Survey on AI-Generated Plagiarism Detection: The Impact of Large Language Models on Academic Integrity

Article *in* Journal of Academic Ethics · November 2024

DOI: 10.1007/s10805-024-09576-x

CITATIONS
21

READS
998

4 authors, including:

Luis Miralles Pechuan
Technological University Dublin - Grangegorman
77 PUBLICATIONS **725** CITATIONS
SEE PROFILE

David Lillis
University College Dublin
99 PUBLICATIONS **1,052** CITATIONS
SEE PROFILE

Journal of Academic Ethics
https://doi.org/10.1007/s10805-024-09576-x

Survey on AI-Generated Plagiarism Detection: The Impact of Large Language Models on Academic Integrity

Shushanta Pudasaini[1] · Luis Miralles-Pechuán[1] · David Lillis[2] · Marisa Llorens Salvador[1]

Accepted: 19 September 2024

Abstract
A survey conducted in 2023 surveyed 3,017 high school and college students. It found that almost one-third of them confessed to using ChatGPT for assistance with their homework. The rise of Large Language Models (LLMs) such as ChatGPT and Gemini has led to a surge in academic misconduct. Students can now complete their assignments and exams just by asking an LLM for solutions to the given problem, without putting in the effort required for learning. And, what is more worrying, educators do not have the proper tools to detect it. The more advanced AI tools become, the more human-like text they generate, and the more difficult they are to detect. Additionally, some educators find it difficult to adapt their teaching and assessment methods to avoid plagiarism. This paper is focused on how LLMs and AI-Generated Content (AIGC) have affected education. It first shows the relationship between LLMs and academic dishonesty. Then, it reviews state-of-the-art solutions for preventing academic plagiarism in detail, including a survey of the main datasets, algorithms, tools, and evasion strategies for plagiarism detection. Lastly, it identifies gaps in existing solutions and presents potential long-term solutions based on AI tools and educational approaches to address plagiarism in an ever-changing world.

Keywords Artificial intelligence generated content · Large language models · Academic cheating · Plagiarism

✉ Shushanta Pudasaini
D23129142@mytudublin.ie

Luis Miralles-Pechuán
luis.miralles@TUDublin.ie

David Lillis
david.lillis@ucd.ie

Marisa Llorens Salvador
marisa.llorens@TUDublin.ie

1 Technological University Dublin, Dublin, Ireland

2 University College Dublin, Dublin, Ireland

Published online: 04 November 2024

Introduction

ChatGPT is a revolutionary conversational engine based on Large Language Models (LLMs). Since it was released on November 30, 2022, ChatGPT has significantly impacted many industries and domains (Kalla & Smith, 2023), including academia (Sohail et al., 2023). The capabilities of highly advanced LLMs have affected the academic world in various ways. For example, students are using ChatGPT to complete their homework assignments (Tossell et al., 2024) and to pass challenging online exams like the Graduate Record Examination (GRE) and Scholastic Assessment Test (SAT) (Varanasi, 2023).

A large survey conducted over 12 years at 24 universities in the US showed that almost 95% of the students admitted committing plagiarism at least once in their academic studies(ArgaAssociation, 2019). Such academic misconduct has raised concerns about the current evaluation systems used in educational institutions. Lecturers in universities are struggling to detect students' misconduct related to plagiarism (Gaurdian, 2023). For example, according to a survey by Intelligent.com conducted in May 2023, among 3,017 high school and college students, nearly one-third of the students admitted to using ChatGPT to complete their homework (Westfall, 2023). The dependency of students on such tools hampers their creativity and learning skills (Smolansky et al., 2023). Due to such threats, several universities have decided to ban the usage of ChatGPT while submitting assignments (Nolan, 2023).

Apart from students, researchers have also been misusing this technology. Anyone with limited knowledge and research experience can use ChatGPT to write advanced academic content. ChatGPT has reached high-impact journals. For example, a recent publication in Nature states that at least four preprints were submitted to their journal including ChatGPT as a co-author (Stokel-Walker, 2023b). There have also been a few academic papers in notable journals that included some ChatGPT prompts by mistake (Bin-Nashwan et al., 2023) which shows how much LLMs are used. The disruption created by this phenomenon in scientific publishing has been such that journals have strictly banned listing ChatGPT as a co-author (Ian, 2023). This fact may be explained as ChatGPT may present flawed and fabricated research and set out new authorship guidelines for AI-generated text (Harker, 2023), also ChatGPT does not have legal responsibility in case of mistake (Bin-Nashwan et al., 2023). Due to such problems, LLMs have been a major disruptive factor in academia.

Traditionally, plagiarism was mostly done by presenting a document that included paragraphs from other sources or authors without being referenced to. However, with the emergence of LLMs, students can now generate text and complete their assignments entirely by using prompts such as "Generate an essay about the occupation of the Roman Empire in Ireland". In this study, the act of using text LLMs-generated and claiming it as their work is referred to as AI-generated plagiarism. This research explores the impact of LLMs on academia and how to detect AI-generated plagiarism.

Since the introduction of ChatGPT, research has been conducted on both sides: developing more intelligent LLMs and developing models that can detect such AI-generated content. However, the detection always goes one step behind. The text generated by AI-based algorithms has been coined as Artificial Intelligence Generated Content (AIGC) (Liu et al., 2023). From the 94 million training parameter ELMO model released in 2019 to the 1.76 trillion GPT-4 model released in 2023, the LLMs' size and capabilities have been increasing very quickly (Xi et al., 2023). The projected rapid development of highly capable LLMs and the quality of their outputs implies that AIGC will be increasingly difficult to detect in the coming years (Zhao et al., 2023).

Many tools have been developed to detect AIGC such as DetectGPT (Mitchell et al., 2023), RADAR (Hu et al., 2023), Ghostbuster (Verma et al., 2023), GPT-Sentinel (Chen et al., 2023). Also, OpenAI, the creator of ChatGPT, introduced its AIGC detection tool two months after its release in 2020. However, OpenAI states that the detector is not fully reliable (Kirchner, 2023a). Similarly, several AIGC detector tools and software such as CopyLeaks, Turnitin, GPTZero, and Crossplag have been released for the general use of the public to identify AI-generated content. On the other hand, different techniques to attack or evade such AIGC detectors have also been developed and are an active area of research (Cai & Cui, 2023). Evasion techniques such as prompt engineering i.e., optimizing the prompt in LLMs to get required results (Lu et al., 2023), recursive paraphrasing i.e., paraphrasing multiple times (Krishna et al., 2024), authorship obfuscation (Macko et al., 2024), and sentence or word substitution have been developed to point out the failures in the AIGC detector tools.

The main contribution of this research is a comprehensive survey of existing algorithms, tools, datasets, and evasion strategies developed for addressing academic misconduct, particularly in plagiarism detection and AIGC detection. The paper addresses the growing problem of academic misconduct due to the rise of LLMs like ChatGPT, which students and researchers increasingly use to generate content that is difficult to detect as plagiarized. It identifies the increasing issues of academic misconduct since the emergence of LLMs and provides a comprehensive review of both technical and non-technical solutions for AIGC detection. Through a simple experiment, the paper demonstrates the unreliability of existing tools in detecting misconduct when evasion techniques are used. Based on this experiment and a survey, the paper discusses the reliability and feasibility of addressing the problem, explores various educational solutions to combat plagiarism, and examines the adoption, ethical implications, and trustworthiness of LLMs in academia.

This paper is organized as follows. Section "Large Language Models and Academic Misconduct" identifies different problems due to academic dishonesty with the emergence of LLMs and shows how such LLMs have affected academia by presenting the different ways such tools are making academic plagiarism easier. A survey of existing algorithms, datasets, and tools for detecting academic cheating methods such as plagiarism and generating text with AI is performed in Section "Detecting AI-Generated Plagiarism". Different evasion techniques have also been discussed in Section "Detecting AI-Generated Plagiarism". Based on the detection algorithms and evasion strategies survey, the limitations and gaps observed in current solutions have been discussed in Section "Limitations and Gaps in Current Solutions". Finally, the feasibility of technical solutions for AIGC detection and other alternative educational solutions have been discussed in Section "Discussion".

Large Language Models and Academic Misconduct

This section discusses the impact of LLMs in academia, how they have been utilized for academic misconduct, and their proliferation, resulting in increasingly human-like text that seriously threatens academic integrity.

Rise and Evolution of Large Language Models

Language Modelling (LM) began in the 1990s with statistical learning methods by building word prediction models called N-gram language models based on Markov assumption which assumes the future state of a process depends entirely on its current state (Kuhn et al., 1994).

However, these N-gram language models suffered from issues like increased computational complexity and overfitting because of high dimensional data, also called the curse of dimensionality. The focus then shifted towards Neural Language Models (NLMs), which utilized deep learning (DL) architectures like multi-layer perceptron (MLPs) (Blat et al., 2005) and recurrent neural networks (RNNs) (Chelba et al., 2017) to characterize the probability of word sequences.

After that, Pretrained Language Models (PLMs) such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018) were introduced, which could capture the context-aware representation of any text using the transformer architecture (Vaswani et al., 2017). LLMs have been developed based on pre-existing PLMs using reinforcement learning from human feedback (Li et al., 2023) on top which have completely transformed the landscape of text generation. LLMs like ChatGPT and Gemini are now readily available to the public through products from tech giants such as OpenAI and Google.

The growth of LLMs in size and capabilities at a staggering pace has been reshaping the AI landscape over the last decade (Simon, 2024). This is evident in the timeline of various GPT models released by OpenAI in recent years, as depicted in Fig. 1. There is currently, a race to develop LLMs with superior capacities with a higher number of training parameters going on between big tech companies. With OpenAI releasing its series of models under the GPT family as shown in Fig. 1, Meta has also been releasing several LLMs such as Alpaca, Mistral, and Vicuna under the LLaMA family tree (Zhao et al., 2023). Similarly, the release of LLMs such as PaLM, Gemini, and Gemma from Google, Claude models from Anthropic, and the Command R model from Cohere shows that the development of powerful LLMs is a big race.

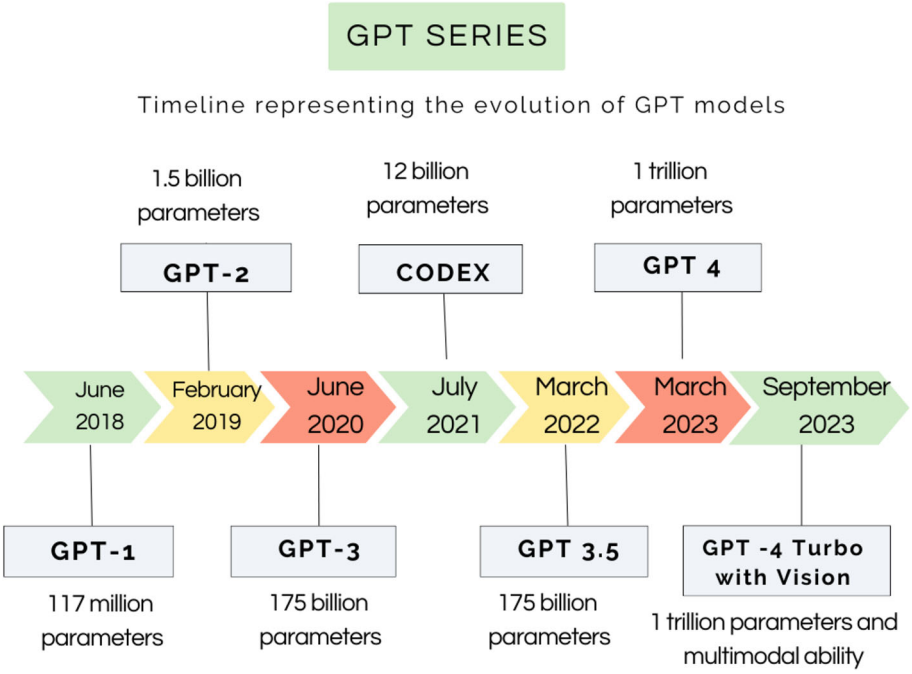

Fig. 1 Timeline indicating the release date and parameter of different GPT models by OpenAI (Kalyan, 2023)

Open-source platforms like HuggingFace are also in the race to bring out more powerful and intelligent LLMs. Currently, HuggingFace has 535,131 models, over 250,000 datasets, and more than 250,000 spaces uploaded on their platform (Gillham, 2024). This indicates a potential scenario where LLMs will continue to improve over time and can generate more human-like text in the future, making AIGC detection more challenging.

