be complete, including any imports needed. Content should be stored in an appropriate data structure. An efficient method should be used to generate the slides	This produces source code which can be run to generate a PowerPoint presentation which is 6 slides long A further request "Provide a script for the presentation. Split this by slide and give detailed examples." Gives a 724 word script, starting: "Good morning everyone, and welcome to our presentation on the ethics of air traffic control systems."

would receive a probability score or a yes/no response indicating if this was written by AI, but with little ability to explain the reason for the decision. Although such technology may perhaps have some use as a conversation starter with students, the accuracy of their results can be questioned, and the risk of these tools being used to falsely accuse students of academic misconduct appears great. If detection is to be considered as a potential solution to AI aided academic misconduct, alternative detection approaches are needed.

The method of digital watermarking is one that is used to add a hidden signature or mark to a digital document. The watermark is usually embedded in a way that is difficult for someone to remove or alter. Usually, the watermark is detected only using specialist software. One example would be to change the pattern of spaces between words and characters in a file in a way that is unique, but not easily visible to humans. That particular approach is unlikely to work for ChatGPT generated content, where extra spaces are likely to be removed.

It has been noted that ChatGPT has an alternative method of watermarking in development (Aaranson 2022), although the information provided is technically complex. This section presents a simplified example of the proposed approach, complete with an example. ChatGPT has been used to generate much of the data needed. This removed even the need for a computer programmer to manually write source code to conduct the small scale study presented and to process the results.

The approach relies on the idea of using a seed number as part of the generation process. As discussed earlier, the seed number is usually random. Detection depends on the fact that, if the same seed number is used with the same input data, the same output should always be produced. For this type of digital watermarking, the seed number selected is kept secret by ChatGPT creators OpenAI and never released.

A high-level process for both text generation and text detection could work as indicated in Table 4. This method is deliberately simplified and it is noted that there are additional complexities with applying this at scale.

The generation method uses non-overlapping 5-grams, but the detection method uses overlapping 5-grams. To illustrate the difference, consider the phrase: "The friendly robot greeted the visitors with a cheerful beep and a wave of its metal arms."

Ignoring punctuation, the non-overlapping 5-gram would be:

- "The friendly robot greeted the"

**Table 4** Sample digital watermarking processes

Generation	<ol style="list-style-type: none"> <li>1. Generate the text using the current ChatGPT process and a random seed</li> <li>2. Split the text into 5-grams (non-overlapping sequences of 5 consecutive words)</li> <li>3. Where possible, remove the 5<sup>th</sup> word in each 5-gram, then generate an alternative word that the language model behind ChatGPT indicates will fit, but which will not change the meaning of the original text. Replace the previous word with the new word. Replacing the word may not always be possible. For example, it might not be sensible to find an alternative if the word is “the”. The replacement should use the secret seed so that the word generated with the same input is always consistent</li> <li>4. Return the text with the replacement words to the users as the output</li> </ol>
Detection	<ol style="list-style-type: none"> <li>1. Split the text into overlapping 5-grams</li> <li>2. Iterate through the 5-gram. In each case, remove the 5<sup>th</sup> word from the 5-gram and then use the language model behind ChatGPT to identify what the expected word should be. This process uses the secret seed. Record if the removed word and expected word is the same</li> <li>3. Calculate a text generation score based on: the number of n-grams with a matching word / the total number of n-grams</li> </ol>

- “visitors with a cheerful beep”
- “and a wave of its metal”

The overlapping 5-gram would be:

- “The friendly robot greeted the”
- “friendly robot greeted the visitors”
- “robot greeted the visitors with”
- “greeted the visitors with a”
- “the visitors with a cheerful”
- “visitors with a cheerful beep”
- “with a cheerful beep and”
- “a cheerful beep and a”
- “cheerful beep and a wave”
- “beep and a wave of”
- “and a wave of its”
- “a wave of its metal”
- “wave of its metal arms”

The detection stage needs to consider more 5-grams in case the student has added new words to the text. For example, even adding a two-word student name to the start of the document would change the non-overlapping 5-grams that would be produced.

Providing the replacement word choices are sensible, the addition of the secret digital watermark should not be obvious to the user.

#### Example of digital watermarking

It may be more instructive to see an example of this process in action.

To aid with this ChatGPT was used first to generate a sample text, then to suggest appropriate word replacements. A prompt was given to avoid the need to generate code.

The first prompt and the text produced is shown in Table 5. Note that, despite attempts to reduce the length of the text, the produced text was still 51 words long.