In recent years, DL models have been capable of generating different types of data. For example, DALL-E 3 from OpenAI can generate images, while models like ChatGPT, Gemini, and Perplexity can generate text. AudioGen and MusicGen from MetaAI can generate audio, and GPT-4 from OpenAI can generate multimodal data, meaning it can handle different types of data. The language models require simple, optimized prompts so that users can get the desired data. These models can generate human-like coherent and diverse texts (Pallagani et al., 2023).

With the rise of LLMs, the branch of AI able to generate new content called Generative AI is growing on a large scale. For example, the market size of generative AI was reported to be 44 billion USD in 2023 and is projected to reach 66.62 billion USD in 2024 (Chui et al., 2023). Also, OpenAI, one of the leading companies in this field, reported in 2023 that 100 million people use ChatGPT every week and over 2 million developers are building generative AI applications from the API service provided by the company (Porter, 2023).

LLMs offer remarkable capabilities such as brainstorming, generating counterarguments, creating summaries and abstracts, and correcting grammar. LLMs like ChatGPT can rephrase text and produce text which is claimed to be almost indistinguishable from human writing (Chui, 2024). These models can handle simple tasks such as airthmetic calculations as well as complex tasks like writing a research paper on a challenging problem (Stokel-Walker, 2023a) and coding tasks. Rehan Haque, founding director of Metatalent.ai, suggests that we have reached a point where entire programming projects can be done using LLMs and then another chain of AI tools can be used to make the project AI undetectable (Whalen et al., 2023).

Academic Misconduct

According to a survey in the USA conducted in 1986, 82% of undergraduate students admitted to some form of misconduct while submitting their assignments (Stern & Havlicek, 2024). Academic misconduct refers to actions that violate the originality of academic work, such as ghostwriting, plagiarism, data fabrication, deceit, and generation using Artificial Intelligence (AI). Among these, generation using AI is the most frequent and recent form of misconduct (Nerdynav, 2024).

There are commercial plagiarism detection tools that are very efficient but cannot detect AIGC. Along with LLM's capability to generate human-like text and their widespread accessibility, several forms of academic misconduct have been growing (Huallpa et al., 2023). With the rise in LLMs, plagiarism detection has also been more difficult. One can use LLMs to generate new text as well as remove plagiarism in a given text. Lee et al. (2024) explored the duality of LLMs, presenting how LLMs can be used for both plagiarism detection and generation, highlighting that LLMs can surpass current commercial plagiarism detection tools. In recent years, a lot of research has been done on advancing LLMs, plagiarism detection, and LLM-generated text detection. The bibliometric data representation on "Plagiarism Detection" and "Large Language Models" is represented in Fig. 2 indicating a huge impact of LLMs in academia.

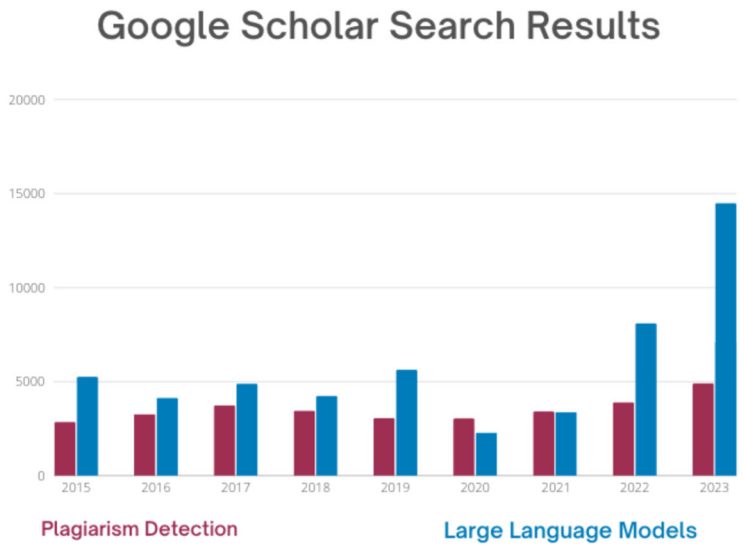

Fig. 2 Bar chart of the bibliometrics data related to plagiarism detection and LLM generated text detection

LLMs have led to a significant amount of detected AIGC on the internet. Some estimates suggest that in 2024, more than 10% of internet content is already AI-generated (Originality.AI, 2024). AIGC presents various threats and challenges, including ethical concerns, harmful or inappropriate content, bias, over-reliance, misuse, digital divide, academic misconduct, security, and privacy (Nah et al., 2023). However, the scope of this research focuses on AI-generated textual content, its impacts on academia, solutions developed to detect such AIGC, the reliability of those solutions, and alternative solutions to manage those impacts on academia.

Detecting AI-Generated Plagiarism

In this section, we review the existing technical solutions for academic misconduct. First, we discuss plagiarism and its typology, then discuss the data, algorithms, and tools developed for plagiarism detection along with the evasion techniques to fool such AIGC detectors. Then, we discuss how LLMs have impacted plagiarism using a simple experiment.

Types of Plagiarism

The American Historical Association (AHA) formally defined plagiarism in 1987 as failure to acknowledge the work of another (Eisner & Vicinus, 2008). Plagiarism, a part of academic cheating, is much more widespread than usually recognized according to a few studies (Martin, 1992).

The typology of plagiarism may vary according to data type or level of difficulty. Foltýnek et al. (2019) presented different typologies defined in several research papers and put forward a new typology for plagiarism according to the level of obfuscation as character-preserving plagiarism, syntax-preserving plagiarism, semantics-preserving plagiarism, idea-preserving

plagiarism, and ghostwriting. The plagiarism type may also vary according to the data types, such as code plagiarism and text plagiarism.

The main types of plagiarism are shown in Fig. 3. In this research, we specifically focus on AI-Generated Plagiarism which shares characteristics with other types of plagiarism as LLMs are capable of evading traditional detection algorithms also.

Academic misconduct was a serious issue even before LLMs existed because of easy access to information and other works on the internet (Ison, 2016). Over the years, different solutions for Author Identification (Kestemont et al., 2019), and Plagiarism Detection (Patel et al., 2011) have been developed. These solutions have been used by academic institutions and publishers (Turnitin, 2024). However, these robust solutions for traditional plagiarism have not been able to perform successfully when applied to detect AI-generated plagiarism (Ed, 2023).

The introduction of LLMs has made the problem of academic misconduct even more serious because of the different use cases of LLMs to perform academic misconduct, such as completing student assignments, passing online examinations, paraphrasing texts to escape plagiarism, and generating academic content for research papers.

To tackle the academic misconduct raised due to LLMs, two major technical solutions were identified in our survey which are represented in Fig. 4 as a taxonomy. Traditional plagiarism detection refers to the detection of plagiarism types represented in Fig. 3 except AI-generated plagiarism whereas LLM-generated text detection refers to the detection of AI-generated plagiarism. The LLM-generated text detection solutions are further categorized into watermarking-based, zero-shot-based, and training classifier-based approaches.

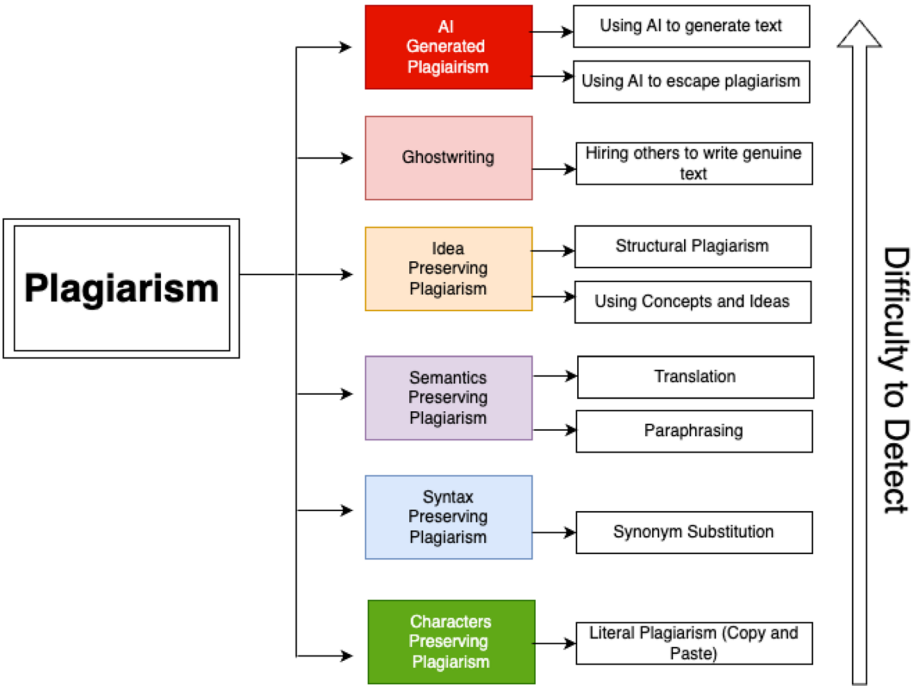

Fig. 3 Diagram representing different types of plagiarism along with the difficulty level in detecting the type of plagiarism

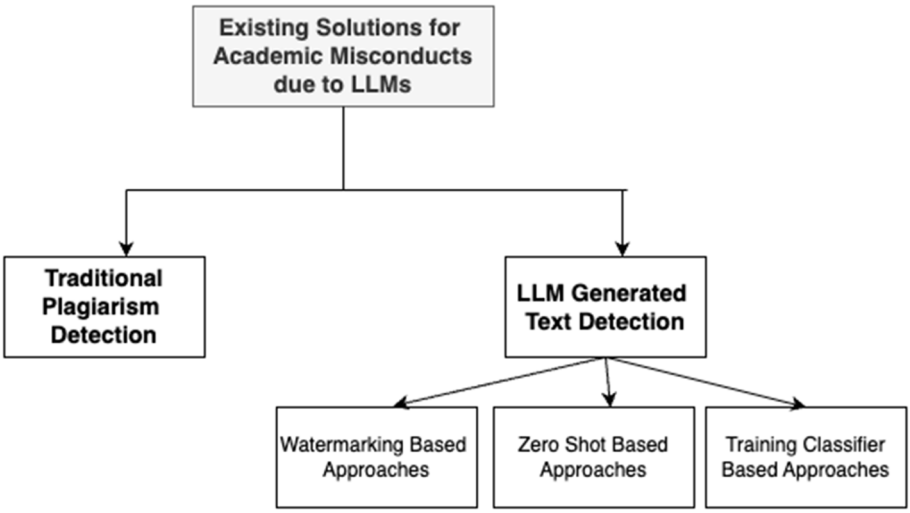

Fig. 4 Taxonomy of the survey done on existing solutions for academic misconduct due to LLMs

Traditional Plagiarism Detection

Plagiarism detection involves both the identification and the prevention of plagiarism through automated systems (Guillén-Nieto, 2022). Plagiarism detection may be categorized based on different factors. Such as the number of languages used, plagiarism detection may be monolingual or cross-lingual (Alzahrani et al., 2012). Likewise, plagiarism detection may be extrinsic or intrinsic. If plagiarism is detected only using the text itself it is termed intrinsic plagiarism detection. In contrast, if plagiarism is detected compared to other text, it is termed extrinsic plagiarism detection (AlSallal et al., 2019).