**Table 5** Sample text generation prompt and text

Prompt	Write me a concise description of how robots work aimed at a non-specialist. The description should be between 30 and 40 words long, including filler words. Do not exceed 50 words, including filler words
Generated Text	Robots are machines that can be programmed to perform tasks automatically. They typically consist of a body, sensors to gather information about the environment, and a control system to interpret the information and direct the robot's movements. Some robots are able to learn and adapt to new situations through artificial intelligence

**Table 6** Prompt to generate replacement words

<p>You are operating as a helper to consider different ways of writing the description. You will suggest new words that could replace the current words in the description. If the replacement word you suggest is substituted for the current word, the sentence should still make grammatical sense and preserve the current meaning of the description</p> <p>You will not suggest replacements for the word "robots"</p> <p>You will not suggest replacements for any word that has the substring "robot"</p> <p>You will not suggest replacements for function words and filler words</p> <p>Make a table with 5 columns. The first column should contain every word from the previous description, including filler words and any words that you are not suggesting replacements for. Do not consider punctuation. The second column should be the position number of that word in the description</p> <p>Put suggested replacement words in columns 3 to 5. If you have been told to not suggest replacements for the word, leave those cells blank</p> <p>Do not list the word being replaced in columns 3 to 5. It is not valid to replace a word with itself</p> <p>Do not list any repeated words in columns 3 to 5</p> <p>Only suggest valid replacement words. You may suggest 0, 1, 2 or 3 replacement words, so long as they are good replacements</p> <p>Show the first 30 rows of the table</p>
--

The second prompt is shown in Table 6. This prompt is more complex, as attempts were made to ensure that the suggested word replacements were sensible and that the key topic word was not replaced. For larger scale work, this would be completed more programmatically. The prompt succeeded in generated a table, but as this is beyond the maximum output length for ChatGPT, it was necessary to generate the first 30 rows, then the remaining rows. The collected output is shown in Table 7.

For the purposes of the example, assume that the secret seed always maps to the Replacement 1 column. The generation method from Table 4 would essentially replace words 5, 10, 20, 30, 45 and 50 (6 words in total).

Table 8 shows the originally generated text and the digitally watermarked text. The different words are shown in bold. The words that were unchanged, but in the 5<sup>th</sup> position in the 5-gram are shown underlined.

Now, consider this process being used for detection purposes on digitally watermarked text submitted by a student. Still using the digitally watermarked text in Table 8 as an example, detection has to consider that additional words, such as the student's name or an essay title, may have been added before the text, shifting the words along. To avoid missing any matches, all 47 different (overlapping) 5-grams within the text need to be considered.

In each 5-gram, the first four words are used along with the secret seed to predict the fifth word. The predicted fifth word is then checked against the actual fifth word, as submitted by the student. In the example, there will be a minimum of 10 matches from

**Table 7** Suggested replacement words

<i>Word</i>	<i>Position</i>	<i>Replacement 1</i>	<i>Replacement 2</i>	<i>Replacement 3</i>
Robots	1			
are	2			
machines	3	devices	gadgets	systems
that	4			
can	5	may	might	could
be	6			
programmed	7	configured	set up	coded
to	8			
perform	9	complete	carry out	execute
tasks	10	jobs	functions	operations
automatically	11	unmanually	unconsciously	independently
They	12			
typically	13	usually	commonly	generally
consist	14	comprise	contain	include
of	15			
a	16			
body	17	frame	chassis	structure
sensors	18	detectors	probes	transducers
to	19			
gather	20	collect	obtain	acquire
information	21	data	knowledge	facts
about	22	regarding	concerning	on
the	23			
environment	24	surroundings	atmosphere	context
and	25			
a	26			
control	27	regulation	governance	management
system	28	platform	technology	setup
to	29			
interpret	30	understand	analyze	explain
the	31			
information	32	data	knowledge	facts
and	33			
direct	34	guide	lead	steer
the	35			
robot's	36			
movements	37	actions	behaviors	motions
Some	38			
robots	39			
are	40			
able	41	capable	competent	qualified
to	42			
learn	43	study	acquire	gain
and	44			
adapt	45	change	modify	adjust
to	46			
new	47	different	alternative	unfamiliar
situations	48	circumstances	conditions	scenarios
through	49	via	with	by means of

**Table 7** (continued)

<i>Word</i>	<i>Position</i>	<i>Replacement 1</i>	<i>Replacement 2</i>	<i>Replacement 3</i>
artificial	50	synthetic	digital	virtual
intelligence	51			

**Table 8** Comparison of originally generated text and digitally watermarked text

<i>Original</i>	<i>Digitally Watermarked</i>
Robots are machines that <b>can</b> be programmed to perform <b>tasks</b> automatically. They typically consist <u>of</u> a body, sensors to <b>gather</b> information about the environment, <u>and</u> a control system to <b>interpret</b> the information and direct <u>the</u> robot's movements. Some robots <u>are</u> able to learn and <b>adapt</b> to new situations through <b>artificial</b> intelligence	Robots are machines that <b>may</b> be programmed to perform <b>jobs</b> automatically. They typically consist <u>of</u> a body, sensors to <b>collect</b> information about the environment, <u>and</u> a control system to <b>understand</b> the information and direct <u>the</u> robot's movements. Some robots <u>are</u> able to learn and <b>change</b> to new situations through <b>synthetic</b> intelligence

the 47 5-grams, namely the words already shown in Table 8 as bolded or underlined. The bolded words are there due to the digital watermarking and so could be expected to increase the number of matches above those obtained from human written text, although the detection engine would not know why the match occurred. There will likely be some further matches generated purely by chance when processing the other 5-grams with the secret seed value.