Plagiarism detection can also be categorised according to their approach. N-gram-based, vector-based, syntax-based, semantic-based, fuzzy-based, structural-based, and stylometric-based (Khaled & Al-Tamimi, 2021). The amount of text on the internet is growing due to the LLMs, and the concern of plagiarism detection is even more serious. Many research conferences and workshops have been held to solve plagiarism detection (Eriksson & Karlgren, 2012; Stein et al., 2011; Potthast et al., 2009) across Europe and other continents (Chaika et al., 2023) such as Plagiarism Analysis, Authorship Identification, and Near-duplicate Detection (PAN), and International Conference Plagiarism.

A typical plagiarism detection algorithm involves feature engineering, classification models, and text-matching similarity metrics. The most common features used in plagiarism detection algorithms are frequency of characters, average word length, average sentence length, Word N-grams frequency, part of speech, synonyms, and hypernyms (Chitra & Rajkumar, 2016). Plagiarism is mainly evaluated based on textual similarity with other reference textual contents. To calculate such similarity, researchers most commonly used Hamming distance, Levenshtein distance, and Longest common subsequence distance as string similarity metrics. They frequently employed the Jaccard coefficient, Cosine coefficient, Manhattan distance, Euclidean distance, Matching coefficient, and Dice coefficient as vector similarity metrics.

Table 1 shows the different algorithms developed along with the datasets, models, and evaluation metrics used for traditional plagiarism detection.

Table 1 Summary of different open source datasets and algorithms developed to solve plagiarism detection along with the paper title, model used, and metrics

Dataset	Paper title	Model used	Metrics
PAN dataset	A New online plagiarism detection system based on DL (El Mostafa Hambi & Benabbou, 2020)	Doc2Vec, SLSTM, CNN	Accuracy
	Exploration of fuzzy C means clustering algorithm in external plagiarism detection system (Ravi et al., 2016)	Fuzzy C means clustering algorithm	Precision, Recall
MRPC dataset	Plagiarism detection using machine learning-based paraphrase recognizer (Chitra & Rajkumar, 2016)	Feature extraction, SVM	Accuracy
	A population-based plagiarism detection using DistilBERT-generated word embedding (Yuqin & Liu, 2023)	DistilBERT, LSTM, Clustering	Precision, Recall, F1, G-means
Custom dataset	Will ChatGPT get you caught? Rethinking of plagiarism detection (Khalil & Er, 2023)	ChatGPT	False negatives
Arabic plagiarism dataset	Arabic plagiarism detection using word correlation in N-Grams with K-Overlapping approach (Alzahrani, 2015)	N gram similarity matching	Recall, Precision, Granularity, Plagdet

Along with such research studies, several plagiarism detection tools have been developed and made available online. For example, Abdelhamid et al. (2022) compared the performance of eight different plagiarism detectors online on different levels of plagiarism in the text. Turnitin, iThenticate, CopyLeaks, Duplichecker, Grammarly, PlagScan, and Quillbot are some online tools to detect plagiarism. Among plagiarism detectors online, Turnitin is the tool mostly favoured by academic institutions (Mphahlele & McKenna, 2019).

A major problem with protecting originality in academic settings is using evasion techniques to avoid existing plagiarism detection tools. For instance, Elkhatat et al. (2021) experimented with four of the most commonly used evasion techniques applied to top plagiarism detectors available online and found that they could not detect plagiarised text created using simple evasion techniques. Table 2 presents the evasion techniques that students mostly use to fool existing online plagiarism detection tools, along with their short description and the result of checking those tools through every technique.

Traditional Plagiarism Detection in the LLMs Era

Plagiarism detection has shifted its focus since the rise of LLMs. LLMs can be used as a tool to test plagiarism scores in a text and also as a tool to paraphrase textual content allowing

Table 2 Summary of evasion techniques for plagiarism detection and result while testing on plagiarism detectors from the experiment done by Elkhatat et al. (2021)

Evasion technique	Description	Successful Plag. Detectors	Unsuccessful Plag. Detectors
Imaged texts	Convert all text to image and export files as PDF	None	Turnitin, iThenticate, Copyscape, PlagAware, StrikePlagiarism.com, Unicheck, Check-For-Plag, Blackboard-SafeAssign
Quoted	Insert invisible white quotation marks for all paragraphs	Blackboard-SafeAssign, Copyscape, PlagAware, StrikePlagiarism.com, Unicheck, Check-For-Plag	Turnitin, iThenticate, PlagScan
Letter-like symbols	Replace most common letters like a, e, o with their Latin small letter	Turnitin, iThenticate, PlagAware	Blackboard-SafeAssign, Copyscape, PlagScan, Unicheck, Check-For-Plag
Invisible letters	Replace spaces within words with any white color letter	StrikePlagiarism.com	Turnitin, iThenticate, Copyscape, PlagAware, Unicheck, Check-For-Plag, Blackboard-SafeAssign

students to evade plagiarism detection even more easily. This duality helps students estimate the likelihood of getting caught and manipulate the copied text to fool plagiarism detection tools.

Biörck and Eriksson (2023) performed four experiments using prompt engineering to determine the efficiency of ChatGPT as a plagiarism testing tool and found the fourth prompt to work very well. In another experiment, Khalil and Er (2023) evaluated 50 essays generated by ChatGPT and found that ChatGPT demonstrated superior performance in plagiarism detection compared to traditional tools like iThenticate.

Combined use of both ChatGPT for text generation and paraphrasing tools such as Quillbot to paraphrase the generated text can easily fool existing plagiarism and AIGC detection tools.

Instead of copying the content written by others and paraphrasing the content using several tools to fool plagiarism detectors, students can generate textual content from ChatGPT and apply paraphrasing to make the content difficult to detect by AIGC plagiarism detection tools. To demonstrate this, we simulated a simple experiment in which a student was assigned to write an article on quantum computing, including its use cases.

We generated the essay using ChatGPT, paraphrased the generated text using QuillBot, and tested its originality. Similarly, we copied an article from an online blog from Investopedia and copied that content to ChatGPT to paraphrase and again tested the originality of the text. The process of the experimentation is shown in Fig. 5.

The testing used the plagiarism detection tool Turnitin and the AIGC detection tools GPTZero and DupliChecker. The output text was identified as being 100% unique by the PlagScan tool, which shows that students can easily complete their assignment using such a combination of tools, and lecturers will not be able to find out. The steps of the sample are shown in shown in Fig. 6.

AI Generated Content Detection

The major objective of AIGC Detection is identifying if a given piece of text has been generated using an AI system or written by a human. It is very difficult even for a human to differentiate because of the high capability of LLMs nowadays to produce more and more human-like text. In an experiment, English instructors hired to differentiate AI-generated and human-written essays were only able to get 67% accuracy (Liu et al., 2023). Similarly, in another experiment, six undergraduate and PhD students were asked to distinguish between 50 documents as AI-generated or human-written. They could only achieve 59% accuracy (Verma et al., 2023). Thus, it is difficult even for humans to differentiate AI-generated text from human-written text.

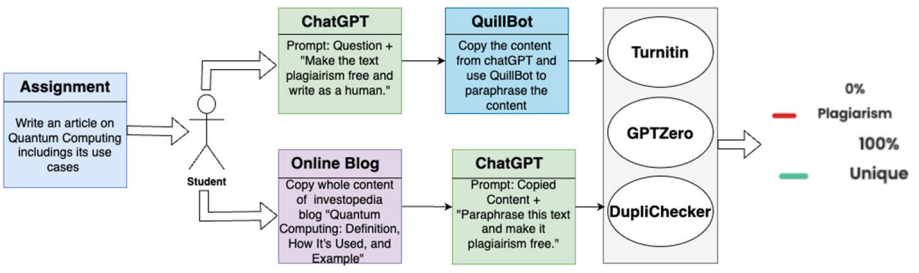

Fig. 5 Diagram representing how ChatGPT and paraphrasing tools can be used to complete assignments without being detected

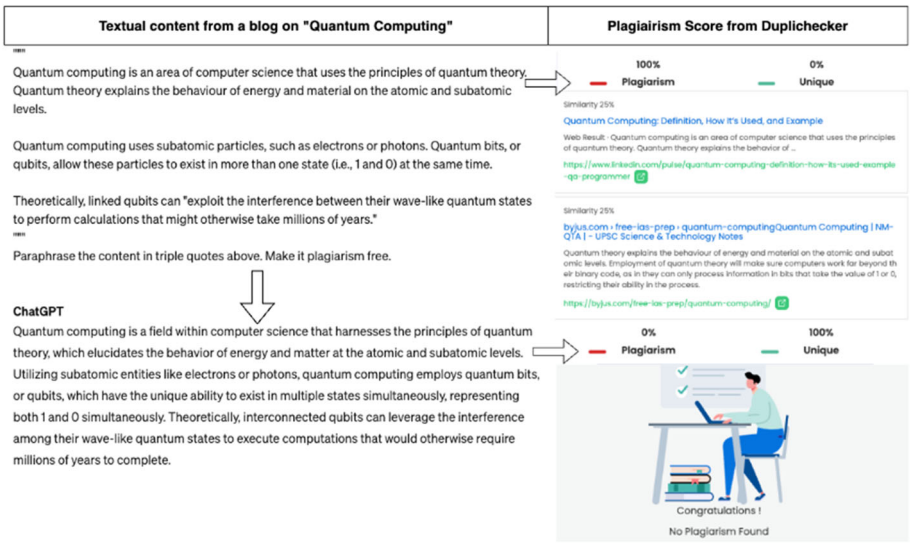

Fig. 6 Diagram representing the data flow for using ChatGPT for paraphrasing to fool plagiarism detection tool Duplichecker. The text taken from a blog is paraphrased using ChatGPT and plagiarism is tested. The Duplichecker tool shows no plagiarism

Shijaku and Canhasi (2023) found that distinguishing text generated by language models from human-written text is even more challenging than differentiating between texts generated by different language models. Additionally, Herbold et al. (2023) found that LLMs produced higher-quality argumentative essays than those written by German high school students in an online forum. Consequently, detecting AIGC is a highly complex task.

According to a quantitative analysis performed between human essays and ChatGPT written essays, human essays have more spelling and grammar errors and personal experiences. In contrast, machine essays have more similar examples and repetitive expressions (Liu et al., 2023). The machine-generated text has a more complex syntactic structure and uses more normalization, whereas human essays tend to be more lexically complex (Liu et al., 2023; Herbold et al., 2023).

Similarly, human-written text tends to have higher perplexity than AI-generated text (Liao et al., 2023). To demonstrate such linguistic differences in human-written and AI-generated text, a human was told to write a paragraph on “cow”, and the prompt “Write a paragraph on Cow” was injected in ChatGPT. The grammatical mistakes and use of personal experience can be seen in the human-written text. In contrast, syntactically complex sentence structure and repetition of words can be seen in the text generated from ChatGPT, shown in Fig. 7.

Open Source Datasets for AIGC Detection

As the term “Garbage In, Garbage Out” in machine learning (ML) represents, high-quality data is vital for efficient ML research (Geiger et al., 2020). AIGC detection, when modelled as a straightforward binary classification problem, requires a high-quality, balanced labelled dataset comprising a wide variety of human-written and AI-generated texts. However, when AIGC detection is solved from other approaches, such as zero-shot detection methods, a dataset may not be required for training (Xian et al., 2017). Building a high-quality dataset

Written by Human	AI Generated Content
The cow is the national animal of the country called Nepal. The people of Nepal worship cows, but in Western countries, they farm the cows for eating purposes which if we do in Nepal we get a jail sentence. Cows are very important for people's lives because they feed on unnecessary grass and they provide us with milk which we create various products like cheese, yoghurt and everyone's favourite ice cream.	The cow, revered in many cultures worldwide, is a gentle and iconic creature, known for its placid demeanor and vital contributions to agriculture and society. Belonging to the Bovidae family, cows are domesticated mammals primarily raised for their milk, meat, and leather. Their large, sturdy bodies are typically adorned with a distinctive pattern of black and white markings, though variations exist across different breeds.