This number is deliberately not being presented here as any form of percentage. This percentage-based approach causes confusion with existing software used to check for similarity and identify possible plagiarism. It is not clear how the results of the detection process would be presented to an end user, whether an educator, assessor or student, but explainable aspects of this process need to be considered. It may, for example, be the case that multiple matches in close proximity provide a greater indication of the use of generative AI than sporadic matches throughout a longer document. Such matters will have to be taken into account as part of the presentation.

### Practicalities of this approach

Based on the example presented, the suggested digital watermarking method for automatically generated text, based on Aaranson (2022) shows technical promise. Much of the devil here is in the detail. Just getting this technology to a level where users understand what is happening, can interpret the detection tool output, and feel able to trust that output, is going to be difficult.

The example given is simplified for demonstration purposes, with deliberately short text. In reality, the algorithmic methods outlined in Table 4 involve additional steps and checks. It is likely that the replacement phase will be done at the same time as generation, not afterwards. The process of splitting the text will not consider things as words and 5-grams, but will use whatever method of tokenisation is internal to the model. The details are unlikely to be made public.

If the 5-gram approach presented was used and known, in practice a student could defeat attempts at detection by themselves changing every 5<sup>th</sup> word of the text.

A large volume of data will be needed to develop the final algorithms and to calculate the thresholds at which to determine that text is likely to have been produced by ChatGPT.

There is a further question to be asked regarding who will have access to the detection component of the process. This is likely to be sold as a commercial offering. If students can access it, they may be able to work out how to change the generated text to avoid detection. Others with access may offer paid services disguising text to avoid detection. Educational institutions wishing to use a detection service supplied by OpenAI will have to integrate this into their processes and consider the budgetary implications. Although detection may be technically possible, this is not a failsafe or fool proof solution. Universities may be better placed focusing their attention into alternatives to detection.

### **The future**

The pace of change with artificial intelligence, machine learning and text generation technologies is phenomenal. This paper provides an overview of emerging technology, correct at the time of writing, but with continual developments and competitors on the horizon. A continual step change in the abilities of language models appears likely, one formed with incremental improvements based on ever-increasing data and access to records of user interactions. The principles behind the technologies are likely to remain consistent with the ideas expressed in this paper.

The underlying message from this paper must be for educators and the academic integrity community to engage with emerging artificial intelligence technologies, not to ignore them or to fight against them. The discussion about how to do this is one to be had sooner rather than later. But as with the way that contract cheating has been addressed, it is proposed that multiple approaches will be needed.

As a starting point for discussion, this paper proposes seven areas for exploration, all of which should be interlinked.

*Policy development* – quality assurance bodies and individual universities are beginning to consider how to address artificial intelligence use in the classroom, although this is always a challenging task when the technology itself continues to develop. It is hard to define what good policy in this area should look like, and if this should encourage the use of AI to be completely banned or embraced. Policies need to be developed that are agile to change and which consider how AI will be used externally in the future. More regional consistency across university approaches to policy are also needed, so universities need to be supported to work collaboratively to develop the next generation of academic integrity related policies.

*Student training* – clear training needs to be available for students to understand what use of artificial intelligence is acceptable in different situations, how to acknowledge this use and the ethical implications of such technology. Students also need to realise how to use this technology and to see its limitations, including the potential for wrong answers to be produced, something which could be dangerous if, for example, these answers were used as medical advice.

*Staff training* – staff need to be supported to understand the implications of technology and to realise that artificial intelligence solutions are not magic or a black box. All staff in the educational community, from senior leadership, through to teaching staff

and professional support, need to be aware of the necessary changes to educational processes. It is recognised that the number of people available to provide quality training in this field is limited, so a train-the-trainer style of programme may also be necessary here.

*Discipline specific interventions* – too often, education considers all disciplines as one and the same. In reality, there are different styles of assessment, the needs of employers in various fields differ and professional and accreditation bodies have their own requirements. A computer science student, for example, may need to not only understand and embrace this technology, but may also have to be actively involved in developing its next generation. A history student may need to be alert to misrepresented knowledge and date inaccuracies. A medical student could have to use AI to find diagnostic information published only in obscure papers. Since many of the details of current teaching and assessment are best known by academics in their own fields, that is where much of the discussion has to take place, all aided by wider principles that should be developed to be applicable more generally.