Fig. 7 Figure showing an example of a ChatGPT generated and human-written text

is a very laborious process that involves data scraping from various sources, data cleaning, data preprocessing, and data annotation.

The human-written part of most open-source AIGC detection datasets was built using three major techniques discussed below.

- By directly extracting from other open-source datasets such as the wikiHow text dataset (Koupaee & Wang, 2018). This technique was used to build the GPT-Sentinel dataset (Chen et al., 2023).
- Using an available corpus such as in-class or homework exercises, Test of English as a Foreign Language (TOEFL) writing tasks, and GRE writing tasks were used to build ArguGPT dataset (Liu et al., 2023).
- Manually collecting observations from online sources such as the CHEAT dataset built by searching and extracting human-written abstracts from IEEE Xplore (Yu et al., 2023).

The AI-generated text part in those open-source datasets is built using different prompting strategies applied to OpenAI models using the OpenAI API service (Chen et al., 2023; Yu et al., 2023). Some papers claim very high results in AIGC detection using specific custom datasets, however the datasets used are not open-sourced (Chen et al., 2023). The datasets used in different research papers to solve AIGC detection are presented in Table 3.

A few gaps were identified in the current survey of datasets used for AIGC detection. Firstly, there is no established benchmark dataset for this problem. Researchers have been developing custom datasets independently. As a result, the evaluation and testing of these datasets are also based on their own data, leading to impartial comparisons between different research implementations. Furthermore, the various datasets created by different researchers often contain textual data from specific domains. Additionally, these datasets typically include straightforward ChatGPT-generated content, making them susceptible to evasion techniques like paraphrasing.

To address these issues, developing a benchmark dataset that encompasses a diverse range of fields and includes modified ChatGPT responses is crucial. This would ensure a more comprehensive evaluation of AIGC detection methods. Solutions for AIGC detection should then be assessed using this benchmark dataset to ensure consistency and fairness in performance comparisons.

Table 3 Summary of different open-source datasets developed for AIGC detection along with the source of the human-written data part, number of instances, and publisher

Dataset	Number of instances	Human written text source	Publisher
ArguGPT (Liu et al., 2023)	4,708	Student essays from TOEFEL	HuggingFace
Custom Dataset	252	Student essays from TOEFEL	Github
Human ChatGPT comparison corpus (HC3) (Guo et al., 2023)	24,300	Open-domain, financial, medical, legal, and psychological areas	HuggingFace
CHEAT (Yu et al., 2023)	35,304	Abstracts from papers	PaperswithCode
DagPap22 (Kashnitsky, 2022)	26,637	Elsevier papers	Github
Ghostbuster Dataset (Verma et al., 2023)	12,500	Subreddit posts	Github
M4GT-Bench Dataset (Wang et al., 2024)	138,465	Wikipedia, Reddit, Arxiv, and PeerRead	Arxiv
M4 Dataset (Wang et al., 2024)	147,895	Wikipedia, Reddit, Arxiv, PeerRead, Urdu-news, and RuATD	Github
GPT-Sentinel (Chen et al., 2023)	3,152,979	OpenGPTText, OpenWebText and ChatGPT	Arxiv

Watermarking Based Approaches

Watermarking-based approaches embed subtle signatures, such as a cryptographic pseudorandom function, into LLM-generated text. These signatures can later be decrypted to verify whether the text was produced by a specific LLM. In text generation models like LLMs, both input and output are always in the form of tokens. These models generate a probability distribution for the next predicted token, based on the preceding sequence of tokens. Typically, the next token is sampled randomly from this distribution according to a parameter called temperature.

To introduce watermarking, instead of selecting the next token purely at random, a cryptographic pseudorandom function can be used. This function subtly influences the token selection process in a way that is imperceptible to end users but allows for subsequent decryption to determine if a specific LLM generated the text.

The process of watermarking involves the following steps:

1. **Token Probability Distribution**: The LLM generates a probability distribution over the potential next tokens based on the previous token sequence.
2. **Cryptographic Pseudorandom Function**: Instead of randomly selecting the next token, this function is applied to subtly alter the probability distribution.
3. **Token Selection**: The next token is then selected based on the modified distribution.
4. **Decryption for Verification**: To verify if a text was generated by the LLM, the embedded signature can be decrypted, revealing whether the text originated from the particular LLM.

Watermarking in natural language has been employed for purposes such as information hiding even before the release of ChatGPT (Topkara et al., 2006). Techniques like morphosyntactic alterations and synonym substitution have been used for watermarking natural language (Meral et al., 2009; Hao et al., 2018). At a workshop on LLMs and Transformers, Scott Aaronson revealed his proposal for a watermarking scheme based on the "Gumbel Softmax Rule." This scheme was later implemented by Hendrick Kirchner, a scientist from OpenAI, and was effective even with a few hundred tokens (Aaronson, 2022).

Traditional watermarking techniques face the challenge of preserving semantic meaning between the original and watermarked texts, often resulting in significantly different meanings. To address this, Yang et al. (2022) introduced context-aware lexical substitution for watermarking, achieving a Z-score of 2.19, which indicates the watermark strength. Higher Z-scores reflect greater deviation of watermarked text features from the mean values found in human-written text in a controlled manner. Additionally, Abdelnabi and Fritz (2021) introduced the first end-to-end Adversarial Watermarking Transformer (AWT) model just a month after the release of GPT-2, achieving a Z-score of 2.73. Further advancements were made by Yang et al. (2023), who developed a framework for watermarking black-box language models, achieving a Z-score of 3.63.

Kirchenbauer et al. (2023) proposed a watermarking technique in June 2023, which reduced the False Positive Rate (FPR) to 0 and could be implemented without prior knowledge of any LLM parameters or its Application Programmable Interface (API). However, in October 2023, Zhao et al. (2023) demonstrated that this watermarking technique failed under paraphrasing attacks using models such as ChatGPT, DIPPER-1, DIPPER-2, and BART. Zhao et al. (2023) also proposed a new watermarking solution called Unigram-Watermark, which outperformed Kirchenbauer et al. (2023) method in resisting paraphrasing attacks.

Despite these advancements, watermarking techniques continued to struggle with balancing detectability and maintaining the semantic integrity of the generated text (Huo et al., 2024). To address this, Huo et al. (2024) introduced a multiobjective optimization (MOO)

approach for watermarking, which showed improved performance and robustness against copy-paste and paraphrasing attacks.

A new watermarking framework, called WaterMax, was introduced by Giboulot and Teddy (2024). This framework achieves high detectability while leaving the LLM untouched, thereby maintaining the quality of the generated text. WaterMax represents a significant advancement, outperforming previous watermarking techniques. A summary of different research papers on AIGC detection based on watermarking techniques is presented in Table 4.

Zero-shot Based Approaches

Zero-shot approaches involve pretrained neural network models predicting unseen classes without specific training for those classes (Davison, 2020). Unlike traditional text classification models that rely on large sets of labelled datasets for training, zero-shot learning leverages the ability of pretrained language models to classify new text observations even in entirely new datasets (Pushp & Srivastava, 2017).

Giant Language Model Test Room (GLTR) is a pioneering tool in AIGC detection based on zero-shot learning. Developed by Gehrmann et al. (2019), GLTR detects whether the text is machine-generated by providing a visual footprint of the text, thereby supporting its predictions. Mireshghallah et al. (2023) conducted experiments to determine if other language models could detect machine-generated text from different models. They discovered that smaller language models like OPT-125M were more effective AIGC detectors than larger models such as GPTJ-6B.

In the AIGC detection domain, DNA-GPT was introduced and claimed to outperform the OpenAI text detector on four English and German datasets (Yang et al., 2023). The core concept of DNA-GPT is to compare the probability divergence between actual tokens and generated tokens.

Detect-GPT marked another significant advancement in AIGC detection, increasing the state-of-the-art zero-shot detection AUROC from 0.81 to 0.95 (Mitchell et al., 2023). Detect-GPT is based on the principle that machine-generated text generally resides in regions of negative curvature in the log probability generated by the model (Mitchell et al., 2023). Recently, Fast-DetectGPT further improved the AUROC to 0.98 and achieved a detection speed 340 times faster than DetectGPT (Bao et al., 2023). Fast-DetectGPT introduced the concept of conditional probability curvature to highlight the vocabulary selection differences between machine-generated and human-written text (Bao et al., 2023). Table 5 presents a summary of research papers on AIGC detection based on zero-shot learning, along with their authors, publication years, and evaluation metrics.

Training Classifier Based Approaches

AIGC detection can be framed as a binary text classification problem, where the task is to categorize text as either AI-generated or human-written. Following the release of ChatGPT in November 2022, OpenAI responded by developing the OpenAI Text Classifier, which was introduced in January 2023 (Wiggers, 2023).

Liu et al. (2023) proposed ArguGPT, a classification model trained on sentence-level and essay-level data from TOEFL and GRE tasks. This model achieved an accuracy of 90%. Oghaz et al. (2023) explored a range of machine learning algorithms, including Multinomial Naive Bayes, Random Forest, Support Vector Machines (SVM), and K-nearest Neighbors

Table 4 Summary of AIGC Detection Research Papers Based on Watermarking along with the metrics used to evaluate the approach

Paper	Authors	Year	Evaluation metrics
Watermarking Text Generated by Black-Box Language Models (Yang et al., 2023)	Xi Yang et al.	2023	ROC, AUC, Semantic similarity, METEOR score
Tracing Text Provenance via Context-Aware Lexical Substitution (Yang et al., 2022)	Xi Yang et al.	2021	Semantic Relatedness (SR), Semantic Similarity (SS)
Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding (Abdelnabi & Fritz, 2021)	Sahar Abdelnabi and Mario Fritz	2022	Bit accuracy, METEOR score, SBERT distance
A Watermark for Large Language Models (Kirchenbauer et al., 2023)	John Kirchenbauer et al.	2023	FPR, TNR, TPR, FNR
Provable Robust Watermarking for AI-Generated Text (Zhao et al., 2023)	Xuandong Zhao et al.	2023	TPR, F-Score
Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models (Huo et al., 2024)	Mingjia Huo et al.	2024	FPR, Semantic coherence
WaterMax: breaking the LLM watermark detectability-robustness-quality trade-off (Giboulot & Teddy, 2024)	Eva Giboulot, Teddy Furon	2024	P values, ROC, Tamper resistance

Table 5 Summary of AIGC Detection Research Papers Based on Zero-Shot Learning along with the metrics used for evaluation by the approach

Paper	Author	Year	Evaluation Metrics
GLTR: Statistical Detection and Visualization of Generated Text (Gehrmann et al., 2019)	Sebastian Gehrmann et al.	2019	AUROC
Smaller Language Models are Better Black-box Machine-Generated Text Detectors (Mireshghallah et al., 2023)	Niloofar Mireshghallah et al.	2023	AUC
DNA-GPT: Divergent N-Gram Analysis for Training-free Detection Of GPT-Generated Text (Yang et al., 2023)	Xianjun Yang	2023	AUROC, TPR
DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature (Mitchell et al., 2023)	Eric Mitchel et al.	2023	AUROC
Fast-DetectGPT: Efficient Zero-shot Detection of Machine-Generated Text via Conditional Probability Curvature (Bao et al., 2023)	Guangsheng Bao et al.	2024	AUROC, Speed
Does DETECTGPT Fully Utilize Perturbation? Bridge Selective Perturbation to Fine-tuned Contrastive Learning Detector would be Better (Mao et al., 2024)	Shengchao Liu et al.	2024	Accuracy, F1 Score
Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text (Hans et al., 2024)	Abhimanyu Hans et al.	2024	Precision, F1 Score, FPR

(KNN). They also employed deep learning techniques such as Bidirectional Long Short-Term Memory (LSTM) networks, DistilBERT, and RoBERTa. Their experiments, conducted on a custom dataset of question answers related to computer science, artificial intelligence, and cybersecurity, demonstrated that the RoBERTa-based deep learning model performed best, achieving an F-score of 0.992 and an accuracy of 0.991 (Oghaz et al., 2023).

Text classification algorithms generally utilize common preprocessing techniques, such as tokenization, stemming, lowercasing, and stopword removal. TF-IDF is frequently used for feature extraction (Shijaku & Canhasi, 2023). Shijaku and Canhasi (2023) integrated the XGBoost algorithm with TF-IDF and handcrafted features, visualizing the results with SHAP analysis to highlight the contribution of each parameter to model predictions. Katib et al. (2023) introduced the Tunicate Swarm Algorithm with Long Short-Term Memory Recurrent Neural Network (TSA-LSTMRNN) for AIGC detection. This model achieved an F-score of 93.17% and an accuracy of 93.83% on datasets containing human- and ChatGPT-generated text, respectively (Katib et al., 2023), using the CHEAT dataset (Yu et al., 2023).

Antoun et al. (2023) implemented various transformer-based algorithms, including RoBERTa, ELECTRA, CamemBERTa, and XLM-R, on a modified Human ChatGPT Comparison Corpus (HC3) dataset that includes both English and French texts. They achieved an impressive F-score of 99.86%. Similarly, GPT-Sentinel, a pioneering model utilizing transformer-based networks like RoBERTa and the Text-to-Text Transfer Transformer (T5), achieved over 97% accuracy (Chen et al., 2023). The recent Ghostbuster algorithm claims to have reached a 99.0 F1 score, surpassing previous models by 5.9 F1 points (Verma et al., 2023). A summary of research on AIGC detection through classifier training is presented in Table 6.

Detection Tools

With the advancement of algorithms for AIGC detection, several tools leveraging these state-of-the-art models have become publicly available. OpenAI initially released the OpenAI Text Classifier, but it was later discontinued due to its low accuracy rate (Kirchner, 2023b). Dreamsoft Innovations introduced GPTKit, which claimed to achieve 93% accuracy. Following this, DetectGPT was launched, accompanied by an Application Programming Interface (API) for developers. Originality.ai has positioned itself as a leading AI detection tool, claiming high accuracy across various large language models (LLMs) such as ChatGPT, GPT-4, Bard, and Claude 2. Arslan Akram evaluated six different AIGC detection tools using 11,580 samples from diverse datasets and found that originality.ai was particularly effective across the test dataset (Akram, 2023). Based on Akram's experiments, a summary of the most commonly used AIGC detection tools and their accuracy is provided in Table 7.

While many research papers on AIGC detection focus on binary classification (AI-generated vs. human-written), some online AIGC detectors provide results regarding the probability that a given text is human-written or AI-generated. We tested several AI detectors, including Copyleaks, ZeroGPT, Quillbot, Writer, GPTZero, Detecting-AI, Sapling, Undetectable, and Crossplag. All tested detectors, except for Copyleaks, offered probabilistic outputs. Quillbot further distinguished between AI-generated, AI-generated + paraphrased, human-written, and human-written + paraphrased texts. Additionally, Undetectable provided a feature to "humanize" AI-generated text, while Detecting-AI reported predictions from multiple models.

Table 6 Summary of AIGC detection research papers based on training classifier along with the datasets and models used for training

Title	Author	Year	Dataset used	Models used
ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models (Liu et al., 2023)	Yikang Liu et al.	2023	Custom Essay Dataset	SVM, Roberta
Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers (Gao et al., 2022)	Catherine A. Gao et al.	2023	Custom Dataset using abstracts of research papers	ChatGPT
Detection and Classification of ChatGPT Generated Contents Using Deep Transformer Models (Oghaz et al., 2023)	Mahdi Maktab et al.R	2023	Custom Dataset	ML Algorithms, BiLSTM, DistilBERT, Roberta
Ghostbuster: Detecting text written by large language modes (Verma et al., 2023)	Vivek Verma	2023	Custom Dataset	LLMs
LLM-Detector: Improving AI-Generated Chinese Text Detection with Open-Source LLM Instruction Tuning (Wang et al., 2024)	Rongsheng Wang et al.	2024	Custom Chinese Dataset	GPT
Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defence (Krishna et al., 2024)	Kalpesh Krishna et al.	2023	Custom Dataset	DIPPER, open source LLMs
Towards a Robust Detection of Language Model-Generated Text: Is ChatGPT that Easy to Detect? (Antoun et al., 2023)	Wissam Antoun et al.	2023	HC3 Corpus	RoBERTa, ELECTRA, CamemBERT, CamemBERTa

Table 7 Summary of tools available online for AIGC detection, including their current availability, key features, accuracy reported, and URLs. Note: Writer has no API available; all other tools do

Tool	Availability	Feature	Acc (%)	URL
GPTkit	Free up to 2048 characters	Utilizes 6 different AI-based content detection techniques for high accuracy	88	https://gptkit.ai/
GPTZero	Free for 10,000 words per month	Can detect text from ChatGPT, GPT4, Bard, Llama and other AI models	40	https://gptzero.me/
Originality	$0.01/100 words	Can submit longer text, detailed highlighting, faster scans	96	https://originality.ai/
Sapling	1 million chars at $25/month	Shows percentage of AI generation, also has extension	60	https://writer.com/ai-content-detector/
Writer	Free	Provides Software Development Kit (SDK) for developers	48	https://sapling.ai/ai-content-detector
Zylalab	$24.99 per month	Makes use of OpenAI technology	55	https://zylalabs.com

Table 8 Summary of different research papers on AIGC evasion techniques along with evasion technique applied, tested models, and results of the evasion applied

Experiment	Evasion technique applied	Tested models	Results
Hiding the Ghostwriters: An Adversarial Evaluation of AI-Generated Student Essay Detection (Peng et al., 2024)	Paraphrasing, Word substitution, Sentence substitution	ArguGPT, CheckGPT, RoBERTA-Single, RoBERTa-QA, Quality	Significant drop in accuracy and AUC on all models
Bypassing LLM Watermarks with Color-Aware Substitutions (Wu & Chandrasekaran, 2024)	Self color testing-based substitution	Watermarked LLMs	SCTS evasion technique is highly effective for different watermarking schemes
Watermark Stealing in Large Language Models (Jovanović et al., 2024)	Watermark stealing	Watermarked LLMs	For under $50 an attacker can both spoof and scrub state-of-the-art watermarking schemes with a success rate over 80%
Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defence (Krishna et al., 2024)	Paraphrasing, Word substitution, Sentence substitution	Watermarking, GPTZero, DetectGPT and OpenAI text Classifier	Paraphrasing dropped detection accuracy of DetectGPT from 70.3% to 4.6%
Large Language Models can be Guided to Evade AI-Generated Text Detection (Lu et al., 2023)	Substitution-based In-Context example Optimization method (SICO)	GPT3-D, DPT2-D, GPTZero, OpenAI-D, DetectGPT, Log_Rank	Decreased the AUC of six detectors by 0.5 on average
Language Model Detectors are easily optimized against (Nicks et al., 2023)	Reinforcement learning	OpenAI RoBERTa-Large detector	Decreased the AUC of OpenAI detector from 0.84 to 0.62
Stumbling Blocks: Stress Testing the Robustness of Machine-Generated Text Detectors Under Attacks (Wang et al., 2024)	Editing, Paraphrasing, Prompting, and Co-generating	Watermark, OpenAI Detectors, GLTR, DetectGPT	Performance drops by 35% across all attacks
Evade ChatGPT detectors via a single space (Cai & Cui, 2023)	Add a single-space space strategy	ChatGPT, GPTZero, GPT-4	Increased the evasion rate close to 100%

Techniques to Evade AIGC Detection Tools

With the development of various AIGC detection algorithms, new evasion techniques have emerged to bypass these detectors. These techniques differ based on the detection solution employed. For example, in response to watermarking-based AIGC detection systems, techniques such as watermark stealing (Jovanović et al., 2024), Self Color Testing-based Substitution (SCTBS) (Wu & Chandrasekaran, 2024), and language translation (He et al., 2024) have been introduced.

Similarly, for binary classification training-based AIGC detection models, methods like paraphrasing, word substitution, and sentence substitution are commonly used (Peng et al., 2024; Krishna et al., 2024). Additionally, other prevalent evasion techniques include prompting (Lu et al., 2023), adversarial attacks (Peng et al., 2024), single space addition (Cai & Cui, 2023), and reinforcement learning (Nicks et al., 2023).

These evasion techniques have been tested across various AIGC detection models, often resulting in a significant decrease in AUC (Area Under the Curve) and detection accuracy, as detailed in Table 8. The major events in the AI-generated text detection space are highlighted by the timeline represented in Fig. 8.

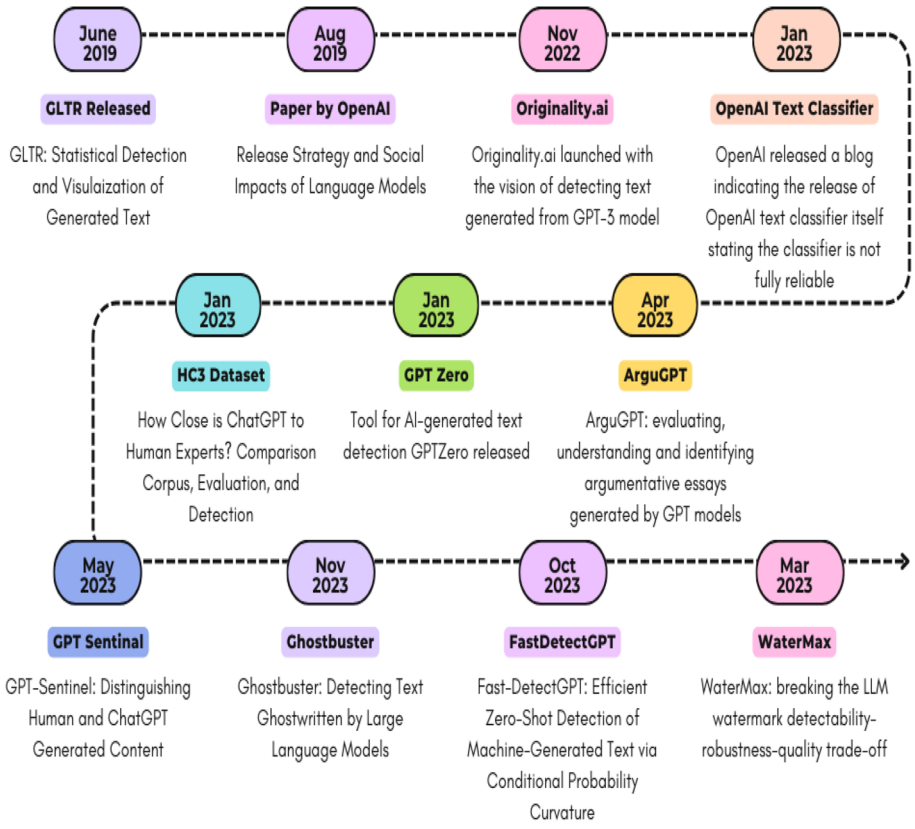

Fig. 8 Major AIGC detection events including the description of top AIGC detection datasets, algorithms, and tools

Limitations and Gaps in Current Solutions

Despite the development of several high-accuracy AIGC detection solutions on test datasets, these tools often fall short in real-world settings. Cases of false positives have been reported (Turnitin, 2023), raising concerns about the reliability of achieving a completely dependable solution. This survey has identified several gaps in existing solutions that, if addressed through further research, could lead to more reliable approaches.

Current Scenario and Reliability of AIGC Detection

Turnitin's 2023 report indicates that over 16,000 academic institutions, publishers, and corporations use Turnitin software to combat plagiarism (Turnitin, 2024). However, 97% of these institutions have not implemented formal policies regarding AI tool usage by students, and 71% of instructors have never used AI writing tools (TimeForClass, 2023). Conversely, 51% of students indicated they would continue using generative AI tools even if prohibited by their institution (TimeForClass, 2023). A survey of 1,147 instructors revealed that while many permit the use of generative AI tools for brainstorming, editing, and structuring assignments, they generally oppose their use for writing significant portions or entire assignments (TimeForClass, 2023).