*Assessment design* – the fitness for purpose of many current assessment methods has to be questioned. If a student can type in an assignment question and get a valid answer within seconds, perhaps that assessment is no longer suitable. This does not mean that all assessment needs to be closed book exams, but alternatives need to be considered, perhaps using artificial intelligence as part of the process, perhaps finding other ways to ensure that students are actively engaged.

*Detection* – this paper has discussed one possible detection technique, but its use depends on a third party and has only been developed for one particular text generation technology. A more general question of detection needs to be explored. This may include human detection. Educators should be encouraged to look at the details of documents, rather than relying on a high level overview. Inaccurate information may need to be questioned. Such information can occur, for example, in the references list of ChatGPT generated papers, where sources are listed with accurate looking references, but where the paper title does not exist, or a Digital Object Identifier (DOI) does not resolve. This is because the reference text has been generated to have the correct format, not to be an existing source. Detection is never a solution, but it can provide a deterrent effect against intentional breaches of academic integrity.

*A student partnership approach* – no changes should ever be made without having students at their forefront, ideally leading the way. Student Unions and other similar international groups have a key role to play here in representing the interests of students. In all, students should be able to see how changes and interventions are beneficial for their future and preparing them for employment.

The idea of a whole community approach is central to academic integrity practice. Here, the community needs to extend beyond education. Text generation tools will not go away. Many in education, especially those working in the computer science fields, are actively working to improve them. Education must be ready to embrace change and must be ready now.

#### Abbreviations

AI	Artificial Intelligence
DOI	Digital Object Identifier
GTP	Generative Pre-trained Transformer

**Acknowledgements**

Not applicable.

**Authors' contributions**

This paper has been conceived and written by the sole author. The author read and approved the final manuscript.

**Authors' information**

Dr Thomas Lancaster is a Senior Teaching Fellow in Computing at Imperial College London. His main research areas relate to academic integrity and contract cheating.

**Funding**

No funding was received for this study.

**Availability of data and materials**

All the relevant data is included within the paper.

**Declarations****Competing interests**

The author has no competing interests.

Received: 19 December 2022 Accepted: 28 March 2023

Published online: 05 June 2023

**References**

- Aaronson S (2022) My AI Safety Lecture for UT Effective Altruism <https://scottaaronson.blog/?p=6823>
- Altman, S (2022) [@sama] ChatGPT launched on wednesday. today it crossed 1 million users! [Tweet] <https://twitter.com/sama/status/1599668808285028353>
- Amigud A, Lancaster T (2019) 246 reasons to cheat: An analysis of students' reasons for seeking to outsource academic work. *Computers & Education* 134:98–107. <https://doi.org/10.1016/j.compedu.2019.01.017>
- Clarke R, Lancaster T (2006) Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites. *Proceedings of 2nd International Plagiarism Conference*. Newcastle, United Kingdom
- Lancaster T (2020) Academic Discipline Integration by Contract Cheating Services and Essay Mills. *J Acad Ethics* 18(2):115–127. <https://doi.org/10.1007/s10805-019-09357-x>
- Lancaster T, Clarke R (2009) Automated essay spinning—an initial investigation. *Proceeding of 10th Annual Conference of the Subject Centre for Information and Computer Sciences*
- Owings S, Nelson J (2014) The essay industry. *Mt Plains J Bus Econ* 15:1–21
- Roe J, Perkins M (2022) What are Automated Paraphrasing Tools and how do we address them? A review of a growing threat to academic integrity. *Int J Educ Integr* 18:15. <https://doi.org/10.1007/s40979-022-00109-w>
- Rogerson A (2017) Detecting contract cheating in essay and report submissions: process, patterns, clues and conversations. *Int J Educ Integr* 13:10. <https://doi.org/10.1007/s40979-017-0021-6>
- Rogerson A, McCarthy G (2017) Using internet based paraphrasing tools: original work, patchwriting or facilitated plagiarism? *Int J Educ Integr* 13(2):1–15 <https://doi.org/10.1007/s40979-016-0013-y>
- Sharples M (2022) Automated Essay Writing: An AIED Opinion. *Int J Artif Intell Educ* 32:1119–1126. <https://doi.org/10.1007/s40593-022-00300-7>
- Walker C (2020) *The white-collar hustle: academic writing & the Kenyan digital labour economy*. University of Oxford, Oxford
- Yorke J, Sefcik L, Veeran-Colton T (2020) Contract cheating and blackmail: a risky business? *Stud High Educ* 47(1):53–66. <https://doi.org/10.1080/03075079.2020.1730313>

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.