This data underscores the need for organizations to establish formal policies governing the use of generative AI writing tools. Alongside such policies, an efficient and reliable detection tool for AI-generated content is crucial. Since the launch of Turnitin's AI writing detection tool on April 4, 2023, over 50 million submissions have been processed globally (Turnitin, 2024). Nevertheless, Turnitin has acknowledged instances of misclassifying human-written text as AI-generated, highlighting the issue of false positives. This scenario suggests that current AIGC detection tools are not yet sufficiently reliable.

Gaps Identified in Existing Solutions

The survey identified several gaps in the current solutions for academic misconduct, focusing on data, methodology, and overall effectiveness. These gaps are detailed below.

Datasets:

- Many algorithms for AIGC detection rely on simplistic datasets, primarily consisting of text generated entirely by ChatGPT.
- High evaluation metrics are often achieved due to testing on these naive datasets, which do not reflect real-world scenarios.
- There is a need for benchmark datasets that include examples of paraphrasing and evasion techniques, which should be developed and made available to the public.

Model Solutions:

- Most models assess the probability of AIGC for entire texts rather than for individual sections.
- There is a need for research focused on detecting AIGC probability at the paragraph or sentence level to enhance accuracy.
- Many models lack explainability in their predictions. Incorporating post hoc model-agnostic techniques, such as SHAP, could help explain predictions, as seen in text classification and sentiment analysis (Mosca et al., 2022).

From a methodological perspective, existing AIGC detection algorithms generally evaluate the entire text to determine if it is AI-generated. However, this approach is vulnerable to various evasion techniques. Instead, developing solutions that focus on the main gist of the text-such as key topics or summaries-might better withstand these evasion tactics.

Moreover, training algorithms on datasets that include examples of evasion techniques could improve their resistance to such tactics. Current approaches, including binary classification, zero-shot techniques, and watermarking, have limitations. Binary classifiers often struggle with new LLM-generated texts, watermarking is ineffective if not all texts are watermarked, and zero-shot approaches falter when faced with evasion techniques. Thus, there is a need for innovative techniques in AIGC detection.

Lastly, most research focuses solely on classifying text as AI-generated or human-written without addressing broader issues such as preventing academic cheating, redesigning assessment strategies, and encouraging ethical AI use in academia. These areas also warrant exploration.

Discussion

It is clear from the experiments and analysis in Section "Detecting AI-Generated Plagiarism" that a reliable solution to prevent academic misconduct due to large language models (LLMs) is crucial. This section explores potential solutions for addressing AI-driven plagiarism, including not only technical approaches but also educational strategies and ethical considerations for LLM adoption that academic institutions should prioritize.

Feasibility of AIGC Detection

The feasibility of detecting AI-generated text was a central focus of a recent survey on various aspects of text generation using contemporary LLMs. Ghosal et al. (2023) mentioned that achieving 100% certainty in detecting AI-generated text is highly unlikely. Raj (2023), a professor at the University of British Columbia, adds that detecting AIGC is challenging because the primary goal of LLMs is to produce increasingly human-like text. Furthermore, even if a highly effective AIGC detection tool is developed, it is likely to be rendered ineffective by the next generation of LLMs.

Walters (2023) observed that while many AI detectors can differentiate between papers generated by the GPT-3.5 model and human-written papers, they struggle with texts produced by the more recent GPT-4 model. Additionally, Weber-Wulff et al. (2023) found that experiments with 12 publicly available AIGC detectors and two commercial systems-Turnitin and PlagiarismCheck-revealed that these tools are often fooled by evasion techniques such as translation and paraphrasing. This suggests that a truly reliable solution for AIGC detection remains elusive and may continue to be out of reach.

Trustworthiness of LLMs

The trustworthiness of LLMs and their generated text presents significant challenges in academic settings. A major concern is the dissemination of false information, such as hallucinations, which undermines the reliability of LLMs for academic use (Sun et al., 2024). Another critical issue is data bias, which can arise from the data used to train LLMs (Brown, 2024). Additionally, Wang et al. (2023) highlights that LLMs are vulnerable to adversarial

attacks, which can lead to the generation of biased or harmful content. Privacy and data security concerns further complicate the adoption of LLMs in academia, raising questions about the trustworthiness of both the models and the content they produce (Farhat et al., 2023).

Alternative Educational Solutions

Following an extensive analysis of over 50 examples of preventive measures against academic cheating, incorporating experiences from various academic institutions, Bylieva et al. (2020) determined that relying solely on technical solutions is insufficient. Technical measures alone lead to an ongoing cycle of countermeasures, which may be ineffective in the long term. Therefore, academic institutions should explore and implement non-technical, alternative educational solutions.

A crucial step in addressing academic misconduct is increasing awareness of issues such as plagiarism and the inappropriate use of AI tools like ChatGPT. A survey of 3,405 students at an Australian university revealed that only 50% had read the plagiarism policy, and many were unclear about what constitutes plagiarism (Gullifer & Tyson, 2014). Additionally, the university's complaint office noted that many grievances were related to insufficient warnings about plagiarism and its consequences (Guardian, 2012). The European Network for Academic Integrity (ENAI) also recommends promoting the ethical use of AI tools in academia. Their guidelines include training for both students and educators on the ethical implications, biases, and limitations of AI tools and providing clear instructions on their appropriate use (Foltynek et al., 2023). Ensuring robust awareness of academic misconduct is essential.

Enhancing current assessment strategies is another effective approach to curbing academic misconduct. Assessments should prioritize evaluating students' critical thinking and problem-solving abilities rather than rote memorization. Given that simple factual questions can be easily answered by language models, assignments should focus on creative projects and open-book exam questions that challenge students to think critically and solve problems (Egan, 2018). Additionally, incorporating visual or interactive elements into assessments, such as analyzing videos or engaging in oral presentations, can help prevent cheating through AI tools (Harper et al., 2021). Timed assignments are another effective strategy, as they limit the opportunity for students to consult external resources like ChatGPT by restricting the time available to complete each question (Holden et al., 2021).

By adopting these alternative educational solutions, institutions can foster a more integrity-driven academic environment and reduce the prevalence of academic misconduct.

Ethical Adoption of LLMs in Academia

One effective solution to address academic misconduct is the ethical and creative integration of AI tools within academic settings. Dr. Amin Davodi advocates for teaching students how to use AI tools for learning purposes rather than as a means of cheating (Bryson, 2023). Universities could permit the use of tools like ChatGPT but with certain restrictions. For example, students might use these tools for reviewing and correcting their work but should not rely on them for direct copy-pasting of content. Implementing a threshold for the proportion of content generated by AI in student assignments can help regulate this practice. Universities and scientific publishers might establish specific guidelines for the acceptable use of AI tools like ChatGPT.

Oravec (2023) suggests designing assignments that promote transparency in the use of AI tools. For instance, assignments could require students to compare their work with ChatGPT's output, review, and revise their submissions accordingly (Oravec, 2023). Such approaches encourage students to use AI constructively and engage in a more meaningful learning process. Additionally, the European Network for Academic Integrity (ENAI) recommends appropriate citation practices for AI tool usage to ensure ethical compliance (Foltynek et al., 2023). Abd-Elaal et al. (2019) also emphasize the need for academic institutions to enhance their plagiarism and fabrication policies, while ENAI advises developing national and institutional guidelines for the ethical use of AI in academia (Foltynek et al., 2023).

The adoption of LLMs in academia also presents several ethical challenges. One major concern is determining authorship and intellectual property rights for texts generated with LLMs, particularly in collaborative environments where LLMs might contribute partially (Meyer et al., 2023). Another issue is addressing biases and ensuring fairness in LLM-generated content. Since LLMs are trained on extensive publicly available datasets, there is a risk of including sensitive or confidential information, which complicates data privacy and security (Sun et al., 2024). Therefore, the integration of LLMs into academic settings must be accompanied by clear guidelines and policies to ensure ethical use and accountability.

The challenge of detecting plagiarism involving LLMs is inherently interdisciplinary, requiring collaboration among experts in computer science, education, ethics, and policy. From a computer science perspective, it is crucial to develop innovative methods and tools specifically designed to identify LLM-generated text. This involves creating algorithms that can effectively differentiate between human and AI-generated content and improve the accuracy of detection systems.

Educators and academic institutions play a vital role in integrating these detection tools into their pedagogical practices. They should design assignments and assessments that are less vulnerable to LLM manipulation, thus preserving academic integrity. By emphasizing critical thinking and original work, educators can reduce the likelihood of students relying on AI tools inappropriately.

Ethical considerations are also paramount. Issues such as data privacy, the potential for misinformation due to LLM hallucinations, and the need to balance surveillance with trust must be addressed. Developing clear guidelines for the ethical use of LLMs in academic contexts is essential to ensure that privacy concerns are managed and that AI tools are used responsibly.

From a policy perspective, establishing standardized regulations is necessary to govern the ethical adoption and use of LLMs in academia. These policies should provide a framework for both the detection of plagiarism and the responsible integration of AI tools into educational settings.

Conclusion

In this research, we explored various methods of academic misconduct and strategies for its prevention, with a particular focus on the impact of large language models (LLMs) on academic cheating. We examined how the advent of generative AI tools has transformed traditional forms of cheating, such as plagiarism. Our analysis included both the development of algorithms and tools designed to detect AI-generated text and plagiarism, as well as the evasion techniques employed to circumvent these systems. Our findings indicate that

current solutions for detecting plagiarism and AI-generated content (AIGC) are insufficiently reliable.

Our survey revealed several gaps in existing datasets and detection solutions for plagiarism and AIGC. Given the ongoing technical arms race between detection technologies and evasion tactics, we propose that academic institutions adopt alternative educational strategies.

Researchers should also prioritize non-technical solutions because the technical war between plagiarism detection and evasion is ongoing. In light of the challenges posed by LLMs in academia, there is a pressing need to develop and implement policies and regulations governing the ethical use of these tools.

Every academic individual should adhere to these policies and regulations. Educators should focus on raising awareness about plagiarism and developing innovative assessment strategies, while students should commit to the ethical use of AI tools in their work.

For future work, several research areas in the field of AIGC detection could be explored. Despite significant efforts, the reliability of current AIGC detection tools remains questionable, partly due to the lack of quality benchmarks. Addressing this issue could involve creating a benchmark dataset for AIGC detection tasks. Additional experiments could explore the effectiveness of combining various methods to enhance the accuracy of detection algorithms. Beyond technical solutions, research into non-technical approaches to AI-based plagiarism and the explainability of AIGC detection models is also crucial. Developing and releasing a quality benchmark dataset, creating hybrid detection models, and improving the transparency of AIGC detection outputs are key areas for future work.

Glossary

The following concepts are essential in understanding our survey on AI-generated plagiarism.

AIGC (AI-Generated Content)	Content created by Artificial Intelligence (AI) systems, including text and media.
AIGC Plagiarism	Present AI-generated content as original work without making any reference.
ChatGPT	A conversational AI model by OpenAI that generates human-like text responses.
DL (Deep Learning)	A machine learning subset using multi-layered neural networks for tasks like image recognition and language processing.
Gemini	A suite of LLMs by Google DeepMind for understanding and generating text.
LLM (Large Language Model)	An AI model trained on vast text data to generate human-like text.
NLP (Natural Language Processing)	A field of AI focused on the interaction between computers and human language, enabling tasks like text analysis and language generation.
OpenAI	An AI research organization known for developing advanced models like ChatGPT.
Plagiarism	Using someone else's work without attribution, claiming it as your own.
Pretrained Language Model	An AI model trained on large text datasets, then fine-tuned for specific tasks.

Acknowledgements This publication has emanated from research [conducted with the financial support of/supported in part by a grant from] Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

References

Aaronson, S. (2022). My ai safety lecture for ut effective altruism. https://scottaaronson.blog/?p=6823. Accessed 19 Mar 2024.

Abd-Elaal, E.-S., Gamage, S., Mills, J. E., et al. (2019). Artificial intelligence is a tool for cheating academic integrity. In *30th annual conference for the australasian association for engineering education (aaee 2019): Educators becoming agents of change: Innovate, integrate, motivate* (pp. 397–403).

Abdelhamid, M., Azouaou, F., & Batata, S. (2022). A survey of plagiarism detection systems: Case of use with english, french, and arabic languages. *arXiv Preprint*. Available at arxiv:2201.03423

Abdelnabi, S., & Fritz, M. (2021). Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 121–140). IEEE. Los Alamitos, CA, USA.

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

AlSallal, M., Iqbal, R., Palade, V., Amin, S., & Chang, V. (2019). An integrated approach for intrinsic plagiarism detection. *Future Generation Computer Systems, 96*, 700–712. https://doi.org/10.1016/j.future.2018.03.044

Alzahrani, S. (2015). Arabic plagiarism detection using word correlation in n-grams with k-overlapping approach. In *Proceedings of the Workshops at the 7th Forum for Information Retrieval Evaluation (FIRE)* (pp. 123–125).

Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism: Linguistic patterns, textual features, and detection methods. https://ieeexplore.ieee.org/abstract/document/5766764. Accessed 07 Mar 2024

Antoun, W., Mouilleron, V., Sagot, B., & Seddah, D. (2023). Towards a robust detection of language model generated text: Is chatgpt that easy to detect? *arXiv preprint* arXiv:2306.05871

ArgaAssociation. (2019). Plagiarism Statistics – Academic Research Guide Association — argassociation.org. Accessed 07 Mar 2024.

Bao, G., Zhao, Y., Teng, Z., Yang, L., & Zhang, Y. (2023). Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint* arXiv:2310.05130

Bin-Nashwan, S. A., Sadallah, M., & Bouteraa, M. (2023). Use of chatgpt in academia: Academic integrity hangs in the balance. *Technology in Society, 75*, 102370.

Biörck, J., & Eriksson, S. (2023). Diva-portal.org. https://www.diva-portal.org/smash/get/diva2:1779786/FULLTEXT01.pdf. Accessed 14 Mar 2024.

Blat, F., Castro, M. J., Tortajada, S., & Sánchez, J. A. (2005). A hybrid approach to statistical language modeling with multilayer perceptrons and unigrams. In *Advances in neural information processing systems 18*. Springer. Retrieved from https://link.springer.com/chapter/10.1007/11551874_25. Accessed 04 Mar 2024.

Brown, N. B. (2024). Enhancing trust in llms: Algorithms for comparing and interpreting llms. arXiv preprint arXiv:2406.01943

Bryson, E. (2023). How To Prevent Students from Cheating with AI. https://ellii.com/blog/how-to-prevent-students-cheating-with-ai. Accessed 04 Apr 2024.

Bylieva, D., Lobatyuk, V., Tolpygin, S., & Rubtsova, A. (2020). Academic dishonesty prevention in e-learning university system. In *World conference on information systems and technologies* (pp. 225–234). Springer.

Cai, S., & Cui, W. (2023). Evade chatgpt detectors via a single space. arXiv preprint arXiv:2307.02599

Chaika, O., Domina, V., Nikolaienko, S., & Fedosii, O. (2023). Zero tolerance to plagiarism in multicultural teamwork: Challenges for english-speaking non-eu and eu academics. *World Journal of English Language, 13*(4), 1–14. https://doi.org/10.5430/wjel.v13n4p1

Chelba, C., Norouzi, M., & Bengio, S. (2017). N-gram language modeling using recurrent neural networks estimation. Retrieved from https://arxiv.org/pdf/1703.10724.pdf

Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Raj, B. (2023). Gpt-sentinel: Distinguishing human and chatgpt generated content. *arXiv preprint*arXiv:2305.07969

Chitra, A., & Rajkumar, A. (2016). Plagiarism detection using machine learning-based paraphrase recognizer. *Journal of Intelligent Systems, 25*(3), 351–359. https://doi.org/10.1515/jisys-2016-0025

Chui, H. C. (2024). Chatgpt as a tool for developing paraphrasing skills among esl learners. Retrieved from https://www.researchgate.net/publication/375799764_ChatGPT_as_a_Tool_for_Developing_Paraphrasing_Skills_Among_ESL_Learners

Chui, M., Hazan, E., Roberts, R., Singla, A., & Smaje, K. (2023). The economic potential of generative ai.

Davison, J. (2020). Zero-shot learning in modern nlp. https://joeddav.github.io/blog/2020/05/29/ZSL.html. Accessed 20 Mar 2024.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*arXiv:1810.04805

Ed, I. H. (2023). Turnitin's ai detector in higher ed expected to have false positives. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/quick-takes/2023/06/01/turnitins-ai-detector-higher-expected-false-positives. Accessed 09 Aug 2024

Egan, A. (2018). Improving academic integrity through assessment design. *Dublin City University*.

Eisner, C., & Vicinus, M. (2008). Originality, Imitation, and Plagiarism. https://library.oapen.org/bitstream/handle/20.500.12657/24007/1/1006126.pdf. Accessed 07 Mar 2024.

El Mostafa Hambi, F., & Benabbou, F. (2020). A new online plagiarism detection system based on deep learning. *International Journal of Advanced Computer Sciences and Applications, 11*(9), 470–478.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2021). Some students' plagiarism tricks and tips for effective check. *International Journal for Educational Integrity, 17*, 1–12. https://doi.org/10.1007/s40979-021-00092-w

Eriksson, G., & Karlgren, J. (2012). Features for modelling characteristics of conversations: Notebook for pan at clef 2012. In *Clef 2012 evaluation labs and workshop - working notes papers*. September 17-20, Rome, Italy: CEUR-WS.org.

Farhat, F., Sohail, S. S., & Madsen, D. Ø. (2023). How trustworthy is chatgpt? the case of bibliometric analyses. *Cogent Engineering, 10*(1), 2222988.

Foltynek, T., Bjelobaba, S., Glendinning, I., Khan, Z. R., Santos, R., Pavletic, P., & Kravjar, J. (2023). *Enai recommendations on the ethical use of artificial intelligence in education*. Springer.

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys (CSUR), 52*(6), 1–42. https://doi.org/10.1145/3345317

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. *BioRxiv,* 2022–12. https://doi.org/10.1101/2022.12.19.521287

Gaurdian, T. (2023). AI makes plagiarism harder to detect, argue academics – in paper written by chatbot | theguardian.com. Accessed 09 Mar 2024.

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). Gltr: Statistical detection and visualization of generated text. *arXiv preprint*. Retrieved from arXiv:1906.04043

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 325–336). ACM.

Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, A. S. (2023). Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv preprint*arXiv:2310.15264

Giboulot, E., & Teddy, F. (2024). Watermax: Breaking the llm watermark detectability-robustness-quality trade-off. *arXiv preprint*. Retrieved from arXiv:2403.04808

Gillham, J. (2024). Huggingface statistics – originality.ai. Retrieved from https://originality.ai/blog/huggingface-statistics

Guardian, T. (2012). Universities need to tell students the rules about plagiarism, says adjudicator. https://www.theguardian.com/education/2012/jun/11/universities-students-rules-plagiarism-adjudicator. Accessed 28 Mar 2024.

Guillén-Nieto, V. (2022). Plagiarism detection: Methodological approaches. In *Language as evidence: doing forensic linguistics* (pp. 321–372). Springer.

Gullifer, J. M., & Tyson, G. A. (2014). Who has read the policy on plagiarism? Unpacking students' understanding of plagiarism. *Studies in Higher Education, 39*(7), 1202–1218.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., . . . Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv Preprint*. Available at arXiv:2301.07597

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., . . . Goldstein, T. (2024). Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint*arXiv:2401.12070

Hao, W., Xiang, L., Li, Y., Yang, P., & Shen, X. (2018). Reversible natural language watermarking using synonym substitution and arithmetic coding. *Computer Speech & Language, 52*, 139–154. https://doi.org/10.1016/j.csl.2018.04.001

Harker, J. (2023). Science journals set new authorship guidelines for ai-generated text. Retrieved from https://factor.niehs.nih.gov/2023/3/feature/2-artificial-intelligence-ethics

Harper, R., Bretag, T., & Rundle, K. (2021). Detecting contract cheating: Examining the role of assessment type. *Higher Education Research & Development, 40*(2), 263–278.

He, Z., Zhou, B., Hao, H., Liu, A., Wang, X., Tu, Z., . . . Wang, R. (2024). Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models. *arXiv preprint*arXiv:2402.14007

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A largescale comparison of human-written versus chatgpt-generated essays. *Scientific Reports, 13*(1), 18617. https://doi.org/10.1038/s41598-023-41872-6

Holden, O. L., Norris, M. E., & Kuhlmeier, V. A. (2021). Academic integrity in online assessment: A research review. In *frontiers in education* (Vol. 6, p. 639814). Frontiers Media SA.

Huallpa, J. J., et al. (2023). Exploring the ethical considerations of using chatgpt in university education. *Periodicals of Engineering and Natural Sciences, 11*(4), 105–115.

Hu, X., Chen, P.-Y., & Ho, T.-Y. (2023). Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems, 36*, 15077–15095.

Huo, M., Somayajula, S. A., Liang, Y., Zhang, R., Koushanfar, F., & Xie, P. (2024). Token-specific watermarking with enhanced detectability and semantic coherence for large language models. *arXiv preprint*. Retrieved from arXiv:2402.18059

Ian. (2023). Science journals ban listing of chatgpt as co-author on papers. Retrieved from https://www.theguardian.com/science/2023/jan/26/science-journals-banlisting-of-chatgpt-as-co-author-on-papers

Ison, D. C. (2016). Academic misconduct and the internet. ResearchGate. https://www.researchgate.net/publication/301234567_Academic_Misconduct_and_the_Internet. Accessed 07 Mar 2024.

Jovanović, N., Staab, R., & Vechev, M. (2024). Watermark stealing in large language models. arXiv preprint arXiv:2402.19361

Kalla, D., & Smith, N. (2023). Study and analysis of chatgpt and its impact on different fields of study. *International Journal of Innovative Science and Research Technology, 8*(3), 827–833.

Kalyan, K. S. (2023). A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 100048. Retrieved from https://doi.org/10.1016/j.nlp.2023.100048

Kashnitsky, Y. (2022). Source code for the coling workshop competition "detecting automatically generated scientific papers". https://github.com/Yorko/fake-papers-competition-data. GitHub repository. GitHub.

Katib, I., Assiri, F. Y., Abdushkour, H. A., Hamed, D., & Ragab, M. (2023). Differentiating chat generative pretrained transformer from humans: Detecting chatgpt-generated text and human text using machine learning. *Mathematics, 11*(15), 3400.

Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., & Stein, B. (2019). Overview of the cross-domain authorship attribution task at PAN 2019. In *Working notes of CLEF 2019: conference and labs of the evaluation forum, Lugano, Switzerland, September 9-12, 2019* (pp. 1–15).

Khaled, F., & Al-Tamimi, M. S. H. (2021). Plagiarism detection methods and tools: An overview. *Iraqi Journal of Science*, 2771–2783.

Khalil, M., & Er, E. (2023). Will chatgpt get you caught? Rethinking plagiarism detection. In *Proceedings of the international conference on human-computer interaction* (pp. 475–487). Springer.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In *Proceedings of the international conference on machine learning* (pp. 17061–17084). PMLR. Baltimore, MD, USA.

Kirchner, J. H. (2023a). New ai classifier for indicating ai-written text. Retrieved from https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

Kirchner, J. H. (2023b). New ai classifier for indicating ai-written text. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text. Accessed 20 Mar 2024.

Koupaee, M., & Wang, W. Y. (2018). Wikihow: A large-scale text summarization dataset. *arXiv Preprint*. Available at arXiv:1810.09305

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2024). Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems, 36*.

Kuhn, T., Niemann, H., & Schukat-Talamazzin, E. (1994). Ergodic hidden markov models and polygrams for language modeling. *IEEE Transactions on Speech and Audio Processing*. Retrieved from https://ieeexplore.ieee.org/abstract/document/389282. Accessed 04 Mar 2024

Lee, J., Agrawal, T., Uchendu, A., Le, T., Chen, J., & Lee, D. (2024). Plagbench: Exploring the duality of large language models in plagiarism generation and detection. Retrieved from arxiv:2406.16288

Li, Z., Yang, Z., & Wang, M. (2023). Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint*arXiv:2305.18438

Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., . . . Liu, T., et al. (2023). Differentiate chatgpt-generated and human-written medical texts. *arXiv Preprint*. Available at arXiv:2304.11567

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., . . . Hu, H. (2023). Argugpt: Evaluating, understanding and identifying argumentative essays generated by gpt models.

Lu, N., Liu, S., He, R., Wang, Q., Ong, Y.-S., & Tang, K. (2023). Large language models can be guided to evade ai-generated text detection. arXiv preprint arXiv:2305.10847

Macko, D., Moro, R., Uchendu, A., Srba, I., Lucas, J. S., Yamashita, M., . . . Bielikova, M. (2024). Authorship obfuscation in multilingual machine-generated text detection. *arXiv preprint*arXiv:2401.07867

Mao, C., Vondrick, C., Wang, H., & Yang, J. (2024). Raidar: Generative ai detection via rewriting. *arXiv preprint*arXiv:2401.12970

Martin, B. (1992). Plagiarism by university students: The problem and some proposals. https://documents.uow.edu.au/~bmartin/pubs/92tert.html. Accessed 07 Mar 2024.

Meral, H. M., Sankur, B., Özsoy, A. S., Güngör, T., & Sevinç, E. (2009). Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language, 23*(1), 107–125. https://doi.org/10.1016/j.csl.2008.02.003

Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P.-C., Gonzalez-Hernandez, G., et al. (2023). Chatgpt and large language models in academia: Opportunities and challenges. *BioData Mining, 16*(1), 20.

Mireshghallah, F., Mattern, J., Gao, S., Shokri, R., & Berg-Kirkpatrick, T. (2023). Smaller language models are better black-box machine-generated text detectors. *arXiv preprint*. Retrieved from arXiv:2305.09859

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning* (pp. 24950–24962). PMLR.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). Shap-based explanation methods: A review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4593–4603).

Mphahlele, A., & McKenna, S. (2019). The use of turnitin in the higher education sector: Decoding the myth. *Assessment & Evaluation in Higher Education, 44*(7), 1079–1089. https://doi.org/10.1080/02602938.2018.1526767

Nah, F.F.-H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative ai and chatgpt: Applications, challenges, and ai-human collaboration. *Information Systems Management*. https://doi.org/10.1080/15228053.2023.2233814

Nerdynav. (2024). Chatgpt cheating statistics & impact on education (2024). Retrieved from https://nerdynav.com/chatgpt-cheating-statistics/

Nicks, C., Mitchell, E., Rafailov, R., Sharma, A., Manning, C. D., Finn, C., & Ermon, S. (2023). Language model detectors are easily optimized against. In *The twelfth international conference on learning representations*.

Nolan, B. (2023). Here are the schools and colleges that have banned the use of chatgpt over plagiarism and misinformation fears. Retrieved from https://www.businessinsider.com/chatgpt-schools-colleges-ban-plagiarism-misinformation-education-2023-1?r=US&IR=T

Oghaz, M. M. D., Dhame, K., Singaram, G., & Saheer, L. B. (2023). Detection and classification of chatgpt generated contents using deep transformer models. *Authorea Preprints*.

Oravec, J. A. (2023). Artificial intelligence implications for academic cheating: Expanding the dimensions of responsible human-ai collaboration with chatgpt. *Journal of Interactive Learning Research, 34*(2), 213–237.

Originality.AI. (2024). Ai content in google search results - originality.ai. Retrieved from https://originality.ai/ai-content-in-google-searchresults. Accessed 09 Aug 2024

Pallagani, V., Muppasani, B., Murugesan, K., Rossi, F., Srivastava, B., Horesh, L., . . . Loreggia, A. (2023). Understanding the capabilities of large language models for automated planning. Retrieved from arxiv:2305.16151

Patel, A., Bakhtiyari, K., & Taghavi, M. (2011). Evaluation of cheating detection methods in academic writings. *Library Hi Tech, 29*(4), 623–640. https://doi.org/10.1108/07378831111189554

Peng, X., Zhou, Y., He, B., Sun, L., & Sun, Y. (2024). Hidding the ghostwriters: An adversarial evaluation of ai-generated student essay detection. *arXiv preprint*arXiv:2402.00412

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Retrieved from arxiv:1802.05365

Porter, J. (2023). Chatgpt continues to be one of the fastest-growing services ever. Retrieved from https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference

Potthast, M., Stein, B., Eiselt, A., Barrón-Cede no, A., Rosso, P., et al. (2009). Overview of the 1st international competition on plagiarism detection. In *CEUR Workshop Proceedings* (Vol. 502, pp. 1–9).

Pushp, P. K., & Srivastava, M. M. (2017). Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint*. Retrieved from arXiv:1712.05972

Raj, A. (2023). Finding the real author with turnitin ai detection. https://techwireasia.com/06/2023/turnitin-ai-detection-tackling-the-issue-of-academic-integrity/. Accessed 07 Apr 2024.

Ravi, N. R., Vani, K., & Gupta, D. (2016). Exploration of fuzzy c-means clustering algorithm in external plagiarism detection system. In *Intelligent systems technologies and applications: Vol. 1* (pp. 127–138). Springer.

Shijaku, R., & Canhasi, E. (2023). *Chatgpt generated text detection*. Publisher: Unpublished.

Simon, J. (2024). Large language models: A new moore's law? Retrieved from https://huggingface.co/blog/large-language-models

Smolansky, A., Cram, A., Raduescu, C., Zeivots, S., Huber, E., & Kizilcec, R. F. (2023). Educator and student perspectives on the impact of generative ai on assessments in higher education. (pp. 378–382).

Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., . . . Mansoor, W. (2023). Decoding chatgpt: A taxonomy of existing research, current challenges, and possible future directions. *Journal of King Saud University- Computer and Information Sciences*, 101675.

Stein, B., Potthast, M., Rosso, P., Barrón-Cedeno, A., Stamatatos, E., & Koppel, M. (2011). Fourth international workshop on uncovering plagiarism, authorship, and social software misuse. In *ACM SIGIR Forum* (Vol. 45, pp. 45–48). ACM New York, NY, USA.

Stern, E. B., & Havlicek, L. (2024). Academic misconduct: Results of faculty and undergraduate student surveys. Retrieved from https://www.jstor.org/stable/45445129

Stokel-Walker, C. (2023a). Chatgpt listed as author on research papers: Many scientists disapprove. Retrieved from https://www.nature.com/articles/d41586-023-00107-z

Stokel-Walker, C. (2023b). Chatgpt listed as author on research papers: Many scientists disapprove. Retrieved from https://www.nature.com/articles/d41586-023-00107-z

Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., . . . Li, X., et al. (2024). Trustllm: Trustworthiness in large language models. arXiv preprint arXiv:2401.05561

TimeForClass. (2023). Tytonpartners.com. https://tytonpartners.com/app/uploads/2023/06/Time-for-Class-2023-Report_Final.pdf. Accessed 28 Mar 2024.

Topkara, U., Topkara, M., & Atallah, M. J. (2006). The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on multimedia and security* (pp. 164–174). New York, USA: ACM.

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & de Visser, E. J. (2024). Student perceptions of chatgpt use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. IEEE Transactions on Learning Technologies.

Turnitin. (2023). Understanding false positives within our AI writing detection capabilities. https://www.turnitin.com/blog/understanding-false-positives-within-our-ai-writing-detection-capabilities. Accessed 28 Mar 2024.

Turnitin. (2024). Turnitin celebrates 25 years in global academic integrity. Retrieved from https://www.turnitin.com/press/turnitin-celebrates-25-years-in-global-academic-integrity. Accessed 09 Aug 2024.

Varanasi, L. (2023). GPT-4 can ace the bar, but it only has a decent chance of passing the CFA exams. Here's a list of difficult exams the ChatGPT and GPT-4 have passed. — businessinsider.com. Accessed 09 Mar 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2023). Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint* arXiv:2305.15047

Walters, W. H. (2023). The effectiveness of software designed to detect ai-generated writing: A comparison of 16 ai text detectors. *Open Information Science, 7*(1), 20220158.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., . . . Schaeffer, R., et al. (2023). Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *neurips*.

Wang, R., Chen, H., Zhou, R., Ma, H., Duan, Y., Kang, Y., . . . Tan, T. (2024). Llmdetector: Improving ai-generated chinese text detection with open-source llm instruction tuning. *arXiv preprint*arXiv:2402.01158

Wang, Y., Feng, S., Hou, A. B., Pu, X., Shen, C., Liu, X., . . . He, T. (2024). Stumbling blocks: Stress testing the robustness of machine-generated text detectors under attacks. *arXiv preprint*arXiv:2402.11638

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., . . . Arnold, T., et al. (2024). M4gt-bench: Evaluation benchmark for black-boxmachine-generated text detection. *arXiv Preprint*. Available at arXiv:2402.11175

Wang, Y. [Yuxia], Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., . . . Nakov, P. (2024). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th conference of the European chapter of the association*

for computational linguistics (Vol. 1: Long Papers)(pp. 1369–1407). The dataset is available at https://github.com/mbzuai-nlp/M4. St. Julian's, Malta: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.eacl-long.83

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., & Waddington, L. (2023). Testing of detection tools for aigenerated text. *International Journal for Educational Integrity, 19*(1), 26.

Westfall, C. (2023). Educators battle plagiarism as 89

Whalen, J., Mouza, C., et al. (2023). Chatgpt: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education, 23*(1), 1–23.

Wiggers, K. (2023). Openai releases tool to detect ai-generated text, including from chatgpt — techcrunch. https://techcrunch.com/2023/01/31/openai-releases-tool-to-detect-ai-generated-text-including-from-chatgpt/. Accessed 21 Mar 2024.

Wu, Q., & Chandrasekaran, V. (2024). Bypassing llm watermarks with color-aware substitutions. arXiv preprint arXiv:2403.14719

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., . . . Zhou, E., et al. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864

Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning: The good, the bad, and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4582–4591). IEEE.

Yang, X., Chen, K., Zhang, W., Liu, C., Qi, Y., Zhang, J., . . . Yu, N. (2023). Watermarking text generated by black-box language models. *arXiv preprint*. Retrieved from arXiv:2305.08883

Yang, X. [Xi], Zhang, J., Chen, K., Zhang, W., Ma, Z., Wang, F., & Yu, N. (2022). Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 11613–11621). New York, NY, USA: AAAI Press.

Yang, X. [Xianjun], Cheng, W., Petzold, L.,Wang, W. Y., & Chen, H. (2023). Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint*. Retrieved from arXiv:2305.17359

Yu, P., Chen, J., Feng, X., & Xia, Z. (2023). Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *arXiv Preprint*. Available at arXiv:2304.12008

Yuqin, J., & Liu, Y. [Ying]. (2023). A population-based plagiarism detection using distilbert-generated word embedding. *International Journal of Advanced Computer Science and Applications, 14*(8). https://doi.org/10.14569/IJACSA.2023.0140827

Zhao, X., Ananth, P., Li, L., & Wang, Y.-X. (2023). Provable robust watermarking for ai-generated text. *arXiv preprint*. Retrieved from arXiv:2306.17439

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., . . . Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint*arXiv:2303.18223